# ACQUIRING AND ANALYSING DATA IN SUPPORT OF EVIDENCE-BASED DECISIONS

## A GUIDE FOR HUMANITARIAN WORK


ICRC

REFERENCE


ICRC

International Committee of the Red Cross
19, avenue de la Paix
1202 Geneva, Switzerland
T +41 22 734 60 01 F +41 22 733 20 57
E-mail: shop@icrc.org www.icrc.org

# ACQUIRING AND ANALYSING DATA IN SUPPORT OF EVIDENCE-BASED DECISIONS

## A GUIDE FOR HUMANITARIAN WORK

# INTRODUCTION
# ACQUIRING AND ANALYSING DATA

Gathering and analysing data – to study the consequences, in humanitarian terms, of crises and to carry out programmes and activities – is an essential element of humanitarian work. All its assistance strategies require the International Committee of the Red Cross (ICRC) to assess situations and the needs of people affected, and to monitor, review and evaluate its humanitarian activities; this helps to ensure the relevance and effectiveness of its humanitarian work and enables it to remain accountable to the people it seeks to help.

Information generated through assessments – and over the course of monitoring, reviewing and evaluating projects and programmes – is used to make various decisions: when and where to prioritize humanitarian action, whom to assist, what type of activities to undertake, etc. Such information can also be the basis for modifying existing programmes and activities, and for recommendations for the future. Methods for acquiring and analysing data must be reliable and efficient; this is critical for ensuring the soundness of the evidence on which decisions are based.[1]

## GENERATING INFORMATION

Information is generated through a series of actions that are performed until a minimum level of clarity on the subject under study is attained. The entire process is iterative: methods, tools and data are refined continuously as patterns and insights emerge. The duration of each action and the amount of time required to achieve the necessary clarity will vary greatly from one exercise to the next. The objective is to accumulate sound (or sounder) evidence that answers the questions that necessitated the process.

Figure 1 provides an overview of the process. The chart shows that, ideally, the level of clarity should increase at every step of the process. However, sometimes the initial steps of gathering and analysing data may not reveal any patterns or generate any insights; in such cases, analysts will have to go back and re-evaluate their design and the sources of their data. As the chart shows, some exercises will involve collection of primary data and others will not.

1 Evidence is, "information which helps to demonstrate the truth or falsehood of a given hypothesis or proposition," (Clark and Darcy, *Insufficient Evidence? The quality and use of evidence in humanitarian action*, 2014).

# GENERATING INFORMATION

Information is generated through a series of actions that are performed until a minimum level of clarity on the subject under study is attained. The entire process is iterative: methods, tools and data are refined continuously as patterns and insights emerge. The duration of each action and the amount of time required to achieve the necessary clarity will vary greatly from one exercise to the next. he objective is to accumulate sound (or sounder) evidence that answers the questions that necessitated the process.

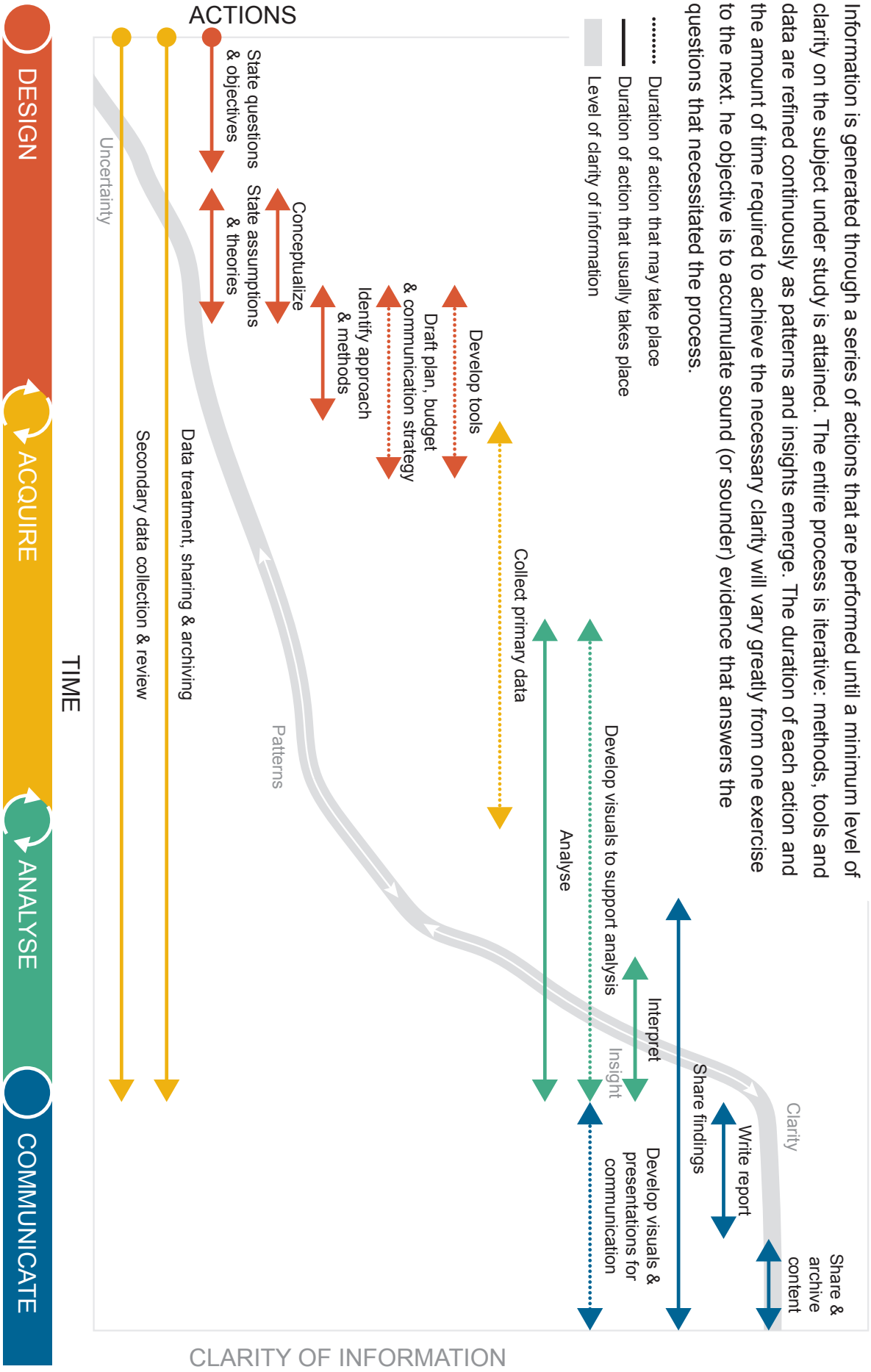



**Figure 1** - Process of generating information

## ABOUT THIS GUIDE

This guide is meant to be both a technical source of reference and a tool for any ICRC staff member working with data and conducting analyses. It was prepared through the lens of the ICRC Economic Security (EcoSec) Unit, and to serve the unit's operational needs. However, many elements are transversal to other disciplines.

It covers qualitative, quantitative and mixed-method approaches to analysis, and data collection and analysis methods widely used in the field. It focuses on assessment, monitoring, reviewing and evaluation activities that involve field visits and/or surveys. The guide does not provide in-depth coverage of secondary-data collation, big-data analysis or reporting.

It reviews various types of data and approaches to analysis (Chapters 1 and 2); then it describes the process of designing an analysis (Chapter 3); this is followed by a consideration of the collection and analysis of primary data (Chapters 3 through 9); the concluding chapter takes up the subject of visualization (Chapter 10).

The guide is intended for humanitarian personnel who are not specialists in data analysis, information management, information technology or database or Web development, but are required to undertake data collection and analysis exercises. It seeks to support them in the performance of key tasks, and covers the basic principles of developing databases, and descriptive and inferential statistics, using standard ICRC software available to all staff members. In some cases, the requirements for collecting and analysing data will be beyond the scope of this guide; more sophisticated software and techniques will be needed, as well as the support of a data analyst or statistician, information management specialist, information technology specialist or database or Web developer. Attention is drawn to these cases throughout the guide.

This is the second version of the guide. The first was produced and made available within the ICRC in May 2015. This second version was revised after an internal ICRC review and with the help of comments and suggestions from field specialists at the Assessment Capacities Project (ACAPS), the World Food Programme's Vulnerability Analysis and Mapping Unit (WFP VAM) and the Harvard Humanitarian Initiative (HHI).

## HOW TO USE THIS GUIDE

The guide has been designed with ease of reference in mind: users do not have to read the document from beginning to end; they can simply refer to the subjects of interest for a given exercise. That said, we recommend that all users read the first three chapters before consulting subsequent chapters. The guide should be supplemented with specific reference materials on assessments – *Assessing Economic Security* (ICRC EcoSec, 2016) – and results-based management – *EcoSec Planning, Monitoring and Evaluation* (ICRC EcoSec 2016) – and with basic training in relevant software packages, such as MS Excel.

Many ICRC tools and guides have also been developed for data collection and analysis, including:

- EcoSec data-collection tools for assessment and monitoring
- the EcoSec sample calculator
- a reference framework for protecting personal data
- a visual identity guide
- a style guide.

These are referred to throughout the guide, and are available on the "Data and Analysis" page at the EcoSec Resource Centre on the ICRC intranet.

## TERMS

The following are some of the terms and definitions used in the guide.

- **Humanitarian work**: All-encompassing term for any type of work done by humanitarian personnel
- **Humanitarian programmes**: Programmes designed to assist or protect populations in need (may involve a variety of activities)
- **Humanitarian activities**: Activities that seek, directly, to assist or protect populations in need (may be part of a broad programme)
- **Data collection and analysis exercise**: Any project involving collection and analysis of data (assessments, monitoring and evaluating activities, etc.)

## ACKNOWLEDGEMENTS

This guide was developed from in-depth research into existing sources of guidance on data gathering and analysis, and on approaches and practices relevant to the ICRC's work. It integrates academic research with the experience of humanitarian personnel to balance sound practices with field realities.

The document makes use of reference materials originating in both international humanitarian and development organizations, to whom we are most grateful for their work in the field.

Special thanks to Patrick Vinck and Phuong Pham with the HHI, Rossella Bottone with the WFP VAM and Patrice Chataigner with the ACAPS project for carefully reviewing and making detailed comments on this document.

# CONTENTS

A guide for humanitarian work ..... 1

**INTRODUCTION ACQUIRING AND ANALYSING DATA ..... 3**
Generating Information ..... 4
About this guide ..... 6
How to use this guide ..... 6
Terms ..... 7
Acknowledgements ..... 7

**CONTENTS ..... 8**

**CHAPTER 1 APPROACHES TO ANALYSIS ..... 13**
Qualitative approaches ..... 14
Quantitative approaches ..... 14
Mixed-method approaches ..... 15
Participatory approaches ..... 15
Choosing an approach ..... 16
Analysis in humanitarian work ..... 17
The analysis team ..... 18
References ..... 20

**CHAPTER 2 DATA AND VARIABLES ..... 23**
Data ..... 26
Variables ..... 27
Indicators ..... 28
Units ..... 29
Data sources ..... 29
Data topics ..... 30
References ..... 31

**CHAPTER 3 ANALYSIS DESIGN ..... 33**
Questions and objectives ..... 34
Broad concepts to concrete data ..... 34
Assumptions and theories ..... 35
Population of interest ..... 36
Secondary-data review ..... 38
Analysis design ..... 39
Error and bias ..... 53
Data protection and ethics ..... 54
References ..... 58

**CHAPTER 4 PRIMARY-DATA COLLECTION ..... 61**
Communication and consent ..... 62
Primary-data-collection methods ..... 64
Level of structure and flexibility ..... 67
Interoperability ..... 68
Primary-data-collection tools ..... 70
Primary-data-collection mediums ..... 84
Putting it into practice ..... 87
References ..... 102

**CHAPTER 5 SAMPLING ..... 105**
Key concepts in sampling ..... 106
Sampling methods ..... 107
Sample size ..... 109
Probability sampling ..... 109
Non-probability sampling ..... 131
Analysis of sampled data ..... 135
Frequently asked questions ..... 139
Note ..... 141
References ..... 142

**CHAPTER 6 DATA TREATMENT ..... 145**
Data processing ..... 146
Quality control ..... 150
Data integrity ..... 153
References ..... 155

**CHAPTER 7 QUANTITATIVE ANALYSIS ..... 157**
Variables and statistics ..... 159
Descriptive statistics ..... 161
Inferential statistics ..... 171
Relationships ..... 173
Comparisons ..... 179
Trends ..... 186
Adding depth to the analysis ..... 191
Outliers ..... 193
Missing values and non-response ..... 193
Reporting statistics ..... 195
References ..... 196

**CHAPTER 8 QUALITATIVE ANALYSIS ..... 199**
Extraction and organization ..... 200
Triangulation ..... 208
Relationships and trends ..... 209
References ..... 216

**CHAPTER 9 COMBINING ANALYSES AND DRAWING CONCLUSIONS ..... 219**
Combining and reanalysing ..... 220
Plausibiliy and validity ..... 222
Interpretation ..... 224
Communication ..... 225
References ..... 226

**CHAPTER 10 VISUALIZATION ..... 229**
Why visualize? ..... 230
Developing visuals ..... 230
Visual design principles ..... 237
Establish a visual hierarchy ..... 238
Balance the level of detail ..... 238
Establish harmony ..... 238
Be accurate ..... 238
Ensure readability ..... 238
Give credit ..... 238
References ..... 239

**ANNEX I TERMS AND DEFINITIONS ..... 241**
A-E ..... 242
F-J ..... 244
K-O ..... 245
P-T ..... 246
U-Z ..... 249

**ANNEX II ANALYSIS DESIGN TOOLS ..... 251**
Conceptual frameworks ..... 252
Logical frameworks ..... 253
Analysis plan template ..... 253

**ANNEX III ADDITIONAL SAMPLING FORMULAS ..... 257**
Basic formula for means or totals ..... 258
Formula for comparison surveys using means or totals ..... 258

**ANNEX IV ANALYSIS AND VISUALS ..... 261**

# CHAPTER 1
# APPROACHES TO ANALYSIS

Assessing and monitoring economic security, as well as other data collection and analysis exercises in support of humanitarian work, may follow a purely qualitative or quantitative approach, or take a mixed-method approach that has both qualitative and quantitative elements. The choice of approach will be determined by the objectives of the exercise and by the informational requirements that should be developed in consultation with those knowledgable of the context or subject matter and historical evidence. Each *approach* will gather certain types of *data* and may employ a variety of *tools and methods* for data collection and analysis.

| THE FOLLOWING DEFINITIONS ARE USED IN THIS GUIDE. | |
|---|---|
| **Approach** | A comprehensive or all-inclusive term for a means of dealing with data or of collecting and analysing them |
| **Data** | Raw, unorganized facts or figures that need to be processed and analysed |
| **Method** | The manner in which data are collected and analysed (for data collection: direct observation, interviews with key informants, household surveys, etc.; for data analysis, quantitative analysis such as descriptive statistics) |
| **Tools** | Instruments that assist in data collection or analysis: questionnaires, mobile phones, etc. |

## QUALITATIVE APPROACHES

**Qualitative approaches** aim to explore and understand phenomena, and are based on the collection of data in their natural setting, from the insider's perspective, through observation and discussion. Qualitative data are mainly descriptive, and collected through such means as open-ended interviews, group discussions and direct observation.

**EXAMPLE**

The Norwegian Refugee Council (NRC) carried out an initial rapid assessment of internally displaced persons (IDPs) in three districts in Yemen in early 2015; its aim was to identify the immediate needs of IDPs in order to prepare a relevant and adequate emergency response. The assessment lasted three days, and employed focus group discussions and interviews with key informants. Indicators for the assessment were based on informants' views concerning key issues related to food, shelter, non-food items and water and sanitation.[2]

## QUANTITATIVE APPROACHES

**Quantitative approaches** seek to confirm specific hypotheses by quantifying data and information. They are usually based on structured surveys or measurements of specific variable (food price, body weight, etc.). Quantitative approaches may use quantitative or qualitative data, and methods of quantitative analysis that aim to quantify – or give numerical expression to – such things as the households that speak a given language and the people who use aid to buy food, or to generalize findings back to the general population of interest (e.g. statistical inference).

**EXAMPLE**

Between 2010 and 2012, ICRC field teams carried out nutritional screening at a number of prisons in a West African country, in order to learn more about the nutritional situation of inmates. The main objective was to monitor, for a certain duration, rates of global acute malnutrition (GAM) and severe acute malnutrition (SAM). The height and weight of each inmate was measured to calculate their body mass index (BMI) over a three-month period. GAM and SAM rates were calculated on the basis of these BMI variables.

2 NRC, April 2015.

## MIXED-METHOD APPROACHES

**Mixed-method approaches** combine quantitative and qualitative methods of analysis. They are usually employed to offset the inadequacy of one type of data and/or method of analysis by supplementing it with the other. The main characteristic of these approaches is pragmatism: they use the most practical means to generate information to the required level of confidence (either measured for statistically significant data or described based on triangulation, the experience of the analyst and a general feeling for the reliability and accuracy of the information).

**EXAMPLE**

In early 2014, the WFP, the Central Statistical Organization of the Government of Yemen (CSO) and the United Nations Children's Fund (UNICEF) jointly conducted the Comprehensive Food Security Survey in Yemen. The survey was designed to provide representative statistics on food security and nutrition, examine the underlying causes of food insecurity and malnutrition, and identify vulnerable areas. The aim was to acquire information on which to base decisions concerning programmes on chronic food insecurity and malnutrition. Various methods were used to collect primary data, such as structured household interviews, interviews with women of reproductive age, anthropometric measurements of children under five, community-level focus group discussions and interviews with traders.[3]

## PARTICIPATORY APPROACHES

**Participatory approaches** are not analytical; the word 'participatory' refers to the various methods for managing and conducting the collection and analysis of data. They are mentioned here because of their frequent use in the humanitarian field. In a participatory approach, responsibility – for decision-making and for defining, collecting and analysing pertinent data – is shared by the team leading the exercise, the participant(s) or community involved and any other parties interested. The team leading the exercise will exert some degree of authority, and play the role of mediator, but opinions will be shared openly.

Participatory approaches normally use a qualitative or mixed-method approach: 'participatory tools' for data collection and other qualitative methods and traditional statistics in data analysis. But these choices will be determined by the way in which the exercise in question is being carried out.

**EXAMPLE I**

In rural areas of Colombia affected by armed conflict, the ICRC does a participatory rural appraisal, which is led by a multidisciplinary team consisting of specialists in agriculture, economic security, weapon contamination and water-and-habitat. The aim of the appraisal is to identify the consequences – in humanitarian terms – of armed conflict and the general problems and needs of people affected. To that end, the community is divided into groups that focus on specific needs and work to identify solutions. The entire community then has several weeks to think about solutions to the various problems and propose pertinent humanitarian activities; afterwards, these proposals are conveyed to the ICRC for consideration. ICRC programmes are designed in line with these recommendations, and later implemented and monitored together with the community.

**EXAMPLE II**

The ICRC's community-based protection activities aim to strengthen the resilience of conflict-affected communities by reducing their exposure to threats and by modifying coping strategies that may be doing more harm than good. As part of the process, the ICRC facilitates workshops with conflict-affected communities that define the main threats to their security, the challenges they are facing and their means of coping with the situation. The community then analyses these data to identify both constructive and harmful strategies, and opportunities and potential solutions. The ICRC's role is to facilitate the analysis, because in this instance, it is the members of the community who are the experts. The main findings and recommendations are made by them.

3 WFP, CSO, UNICEF, 2014.

## CHOOSING AN APPROACH

The choice of approach will depend on the objective(s) of the exercise and on the time and resources available. Whenever possible, a mixed-method approach[4] should be chosen, as there is "no clear distinction between quantitative and qualitative methods, and it is more helpful to consider data collection and analysis as being located on a quantitative-qualitative continuum" (World Bank, 2000, p. 3).

The main difference between qualitative and quantitative approaches is that the former do not seek statistical significance, and thus any attempt to extrapolate their findings will be conjectural to some extent. Qualitative approaches are sometimes thought to be less rigorous than quantitative ones; however, they make up in explanatory power what they lack in statistical definitiveness. Quantitative approaches are often criticized for being inflexible and mechanical, and for their inability to match figures with explanations. The table below highlights the key characteristics of each approach. Wherever pertinent, a mixed-method approach will combine key features of each.

**Table 1** - **Key features of qualitative and quantitative approaches.** ***Adapted from WFP, February 2009; ACAPS, August 2013; and Creswell, 2003.***

| | QUALITATIVE APPROACH | QUANTITATIVE APPROACH |
|---|---|---|
| **Objectives and key questions** | ■ To explore and/or understand phenomena<br>■ Establish in-depth understanding<br>■ Open-ended questions, such as 'How?' and 'Why?' and 'What do I need to look for in more detail?' | ■ To seek precise answers or figures; to confirm assumption (hypothesis) with an identified level of confidence<br>■ Establish a general overview<br>■ Closed sequence of questions, such as 'What?' and 'How many?' |
| **Main features** | ■ Detailed and complete information with contextualization, interpretation and description<br>■ Opinions, explanations, perceptions, predictions, beliefs and desires of the people affected | ■ Precise measurements that often require further explanation<br>■ Objective, reliable and possibly verifiable data, if collected correctly<br>■ Suitable for generalization and further analysis (prediction, correlation, causation, etc.) |
| **Perspective** | ■ Looks at the whole context from within<br>■ Searches for patterns/trends<br>■ Lends itself to community participation | ■ Looks at specific aspects of the context from outside<br>■ Seeks to measure a specific element |
| **Data collection methods** | ■ Memoing, coding and categorization<br>■ Phenomenology<br>■ Ethnography<br>■ Case studies<br>■ Narrative research<br>■ Participatory research[5] | ■ Measurement<br>■ Structured observation<br>■ Surveys |

4 See designs for mixed-method approaches in Chapter 3.

5 These are some of the most widely used methods of participatory research in humanitarian and development studies: Participant Observer', Rapid Rural Appraisal for agricultural research, Participatory Rural Appraisal, Participatory Action Research, and Participatory Impact Assessments for monitoring and evaluation.

| | QUALITATIVE APPROACH | QUANTITATIVE APPROACH |
|---|---|---|
| **Data collection tools** | ■ Notepads, whiteboards, etc.<br>■ Flexible data collection guides with flexible sequence of questions<br>■ Direct observation or recording device such as camera, tape recorder, etc.<br>■ Participatory tools<br>■ The data collector plays a key role | ■ Standard questionnaire, form, etc. with fixed or random sequence of questions<br>■ Technological instruments such as scale, measuring tape, etc. |
| **Data format** | ■ Mostly qualitative data and drawings, timelines, diagrams, etc.<br>■ Geospatial data (more as a visual graphic of patterns, trends, etc.)<br>■ Data can be observed, but not measured<br>■ Mainly textual, but also categorical | ■ Quantitative and qualitative data<br>■ Geospatial data (for measurement)<br>■ Data can be counted or measured<br>■ Mainly numerical and categorical |
| **Analytical methods** | ■ Descriptive words or visuals<br>■ Systematic and iterative process of searching, categorizing and integrating data<br>■ Analysis and interpretation from the subject's perspective (participation)<br>■ Generalizing from a limited number of specific observations or experiences | ■ Descriptive statistics or spatial trends<br>■ Forecasting and making comparisons through statistics<br>■ Analysis and interpretation based on measurement<br>■ Generalizing with quantifiable precision and confidence (inferential statistics) |

## ANALYSIS IN HUMANITARIAN WORK

There are a number of types of analysis used in humanitarian work. Each may employ a qualitative, quantitative or mixed-method approach and can be done at a macro or micro level depending on the information requirements and the decisions that will be taken with the results. Some commonly used analyses in humanitarian work are listed below.

### EXPLORATORY DATA ANALYSIS

An **exploratory data analysis** takes a look through a dataset in an attempt to discover any key characteristics that may not be implicit in the analyst's original assumptions, hypotheses or theories. It can serve as a precursor to an exercise – and used to study data in a secondary-data analysis or information compiled in a desk review– or it can be an initial step in data analysis after all the data have been compiled. The objective of exploratory analysis is to guide the analyst in elaborating hypotheses, theories and/or assumptions, which can then be tested either by using the same data or by collecting fresh data and/or information.

### SITUATION ANALYSIS

A **situation analysis** examines the internal and external, and the direct and indirect, factors that may have some bearing on a situation. It aims to understand the relationship between these various factors, the circumstances that led to the current situation and prospects for the future. Within the context of the ICRC's work, such things as physical security, population movement, food security, and income would qualify as a 'situation'. Situation analysis may also include severity analysis.

### VULNERABILITY ANALYSIS

Vulnerability is the degree to which a person affected lets a given event cause harm. It is a person or population's sentivity multiplied by their lack of ability to cope or adapt. A **vulnerability analysis** normally starts by defining vulnerability in a given context (e.g. Vulnerable to what?) and then identifies specific criteria (called domains)[6] and associated indicators that can be used to measure a population's vulnerability to a certain shock or situation. Vulnerability analysis may be used together with severity analysis, risk analysis and prioritization of activities.

### SEVERITY ANALYSIS

A **severity analysis** builds on vulnerability analysis, but looks more deeply into the scale of a shock by adding intensity and exposure to the analysis. Intensity is the strength of a given shock and exposure is the scale on which people are affected (normally expressed in terms of number of people or geographic area).[7]

### RISK ANALYSIS

A risk analysis looks into the likelihood of an adverse impact should a shock (or threat) occur. As such, it is a projection of possibility or possibilities.

### SCENARIO-BUILDING AND FORECASTING

**Scenario-building** involves drawing up plausible scenarios – normally two or more – based on past and present evidence. It may use process diagrams and chains of plausibility. **Forecasting** involves making predictions on the basis of evidence, past and present. Forecasting is more specific than scenario building: it focuses on one particular theme.

### RESULTS MONITORING

**Results monitoring** normally uses a framework (known as the results monitoring or logical framework) that is set up to describe the expected and actual changes after a project or programme. This framework makes it possible to measure progress in terms of results-based indicators. Results monitoring may include data and analysis from both humanitarian activity and situation monitoring.

Predictive and causal analyses using quantitative methods are less common in humanitarian work because of the sheer difficulty of the methods of data collection required, which are both rigorous and repetitive as well as the number of intervening variables. They may be more commonly done using qualitative methods.

## THE ANALYSIS TEAM

Sound or effective analysis is a collaborative process – at least, a consultative one – and carried out by a team rather than an individual. Someone may act as a leader, but he or she will always look for the support of others.

The composition of an analysis team will vary: a large-scale needs assessment may involve numerous specialists in various areas (health, water-and-habitat, economic security, etc.) whereas an agricultural assessment may need only a team of agronomists.

An analytical team should be made up of people with complementary strengths. The table below is taken from the Data Expeditions section of School of Data's webpage.[8] It lists the 'characters' that take part in a 'data expedition'.[9]

6 See "Conceptualization" in Chapter 3: Analysis design for more information on concepts, domains, indicators and measures.

7 ACAPS, 2016.

8 School of Data is a network of data literacy professionals that provides training in data literacy (http://schoolofdata.org/).

9 Taken directly from the School of Data webpage (http://schoolofdata.org/data-expeditions/); the only changes are to the description of 'scout', which, unlike the School of Data table, mentions primary-data collection and interviews and the skills required to carry out these activities.

| | |
|---|---|
| **SCOUT** | Scouts hunt down data. They may or many not have to be technical specialists: that will depend on the difficulty of obtaining the data. If primary data has to be collected or interviews conducted, they may need to have strong interpersonal skills. |
| **STORYTELLER** | People who are good at finding interesting angles to explore and producing outputs that really speak to the intended audience. Storytellers are particularly key in defining the question and in pulling together the final mission reports at the end. |
| **ANALYST** | Analysts are the ones who crunch the data found by the scouts and test the hypotheses generated by the storytellers. |
| **DESIGNER** | Beautify the outputs and make sure the story really comes through the data. Note: a paper representation of how you would like to present your outputs is just as valid as a fully-fledged interactive graphic produced by a coder (sometimes it even serves as a precursor to an interactive). |
| **ENGINEERS (OPTIONAL)** | Put together the final outputs with help [*sic*] of the group. Engineers are usually somewhere on the technical spectrum but they don't have to necessarily be coders, [*sic*] we've seen loads of great outputs from people who know how to use off-the-shelf tools. |

These 'characters' are representative of the skills needed for a data and analysis exercise in humanitarian work. In certain cases, a single person may have more than one skills; in others, a particularly skill may not be required.

**ROAD TO A SUCCESSFUL ANALYSIS: KEY POINTS TO REMEMBER WHEN YOU PLAN YOUR EXERCISE:**[10]

1. **Qualitative and quantitative approaches are not in competition with one another**. Some people assume that quantitative approaches are more reliable than qualitative, or that qualitative data is easier to collect than quantitative. This is not the case. In fact, effective collection and analysis of qualitative data takes experience and skill; and when quantitative data are not collected or processed correctly, they may be useless. It is the quality of the data and the quality of analysis that is most important; the types of data and analysis required are defined by the decisions that will be based on them.

2. The **quality of the analysis is determined by the quality of the data**. While data-collection methods and data treatment never make headlines like analysis and graphics, they form the basis of the entire process.

3. **Data rarely speak for themselves**. They have to be contextualized and interpreted before they can be used for decision-making, and that is done by people, not computers.

4. There are **no one-size-fits-all options in analysis**. In order to be certain of selecting the best analytical model and methods, each situation deserves proper planning and thought.

5. **Sound and effective analysis is a consultative process**. It is best done in a group setting, and involves people who have experience in the context and a variety of relevant expertise.

6. **Information needs and feasibility should be constantly reviewed.** These needs will evolve with the situation and/or as certain pieces of information illuminate previously grey areas. The feasibility of producing information can also change, because of the time-sensitive nature of certain kinds of information, an initial analysis that reveals the inadequacy of data-collection methods, resource issues, access issues, etc.

10 Based on ACAPS, August 2013.

# REFERENCES

ACAPS, *Compared to What? Analytical Thinking and Needs Assessment*, August 2013. Available at: http://www.acaps.org/img/documents/c-160806-tb-compared-to-what-final.pdf.

ACAPS, *How Sure are You? Judging Quality and Usability of Data Collected during Rapid Needs Assessments*, August 2013. Available at: http://www.acaps.org/img/documents/h-130827-tb-how-sure-are-you-final.pdf.

ACAPS, *Qualitative and Quantitative Research Techniques for Humanitarian Needs Assessment: An Introductory Brief*, May 2012. Available at: http://www.acaps.org/img/documents/q-qualitative-and-quantitative-research.pdf.

ACAPS, Severity Measures in Humanitarian Needs Assessments: Purpose, Measurement, Integration, August 2016. Available at: https://www.acaps.org/sites/acaps/files/resources/files/acaps_technical_note_severity_measures_aug_2016_0.pdf

Consortium of International Agricultural Research Centers, International Fund for Agricultural Development, International Institute of Tropical Agriculture, *Rapid Rural Appraisal Report of Northern Uganda February-March 2014*, July 2014. Available at: http://reliefweb.int/report/uganda/rapid-rural-appraisal-report-northern-uganda-february-march-2014.

Creswell, John W. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 2003. Sage Publications, Inc. Third edition. Available from: http://isites.harvard.edu/fs/docs/icb.topic1334586.files/2003_Creswell_A%20Framework%20for%20Design.pdf.

NRC, *Initial Rapid Assessment: Abyan and Hajjah IDPs (Yemen)*, April 2015. Available at: http://reliefweb.int/report/yemen/initial-rapid-assessment-abyan-and-hajjah-idps-yemen.

School of Data website, 'Guide for Guides': http://schoolofdata.org/data-expeditions/guide-for-guides/.

Voluntary Services Overseas, *Facilitator Guide to Participatory Approaches*, 2009. Available at: http://community.eldis.org/.59c6ec19/.

WFP, *Emergency Food Security Assessment (EFSAs) Technical Guidance Sheet No. 8: Introduction to Qualitative Data and Methods for Collection and Analysis in Food Security Assessments*, February 2009.

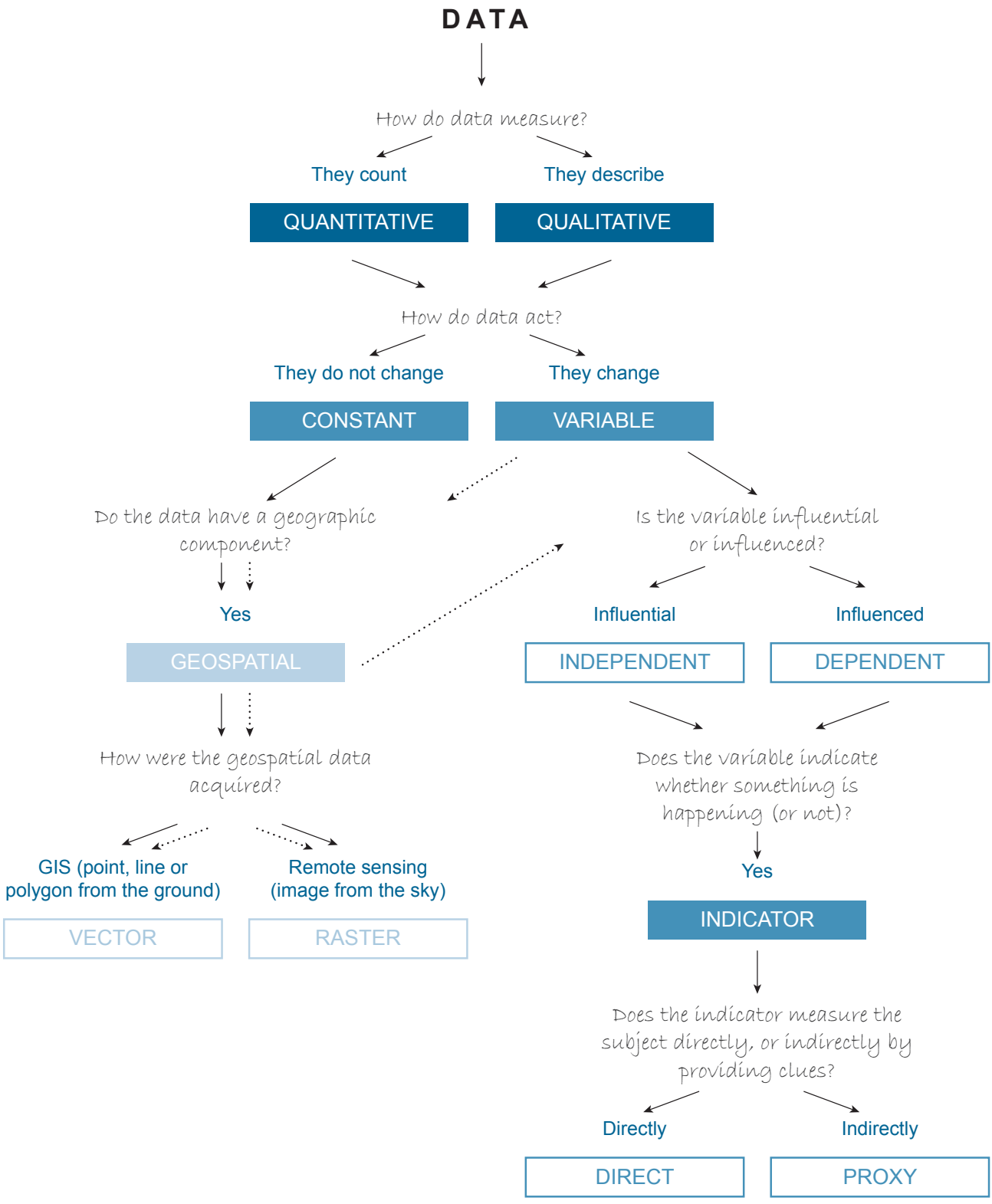
WFP, CSO, UNICEF. *Yemen: Comprehensive Food Security Survey*, November 2014. Available at: http://www.wfp.org/content/yemen-comprehensive-food-security-survey-november-2014.

World Bank, *Integrating Quantitative and Qualitative Research in Development Projects*, June 2000. Available at: http://documents.worldbank.org/curated/en/2000/06/2095601/integrating-quantitative-qualitative-research-development-projects.

# CHAPTER 2
# DATA AND VARIABLES

**Data** are raw, unorganized facts or figures that have to be processed and analysed. **Variables** are a type of data that are subject to change; they are the basis for most analyses carried out to understand patterns, relationships and trends. Data and variables can take various forms: simple and seemingly random or statistical and seemingly complicated. Whichever the case, data and variables are the basis for analysis, but they are of no use until they are processed, analysed and eventually turned into information.

Figure 2 is a simple flow chart that shows a number of different ways of categorizing data. Data can have various features (quantitative variable, proxy indicator, etc.). It is important to understand these features fully; otherwise it will not be possible to choose and carry out the proper type of analysis.




*How do 'data', 'information' and 'evidence' differ from each other?*

**Data** are raw, unorganized facts that have to be processed. **Information** is processed data that are organized and presented in such a way that they can be used. **Evidence** is information that demonstrates the truth or falsehood of something.

# DATA

## QUALITATIVE DATA

**Qualitative data** are descriptive, and range from structured categories to perceptions, opinions, intentions and observations. They can be collected in various ways, through structured surveys or less structured observation, discussion and open-ended questions.

**EXAMPLE**

An EcoSec household economy assessment in northern Mali in 2014 used a household survey to collect data on food consumption and income-earning opportunities. As part of the survey, data was collected on the households' main income-generating activities – agriculture, livestock farming, fishing, salaried jobs, daily labour, business, etc. – for use in cross-tabulations. The data contained in the categorical list of income-generating activities, which was prepared afterwards, is qualitative.

## QUANTITATIVE DATA

**Quantitative data** are numerical, and expressed as statistics, rates, proportions, etc.

**EXAMPLE**

The EcoSec team in Somalia collects data on the market prices of key commodities purchased by poor households at six markets in Puntland and eight markets in the southern and central regions of the country. Commodity prices are quantitative data.

## GEOSPATIAL DATA

**Geospatial data** have a geographic or spatial component. The two main types of geospatial data are geographic information systems (GIS) data and remote sensing data.

**GIS data** are geographically referenced data that can be used in GIS for spatial analysis. GIS integrates hardware, software and data to collect, manage, analyse and display all forms of geographically referenced information.[11] GIS data are developed and shared in the form of layers that can be stacked one on top of another to create maps and analyse their features in relation to their location (geospatial analysis). These layers can be polygons, lines or points, or simple text, shapefiles (.shp), geodatabase files (.gdb) Google Earth files (.kml/.kmz), web map or feature service (wms/wfs), or take some other form.

**EXAMPLE**

As part of a review of an ICRC access to employment programme in Colombia; they wanted, in this connection, to map the total number of beneficiaries by location (in this case, cities in Colombia). GPS coordinates were collected for each site together with shapefiles (.shp) of administrative and international boundaries, with a view to developing a map of the country showing the programme sites.[12] The GPS coordinates and shapefiles are GIS data.

**Remote sensing is** a technique used to acquire information from a distance. Remotely sensed data are normally extracted from satellite images (collected from sensors or cameras on satellites), aerial images (collected from manned or unmanned aircraft) or on-the-ground images, which are then processed, classified, analysed and interpreted. Data can represent objects or phenomena on the earth's surface, in the atmosphere or in the oceans. Remote-sensing analytical techniques depend on the satellite sensor used to collect the images (in the case of satellite images), the resolution of the image (in the case of both satellite and aerial images) and the desired analytical output.

11 Esri website, accessed in April 2015.

12 ICRC Colombia, May 2014.

**EXAMPLE**
Agricultural land in Gaza was severely damaged during the seven-week war in early 2014. In order to gain a better understanding of the extent of the damage, UNOSAT compared high-resolution satellite images from before and after the war, and identified changes in agricultural areas between those dates. The images used are remote-sensing data and the analytical techniques employed are remote sensing.[13]

## VARIABLES

A **variable** is any piece of data that can take on different values, and is subject to change. It is the opposite of a **constant**, a piece of data that does not change. A variable can be a piece of quantitative or qualitative data. Various types of variable are used in humanitarian work, the most common being independent, dependent and control variables.

*What is the difference between a 'variable' and a 'case'?*
A case is one record or one response: one individual, one household, one key informant, etc. Data for one variable can include one or many cases.

### TYPES OF VARIABLES

#### INDEPENDENT VARIABLE

An **independent variable** is one that is not influenced by other measurable variables: for instance, factors (something that influences other variables) or predictors (something that can be used to predict the likely value of something else).

#### DEPENDENT VARIABLE

A **dependent variable** is one that is influenced or 'depends' on another measurable variable. It can also be referred to as a 'response' variable.

**EXAMPLE**
An economic security assessment aims to identify food- and income-insecure households in a community of displaced people. The assessment measures current levels of food consumption and current coverage of essential expenditures. It considers size of household, time since arrival and access to income-generating activities as possible determinants of poor food consumption and inability to cover essential expenditures. In this case, the independent variables are household size, time since arrival and access to income-generating activities; the dependent variables are levels of food consumption and current coverage of essential expenditures.

### MEASUREMENT SCALE

Variables can also be categorized by what they measure. Understanding what the variables in question measure is important; it is only then that the appropriate methods of quantitative analysis can be chosen (see Chapter 7 Quantitative analysis). Measurement scales can be nominal, ordinal, interval or ratio.

13 UNOSAT, September 2014.

**Table 2** - Variable types by what they measure

| VARIABLE | DEFINITION | EXAMPLE |
|---|---|---|
| **NOMINAL** | Attributes are uniquely 'named' with no implied order. Nominal data with only two categories are dichotomous. | ▪ Phone number<br>▪ Red, blue, green<br>▪ Resident, Returnee, IDP, Refugee<br>▪ Man/Woman<br>▪ Yes/No<br>▪ Included/Excluded |
| **ORDINAL** | Attributes can be ranked in a meaningful order with categories | ▪ High, medium, low<br>▪ 0-5, 6-10, 11-15, >15<br>▪ Likert scales |
| **INTERVAL** | Can only have a finite number of real values (whole numbers) and the distance between each number is meaningful but arbitrary (e.g. the distance between 1 and 2 may not be the same as between 2 and 3). Intervals do not have a true 0. | ▪ Temperature<br>▪ Time of the day<br>▪ Oedema |
| **RATIO** | Can have an infinite number of real values, and the value of 0 is meaningful. The distance between two numbers is the same (e.g. one person can be twice as tall as the other, 100 and 200 dollars is the same difference as 200 and 300 dollars, etc.). | ▪ Age<br>▪ Household size<br>▪ Height/weight<br>▪ Income/expenditure<br>▪ # of meals/day<br>▪ Distance<br>▪ Dependency ratio<br>▪ % of expenditure on food |

## INDICATORS

An **indicator** is a variable that indicates something, such as a change or a trend. Indicators are compiled from data, and measured and interpreted through comparison with standard or context-specific baselines, thresholds or target values. Many indicators – but sometimes only one – may be chosen for an exercise.

Indicators can be direct or indirect. **Direct indicators** provide a direct measure of something, but **proxy indicators** provide indirect information. **Proxy indicators** are useful in triangulating or completing data collected to fulfil direct indicators; and in certain cases, while less precise, they may be more reliable if the data are easier to collect than direct indicators.

**EXAMPLE**

The proportion of people's income that is obtained from regular and sustainable sources can be a direct indicator of their access to income. Their use of harmful coping strategies to cover basic needs can be a proxy indicator.

# UNITS

## UNIT OF ANALYSIS

A **unit of analysis**, sometimes also referred to as 'the reporting domain', is defined here as the level at which conclusions will be drawn and information reported. For example, if the object of a report is to draw conclusions about the population of Sierra Leone, the unit of analysis is the population of Sierra Leone. There may be one or many units of analysis, and they may be defined by various distinct characteristics, such as:

- geographic location
- population group
- livelihood zone.

## UNIT OF OBSERVATION

A **unit of observation** is level at which data are collected (i.e. observed). For example, for a household survey, the unit of observation is the household. For a key informant interview, the level of observation may be a particular community or neighbourhood, depending on who the key informant is asked to generalize their answers to. The unit of analysis and the unit of observation must be defined in the analytical design to ensure that relevant data are collected, and appropriate sampling methods used for that purpose, so that data may be reported back to the desired reporting level.

## UNIT OF MEASUREMENT

The **unit of measurement** is that which is used to quantify a variable: kilometres, dollars, litres, etc.

# DATA SOURCES

### PRIMARY DATA

**Primary data** are data collected by the person or persons who will make use of them, such as data collected during a survey or an experiment or by witnessing something.

### SECONDARY DATA

**Secondary data** are data collected by someone other than the user of the data:[14] census data, data from national or other international organizations, data collected separately by a National Red Cross or Red Crescent Society, historical accounts, media reports, etc.

### CROWDSOURCING AND CROWDSEEDING

**Crowdsourcing** is a method of gathering data and/or information by soliciting contributions from a large group of people; it may be confined to a specific community (eyewitnesses or victims during a crisis, humanitarian workers, technical specialists, etc.) or open to the general public. Crowdsourcing techniques can be used on both a larger (e.g. open to anyone) and a smaller scale (e.g. small community group).

**EXAMPLE**

After the earthquake in Haiti in January 2010, humanitarian organizations had an exhaustive list of health facilities, but no data on where these facilities were. At the request of the humanitarian community, CrisisMappers (http://crisismappers.net/) had 'crowdsourced' volunteers in distant locations use high-resolution satellite imagery to verify the location of 102 out of 105 facilities (e.g. their GPS coordinates) in less than two days.[15]

14 Wikipedia, Wikipedia's 'Secondary data' entry, accessed in February 2015.

15 UN Foundation, 2011.

**Crowdseeding** differs from crowdsourcing in this way: providers of information are pre-identified and often trained in gathering and sharing information; this makes it possible to have more control over information providers than in crowdsourcing.

**EXAMPLE**
Voix de Kivus was a project, led by academics from Columbia University, to study the effectiveness of technical tools for collecting representative data in areas with vulnerable populations. A random sample of villages in the eastern Democratic Republic of the Congo were given mobile phones and credit for an 18-month period and asked to report on conflict events as they happened. The events data would be used later to analyse the effects of development projects in the region.[16] The events data were crowdseeding data.

### BIG DATA

**Big data** is a general term used to describe techniques for analysing large data sets from traditional and digital sources that are too large and complex to process and analyse with standard data processing and database management applications. As a result, advanced computing and statistical techniques have been developed to capture, process, understand and analyse these data.

**EXAMPLE**
As part of its Data for Development (D4D) challenge, the mobile network operator Orange shared anonymized phone calls and SMS exchanges between five million mobile phone users between December 2011 and April 2012 with academics; the aim was to foster research into human behaviour and to study in depth how big data, such as phone and SMS records, and their associated attributes could be used in development planning.[17]

## DATA TOPICS

Most data used in humanitarian contexts are related to demographics, geography and specific issues such as health, nutrition and income.

### DEMOGRAPHIC DATA

**Demographic data** are data on a given population: population statistics, gender, age, ethnicity, etc. Governments can collect demographic data during a census and share them with others through their statistics offices (a source of secondary data for the ICRC). Demographic data are also collected by the ICRC during field visits to fill gaps in missing or outdated census data, and as a means to disaggregate data collected during assessment, beneficiary registration, monitoring and evaluation activities.

Personal data are, "any information relating to an identified or identifiable natural person. This may include an identifier such as a name, audiovisual material, an identification number, location data, online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of a person. This also includes data identifying or capable of identifying human remains" (ICRC, 2015).

### GEOGRAPHIC DATA

**Geographic data** are geographically referenced data (such as GIS data or satellite imagery) on specific points of interest (villages, towns, cities, mountains, etc.), political boundaries, roads, lakes, rivers, oceans, land cover, etc. The ICRC usually acquires baseline geographic data from secondary sources, such as government mapping agencies or statistics offices, or academic/research institutions.

16 Van der Windt and Macartan, February 2012.
17 Orange Group webpage, Accessed in March 2015.

**Baseline demographic data** are often linked to baseline geographic data. For example, population statistics are normally reported by administrative units (provinces, prefectures, states, etc.) or major towns/villages. Using a standard set of baseline data across all data-collection exercises in a given area ensures interoperability between datasets, is crucial for spatial analysis to recitify tabular data with GIS systems, proper data management and, in turn, for any statistics that are reported back to a given geographic location. For example, if data are collected from a number of villages located in five different regions, and the analyst would like to draw conclusions for each region, he or she needs to start will a list of all the villages, and of the regions they fall under, to be able to draw a sample and make conclusions by region.

### THEMATIC DATA

**Thematic data** are data on a specific social issue of concern: population movements, incidents of violence, health, nutrition, food consumption, etc. The ICRC uses both primary and secondary sources to collect thematic data; it may also use crowdsourcing, crowdseeding and big data where pertinent and reliable. When they are on a social element (humanitarian needs, future intentions, access to income, etc.), thematic data are even more subject to error and bias; extra care should be taken in collecting thematic data, regardless of whether primary or secondary sources are used.

## REFERENCES

Esri website, 'Understanding GIS: What is GIS?': http://www.esri.com/what-is-gis. Accessed in April 2015.

ICRC Colombia, "EcoSec Executive Brief: Colombia - Access to Employment Programme 2013", May 2014.

ICRC, "Reference Framework on Personal Data Protection: A Handbook", December 2015. Internal document.

Orange Group webpage, 'Data for Development (D4D)': http://www.d4d.orange.com/en/home.

UN Foundation, *Disaster Relief 2.0: The future of Information Sharing in Humanitarian Emergencies*, 2011. Available at: http://www.unfoundation.org/news-and-media/publications-and-speeches/disaster-relief-2-report.html.

UNOSAT, *Damage to Agricultural Areas and Greenhouses, Gaza Strip – Occupied Palestinian Territory*, September 2014. Available at: http://www.unitar.org/impact-2014-conflict-gaza-strip-unosat-satellite-derived-geospatial-analysis.

Van der Windt, P. and Macartan, H., "*Voix des Kivus: Reflections on a crowdseeding approach to conflict event data gathering",* February 2012. Available at: http://petervanderwindt.com/writing/.

# CHAPTER 3
# ANALYSIS DESIGN

The design of an analysis is of direct consequence for the final results. Therefore, the time and effort required for thinking, designing and planning before collecting, analysing and reporting should not be underestimated.

## QUESTIONS AND OBJECTIVES

An exercise should always start with an objective that may be simultaneously or subsequently followed by a list should be drawn up of key questions that have to be answered: What happened? Was anyone injured? Where are they? Do they need help? What are their needs?

In the context of analysis, an **objective** is a statement that outlines the results expected from an exercise; it identifies, clearly, what needs to be understood. An exercise may have many objectives, each serving as a guiding point for information and data needs, collection methods and analysis.

**EXAMPLE**

The following are lists of key questions and objectives for a review, conducted in 2013, of an 'access to employment' programme for urban IDPs in Colombia.

**QUESTIONS**

- What is the current status of the activities in the programme?
- Did the programme do what it was designed to do?
- Does the ICRC need to continue to assist the beneficiaries of this programme? If yes, how?
- How could this type of programme be improved?

**OBJECTIVES**

**To draw conclusions and formulate recommendations on:**

- the evolution of programme activities
- the achievements of the programme
- adjustments required to improve implementation, beneficiary selection and impact

Objectives should be:

- **specific** – indicating what (people, processes, etc.) need to be analysed, where, on what scale and in what depth of detail;
- **relevant** – confined to what needs to be understood in order to take a decision;
- **comprehensive** – include all that needs to be understood in order to take that decision with an acceptable level of confidence; and
- **realistic** – knowing what is feasible, given the information available or collectable and the possible consequences of the conclusions reached by the analysis.

Humanitarian work takes places in rapidly evolving contexts that are also complex and challenging. The subjects of interest are often influenced by many factors, measurable and immeasurable. Attempting to understand every variable may create more questions or, even worse, deepen misunderstanding. It is important that the objectives focus on what is realistic, and that they be based on the decisions that need to be made with the information that is generated.

## BROAD CONCEPTS TO CONCRETE DATA

How do we move from broad questions to identifying the concrete data that we need to collect for analysis? For example, let us say that one of the key questions that a study seeks to answer is this: *Who in the community is most vulnerable?* How should the team proceed? In the social sciences, the first steps involve 'conceptualization' and 'operationalization'.

### CONCEPTUALIZATION

**Conceptualization** is the process of assigning meaning to the concepts contained in the objectives of the study. A **concept** is a word or phrase that suggests a meaning but not a

specific object.[18] For example, in the hypothetical study mentioned above, 'vulnerable' is a **concept**. But, w*hat does 'vulnerable' mean?* That has to be agreed upon before 'vulnerability' can be measured (through data collection and analysis).

Say that in a given context, it is agreed that 'vulnerable' will be defined by the following criteria: head-of-household marital status, access to income, living conditions and neighbourhood crime. These criteria are sub-groups of a concept and are referred to as **dimensions**. Thus, a concept can have many dimensions. In a different context, however, vulnerability might be defined by proximity to the front lines of a war. That concept has only one dimension.

Conceptual frameworks and models are analytical tools that are used to guide the process of conceptualization. See the section on **Analytical design tools** later in this chapter.

### OPERATIONALIZATION

Now that we have defined vulnerability, how do we *measure* it? In the social sciences, this process of measurement is called **operationalization**. Sometimes a concept or dimension can be directly observed or measured (head-of-household marital status, households living close to the front lines of a war, etc.), but measurement can also be less straightforward than that. For example, how should living conditions, if that is one dimension, be measured? Should we do so by observing general conditions during a house visit, measuring space crowding, examining the sources of water or gauging the level of sanitation? Or should we use some other means?

Indicators are used to define how concepts and dimensions will be measured. An **indicator** is a variable that indicates something, such as a change or a trend. Crowding space could be an indicator of living conditions, because it is measurable (for instance, by dividing the number of people living in a shared space by the size of the space). The specific **measures** used to feed an indicator are the most concrete data.


The differences between dimensions, indicators and measures are not clear-cut; in some cases, they may be the same (e.g. number of members in a household is both an indicator and a measure). Importance should be placed on the fact that abstract ideas in a study need to be translated to concrete measures through a process of defining what the concepts are at the conceptual level and what the measures will be at the operational level.

## ASSUMPTIONS AND THEORIES

Once the objectives have been stated clearly, all assumptions and theories pertinent to the exercise should be thought through and, if necessary, set down so that they can be taken into account when determining information needs and methods.

### ASSUMPTIONS

An **assumption** is anything that is accepted, without proof, as true or inevitable.[19] In some situations, there may be so much certainty about particular kinds of information as to preclude the necessity of looking further in those areas; decisions of this sort must be taken carefully – by people who are knowledgeable about the context – as they will significantly affect the process of determining what the information needs are.

18 Saldaña, 2015.

19 Definition taken from Google, accessed in April 2015.

**EXAMPLE**
Based on previous data from the region, a team assumes that the rate of prevalence of global acute malnutrition among children under five is 15%.

Unlike hypotheses in scientific research, assumptions do not have to be tested for truth in an exercise; they may, however, be disproved during an exercise.

### THEORIES

**Theories** differ from assumptions in that they are based on evidence or observation. Theories can be used to make predictions on the value of a variable, relationships among and between variables, trends, likely outcomes, etc. As with assumptions, theories can help to prioritize information needs or direct analyses; they must, however, be made carefully, and by people knowledgeable about the context and the situation.

**EXAMPLE**
Basic price theory states that changes in market prices are determined by changes in supply and/or demand: when supply goes up, prices goes down; and when demand rises, prices fall.

## POPULATION OF INTEREST

The **population of interest** is the group(s) of people on whom one would like to focus the study – and from whom, potentially, generalize its findings.

**EXAMPLE**
A research team would like to learn more about access to income-generating activities for recent Lebanese returnees from Syria. Lebanese returnees in Lebanon are the population of interest whom the researchers would like to study and draw conclusions from.

### POPULATION CHARACTERISTICS

The characteristics of a population of interest can be demographic, economic or social (who they are) or geographic (where they are). These characteristics may be identified through field knowledge/experience, desk review of secondary information or baseline data.

The characteristics that could influence the variables being measured in a study should be taken into account while designing an analysis.: that. In sampling, it may indicate a need for stratification, for example. In a questionnaire, it may indicate the need to collect certain characteristics of a household that can be used for disaggregation during analysis.

**EXAMPLE**
Access to income for small-business owners on the front line of a conflict, which people are fleeing, may be affected differently from that for small-business owners in locations receiving displaced people from the front lines. The analysis may be designed to stratify the sample by business owners on front lines and business owners in host communities.

The population of interest can be defined on one or many levels, depending on the focus and complexity of the study and the homogeneity or heterogeneity of the population in question. Categories should be detailed enough for each category to be homogeneous within and heterogeneous to the other categories. However, they do not need to be so detailed that they have no predicted effect on the outcome of the study or the information requirements. For example, let us imagine that a study is being conducted on the subject of access to clean water in a rural community of animal herders and farmers; and that it is generally accepted that both groups have the same degree of access; as a result, the study may not have to undertake a detailed comparison of animal herders with farmers in this

regard but would just analyse the rural community as a whole (i.e. no need for stratification or separate samples).

## POPULATION MATRIX

A population matrix can assist in outlining population characteristics together, by creating categories and defining what level of categorization (if any) is required for the study.

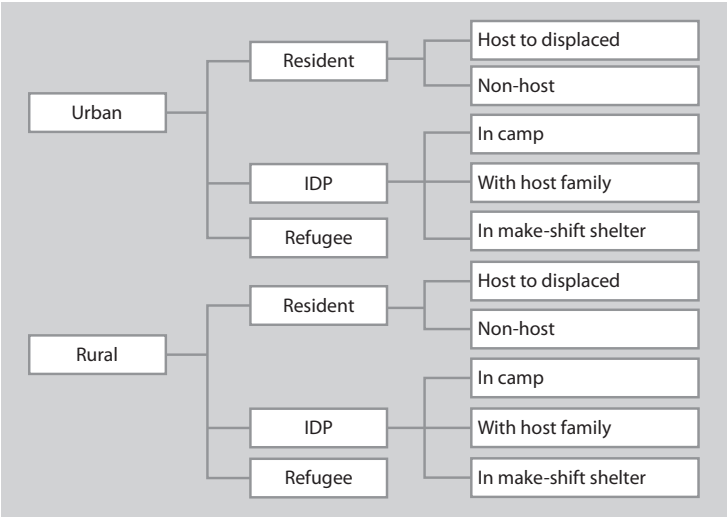



**Figure 3 - A population-of-interest matrix**

### BEST PRACTICES IN DEFINING POPULATION CHARACTERISTICS

- **Definitions should be formulated clearly** so that they are understood by everyone involved in the exercise: from the data collector to the analyst to the reporting officer. Teams should work together to draft definitions, and discuss concrete examples of each population group before moving forward with an exercise.
- **Existing typologies and definitions should be considered first,** in case there are legal meanings (i.e. displaced populations) and to avoid re-inventing the wheel (by not taking into account analytical work that has already been discussed and tested). If existing typologies and definitions are used, then results may be compared with secondary data.

## ONE-GROUP VERSUS TWO-GROUP STUDY

A **one-group study** examines one group individually, with the objective of conducting an intra-group analysis.

A **multi-group study** examines several groups, with the objective, possibly, of comparing them (inter-group comparison) and of doing intra-group analyses of each. When the same variables are measured for each group, using the same methods, this is often called a 'paired design'.

## CROSS-SECTIONAL VERSUS LONGITUDINAL STUDY

A **cross-sectional study** analyses data on a variable for one given period of time. A **longitudinal study** studies the same variable at repeated intervals over time.

| | ONE-GROUP | MULTI-GROUP |
|---|---|---|
| **CROSS-SECTIONAL** | Residents in June | Residents and IDPs in June |
| **LONGITUDINAL** | Residents in June, July, August and September | Residents and IDPs in June, July, August and September |

## SECONDARY-DATA REVIEW

A desk review of secondary data and information should be completed during the design phase, in order to help feed into background and baseline information as well as contextual information that can help to design the analysis, formulate assumptions and theories, and identify appropriate analytical approaches, data sources and design tools. Collection and collation of secondary data may as well begin during this phase, in order to identify if primary-data collection is even necessary.

Any of the following can be regarded as secondary data:

- Baseline or background data, or comparative data (e.g. population census)
- Contextual information (e.g. updated livelihood study)
- Lessons learnt from previous data-collection exercises
- Relevant data sources
- Relevant data-collection tools

In some cases, it may not be necessary to collect data from primary sources: sufficient data and information may be available already, and may only have to be verified.

**SECONDARY DATA AND INFORMATION WILL HELP TO:**

| | |
|---|---|
| ■ Determine if primary-data collection is necessary | If enough secondary information – in terms of quality and scale – is available, it may not be necessary to collect primary data; or the primary-data collection part of the exercise may be confined to triangulating and/or filling gaps in secondary data. |
| ■ Define the criteria to use | Previous exercises can provide data on values before a crisis or during periods of normality, or on other demographic and social indicators (population estimates, livelihood zones, etc.), |
| ■ Identify any relevant assumptions, theories or models | Reliable secondary data can illuminate specific areas of interest. For example, a previous exercise may have shown that most households buy food at the market during the dry season. This information can then be assumed as given and taken into account when designing the exercise. |
| ■ Set up baselines | Secondary data can be used as a baseline for comparison or for designing a sample. |

### BASELINE DATA

Baseline data are any data that can be used as a basis for comparison. They may include population figures, population characteristics (age, gender, displaced or non-displaced people, etc.), population density/infrastructure (e.g. urban or rural), geography (coastal, mountainous, plains, riverine, etc.) and economy/ecology (livelihood zones, natural resources, etc.), impact of disaster (homes damaged, destroyed, etc.), pre-existing vulnerabilities (poverty, malnutrition, access to basic services, etc.) and/or administrative units (provinces, states, districts, etc.).

In studies that require the drawing of a sample from a larger population, it is necessary to disaggregate population data at the relevant level and to estimate proportions for a given indicator (e.g. expected percentage of population without access to income). In evolving emergencies or in countries with outdated census figures, these data are often available only in the form of estimates, which should be subjected to critical study before use; if these estimates are used, the analysis should take into consideration any bias or error they may contain.

The best sources of these data should be identified in-country. Primary data (data collected by the ICRC) and official secondary data (data collected by others) from local authorities and established institutions should be preferred to unofficial secondary data or projections.

### GIS DATA

GIS can be an integral part of a study. In sampling, it is often used to identify the sampling frame and for site selection. It can provide information about a specific location (e.g. all towns within 100 km of the epicentre of an earthquake), political entity (e.g. 54 communes in 12 provinces in 2 regions), geographical feature (coastlines, mountains, plains, rivers, etc.) or phenomena that are directly related to geographical location (livelihood zones, migration routes, urban vs rural areas, etc.). Most people are not trained in GIS; non-GIS specialists can, however, turn to various resources, such as the ICRC Geoportal (https://ext.icrc.org/geoportal) and Google Earth (http://earth.google.com). If more support is required, contact the ICRC GIS Department (GISupport@icrc.org).

## ANALYSIS DESIGN

In the context of this guide, 'analysis design' refers to the overall design, including any models or frameworks and the way in which data will be collected and then combined and analysed. Overall, the design follows the analytical approach, however goes into more detail of *how* that approach will be carried out.

### DESIGN BY ANALYSIS APPROACH

There are a number of different designs and combinations of designs for each analytical approach (see Chapter 1: Analytical approaches). A few of the most commonly used are listed below.

#### QUALITATIVE DESIGNS

Qualitative designs following a qualitative approach seek to study people and things in their natural setting, and to understand or interpret phenomena from the perspective of the subject (people affected, witnesses, local authorities, etc.).[20] Primary data are normally collected through observation, discussion and semi-structured interviews, and may include the use of photography and audio/video recordings. Qualitative designs in humanitarian work normally make use of many sources of data, which are triangulated or combined for analysis. It is a back-and-forth iterative process and data collection, data processing and data analysis often overlap or occur simultaneously. Follow-up and additional data collection are common because insights or gaps in information emerge or are revealed by the data initially collected.

'Qualitative design' is not an established or fixed category in humanitarian work, and many designs that are qualitative may not even be labelled as such. These designs usually take the form of rapid assessments, in-depth assessments, case studies and narratives; and they often make use of elements of formally documented and researched ethnography, phenomenology, case-study research and narrative research.

**Ethnography** is the long-term investigation of a group of people, their customs and culture, and usually entails immersion and participation in the group. The **Rapid Assessment Process** (RAP), now referred to as the **Rapid Qualitative Inquiry (RQI),**[21] is a documented qualitative design developed by James Beebe that aims to collect and analyse data from the point of view of the subject (people affected, witnesses, local authorities, etc.) in a very short period of time. RAPs/RQIs are **ethnographic**; they differ from traditional ethnographies in that they are generally team-based, in order to maximize the amount of information that can be generated in a shorter period of time.

20 Beebe, 2014.

21 Variations of RAP include Rapid Assessment, Rapid Appraisal and Rapid Rural Appraisal.

When used in humanitarian assessments, qualitative designs usually also involve inquiring into the impact of a shock on a given population; this entails employing certain methods of **phenomenological research,** in which the researcher identifies the "essence of human experiences concerning a phenomenon described by participants in a study (Creswell, 2003, p. 15)".

**Case-study research looks** at a single case or explores similar topics (living conditions of urban refugees in two different cities, use of mobile technology in emergencies in four different types of emergency, conditions of a prisoner in two different facilities, etc.) across several cases. Finally, some qualitative designs employed in humanitarian work make use of certain methods of **narrative research,** which looks at the lives of individuals as told through their own stories; these stories are collected through field notes, transcripts of interviews, audio and video recordings, etc.

In a number of contexts where the ICRC works, there is a tendency to use unstructured ad hoc qualitative methods of data collection and analysis. This approach is used mainly to explore a given situation and to fill in gaps in information; it is also used when there isn't enough time to think about the design. However, both qualitative and mixed-method approaches should be guided by an objective and carefully designed; they should be reviewed and revised as necessary throughout the exercise.

## QUANTITATIVE DESIGNS

**Quantitative designs following a quantitative approach** seek to confirm set hypotheses by means of quantified data and information. In all quantitative designs, units of observation (people, households, etc.) are selected methodically, data systematically collected using either a structured form or a questionnaire, and particular attention paid to units of measure to ensure that data can be collated for analysis. Quantitative and/or qualitative data may be collected, with quantitative analytical methods used to quantify such things as percentage of households that speak a given language, number of beneficiaries who used assistance to buy food, etc. or to infer data back to a larger population of interest.

**Descriptive research** aims to describe a variable. Analysis takes the form of descriptive statistics. **Correlational research** aims to explore relationships between variables using statistical data. Analysis is done through cross-tabulations, linear correlation and regression analysis. **Comparative research** aims to compare a variable between groups, across space or over time. Analysis takes the form of statement of differences and ratios, comparison of means, etc.

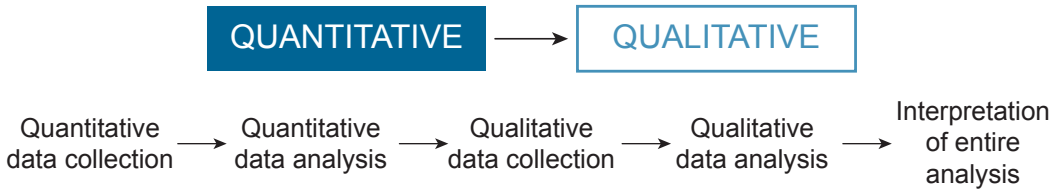
## MIXED-METHOD DESIGNS

Mixed-method designs use a combination of qualitative and quantitative approaches, and thus follow a mixed-method approach. They are considered pragmatic, because the researcher is free to use whatever approach is thought to provide the best understanding of the subject, or appropriate for any given piece of information, or for triangulating or substantiating pieces of information. Mixed-method designs can be classified by the order in which data are collected (sequential or concurrent), the weight or priority given to one of the two approaches (focus on qualitative or quantitative or equally divided between the two) and the point along the process when data are combined for analysis (data collection, analysis or interpretation). Listed below are four designs with graphical representations of the workflow identified by Tashakkori and& Teddlie (2003); these designs are common in humanitarian work (but they may not have the same names). It should be kept in mind that these are only four of the various mixed-method approaches[22] that can be taken; that in practice none of them may be as rigid as implied by a simplified graphic representation; and that they may take many different shapes.

22 See Tashakkori and Teddlie, 2010, for examples of these and other methods.

### SEQUENTIAL EXPLANATORY DESIGN

Data collection and analysis are first quantitative, and then qualitative (to assist in interpreting and explaining the quantitative findings). The primary focus of this design might be to interpret features and relationships. It is particularly useful when quantitative analysis produces unexpected results.

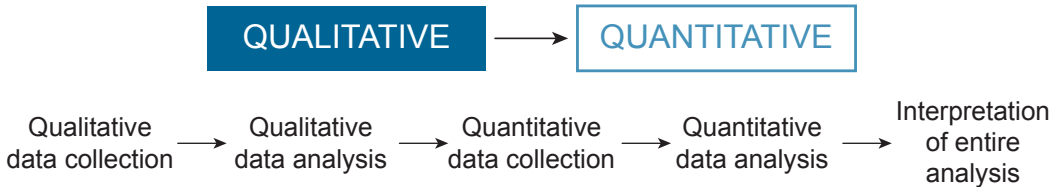



**EXAMPLE**

An initial household-level assessment is done to collect basic information to register beneficiaries for emergency relief assistance. A few vulnerability indicators are also collected for use as a baseline. The data show that expenditure on health care and medicines is high for households in two of the four neighbourhoods assessed. Focus-group discussions are conducted in each neighbourhood, with a view to examining basic household expenses, the factors driving the expenditure on health care and medicines (cost, access, social support, etc.), and the underlying health issues that may be creating the need for medicine and care.

### SEQUENTIAL EXPLORATORY DESIGN

Data collection and analysis are first qualitative, and then quantitative (to assist in interpreting the qualitative findings). The primary focus of this design might be to explore a particular phenomenon.

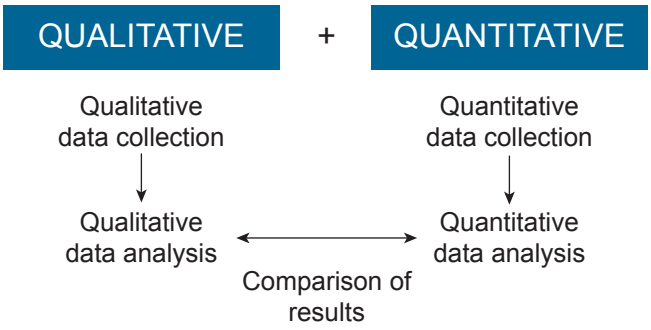



**EXAMPLE**

An initial rapid assessment was carried out after clashes in a village. It consisted of open-ended discussions with key informants and direct observation of the damage. The informants reported that a number of families whose homes were damaged had moved in with other families, and that the crowding in some of these households was 'unacceptable'. A follow-up household-level assessment was then conducted, using structured household interviews. It prioritized information on shelter, as it sought to estimate the number of households living in overcrowded conditions and the state of their original homes to inform decisions on humanitarian assistance.

### CONCURRENT TRIANGULATION DESIGN

Qualitative and quantitative data collection and analysis take place separately at the same time in the same study. Findings are cross-validated as part of analysis and interpretation. One method may be used to offset weaknesses in another.

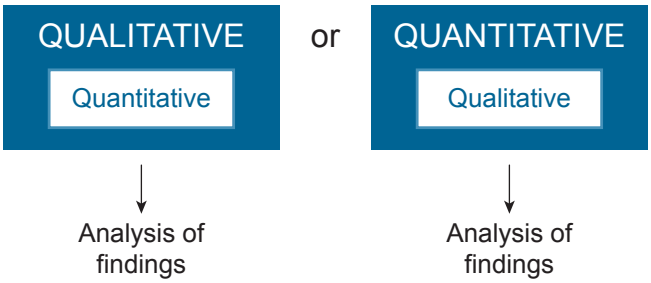



**EXAMPLE**

After a poor harvest, an initial assessment was undertaken to survey the economic impact on households. It consisted of 50 structured household interviews and three focus-group discussions with men, women and local authorities; the aim was to learn more about the consequences of the poor harvest for both the farmers' access to income and the general population's access to food. The household interviews collected data on diversity of diet, main sources of food and coping strategies; the focus groups sought to find out what the contents of the basic food basket were before and after the harvest, evaluate household strategies for obtaining food before and after the harvest, and determine which households were most affected. Data were analysed and then compared to confirm results. For example, the composition of the basic food basket was compared to what households were reportedly eating; and the results of the survey of households affected were compared to the findings on diversity of diet and use of harmful coping strategies.

### CONCURRENT NESTED DESIGN

As in concurrent triangulation, qualitative and quantitative data collection and analysis happen simultaneously; the difference is that one approach – qualitative or quantitative – predominates. Data are analysed together, and one method is used to support the other if the latter cannot produce a complete analysis by itself.




**EXAMPLE**

During the review of a microeconomic initiative, all the participants in the project are interviewed. A structured questionnaire is used for data collection. Responses to closed-ended questions are followed by open-ended enquiries in those cases where the team needs a more detailed explanation. For example, participants are asked if they expect the initiative to continue in the future (a 'yes-or-no' question). If the participant says 'no', he or she is asked to explain why not. Complete answers are transcribed for subsequent use. During data analysis, answers are coded and categorized into like groups (absence of demand in market, competition, lack of interest, other competing responsibilities, etc.) and compared against other key variables either on file or collected during the interviews (business type, business location, net income generated, baseline level of business management skills, etc.); the ultimate aim is to reach an understanding of the various factors that contribute to the economic initiative's continuation from the point of view of the beneficiary.

In such an exercise, responses can even trigger a larger conversion that is off topic (not considered in the design) but relevant. In this case, notes are taken and data are later considered in analysis.

## FEASIBILITY

Analysis design must take into account the feasibility of implementing a particular design at the design stage, and it should be reassessed throughout the process, making adjustments as necessary. Plans for mitigating data-collection challenges can be built into the design (backup sites for site selection, alternative methods for collecting or triangulating data that may not be easy to acquire, etc.). The feasibility of producing information can change for various reasons: the time-sensitive nature of certain information, initial enquiries proving insufficient, resource and access issues, etc.

Ideally, resources and logistics should be determined by information needs, but this may not always be practicable: contextual factors – such as humanitarian access, local practices and customs, or limitations of time and resources – may play at least as big a role.

### ACCESS

Access to people in need may be difficult or non-existent, owing to uncertain security conditions, poor infrastructure, government restrictions, etc. This should be taken into account at the very beginning: in any design that is set up to gather primary data only from areas that are accessible, provision should be made to use secondary sources or other means to collect data in inaccessible areas; the design should also ensure that data collected by these means are analysed with caution

**Figure 4 - A demonstration of an instance of the entire population of interest (dots in grey) being inaccessible (inaccessible areas in red).**

If there is a possibility of some areas becoming inaccessible at some point during the exercise (upsurge in violence, heavy rains, etc.), mitigation plans should be made. For example, in a primary-data collection exercise, backup locations with similar characteristics could be identified in advance, sample sizes increased in consideration of the possibility of lack of response, and flexibility built into schedules to make them more responsive to changing circumstances. Reframing a sample mid-exercise may be required as well.

### LOCAL PRACTICES AND CUSTOMS

Local practices and customs, and cultural and religious occasions, must be kept in mind when planning collection dates (interviews with authorities should be planned for the first four working days of the week because they may not work on Fridays, Friday may be the only market day, households may be fasting during a given holiday, etc.); the choice of interviewees may be restricted, and that should also be taken into account (e.g. it may not be customary for a female head of household to speak on behalf of the entire household). If these matters are not taken into consideration during the planning phase, the exercise will face disruptions, and the result may be data loss and non-response bias. For all these reasons, it is vitally important to work with resident staff and/or the community to design and plan an exercise.

## TIME AND RESOURCES

The total time and resources required will be decided by the amount of existing data readily available, the number of consultations required, the amount of primary data that has to be collected, competing priorities, etc. Estimating the time and resources needed can be done only on a case-by-case basis.

### PRIMARY-DATA COLLECTION

For collecting primary data through field visits and surveys, the amount of time required can be defined by the number of field sites and interviews/discussions (e.g. sample size), the complexity and number of interviews and/or group discussions and the time required for travelling between sites (if there is more than one site). The following simplified formula, adapted from one developed by Action Contre le Faim (ACF), demonstrates the relationship between sample size and time and resources:

$$n_S = n_I \times n_E \times n_D$$

where:

$n_S$ = feasible sample size (given the resources available)
$n_I$ = number of interviews or discussions per day per data collector
$n_E$ = number of data collectors
$n_D$ = number of days

The following table gives an overview of some of the many elements that may have to be taken into consideration when calculating the largest sample size possible. The 'Planning tool' in the 'ICRC sample calculator', which is available at the EcoSec Resource Centre, can be used to calculate this.

| FACTOR | | ELEMENT | DESCRIPTION |
|---|---|---|---|
| $n_Q$ | Number of interviews or discussions per day per data collector | Time available in the day | Working hours minus (travel time + break) |
| | | Time for introductions | Introductions at site |
| | | Time needed per interview/ discussion* | Time needed per interview/ discussion |
| | | Time needed for other interviews/discussions | If there is more than one interview/discussion per site (e.g. focus group and household interview), then the time for each should be calculated separately and included in the total time required for each day |
| $n_E$ | Number of data collectors** | Number of data collectors available | The total number available every day for the duration of the exercise |
| | | Number of data collectors needed per interview/discussion | Decide whether the interview/ discussion should be conducted by more than one person |

| FACTOR | | ELEMENT | DESCRIPTION |
|---|---|---|---|
| $n_D$ | Number of working days | Number of days for field exercise | Total number of days, given the resources and the availability of data collectors |
| | | Number of days required for travel | Travel time between sites |
| | | Number of non-working days | Rest days, holidays and days where people may not be available for the survey |

* There is no mathematical formula for calculating this, as it depends on the type of study being undertaken (e.g. a complete assessment at the household level – depending on the number of households and the number of interviews with key informants – can take anywhere from 30 minutes to 3 hours, while a rapid damage assessment may involve only a few focus-group discussions).

** If there are enough data collectors and transportation options (vehicles, flights, etc.), data collectors can be divided into two or more teams, which would enable greater geographic coverage and entail less travel time.

**MEASURES TO MITIGATE RESOURCE CONSTRAINTS**

| | |
|---|---|
| **Review the resources available** | *Can we get more data collectors?*<br>*Can we add a few extra days?* |
| **Review the amount of data being collected** | *Do we really need all these variables?*<br>*How are we going to use them?*<br>*Are some already available from secondary sources?* |
| **Re-evaluate the indicator** | *If I do not have enough resources to collect a household/individual-level indicator, is there another method I can use to collect this type of information?*<br>*Will the conclusions reached via analysis be good enough?* |

## ANALYSIS DESIGN TOOLS

There are a number of tools that can help in analysis design. Four of those most commonly used in humanitarian work – indicators, criteria, frameworks and analytical plans – are described below.

### INDICATORS

Indicators for **assessments and situation monitoring** must be chosen carefully, because they have to provide a snapshot of the current situation and/or may be compared to the past or what is expected in the future, or enable comparisons to be made between one location and another (e.g. fluctuations and differences in market prices). The same indicator may be employed as an 'outcome indicator'.

In the **monitoring and evaluation of programme activities**, indicators are used to measure the effectiveness of programmes. In monitoring and evaluation, each indicator is classified according to its use in the analysis. There are generally three categories: **process indicators,**[23] **outcome indicators** and **impact indicators**.

23 Some organizations refer to process indicators as 'output' or 'activity output' indicators.

| | WHAT DOES IT MEASURE? | EXAMPLE[24] | WHEN IS IT USED? |
|---|---|---|---|
| **Process indicator** | Measures the implementation (the process and the output) of programme activities | XX number of companies received ICRC contributions for salaries for newly trained urban IDPs | Activity monitoring (e.g. post-distribution monitoring, or PDM) |
| **Outcome indicator** | Measures the short to medium-term effects of programme activities on beneficiaries' lives | XX number of families are able to earn enough income through formal employment | Activity monitoring (e.g. PDM) Activity/ programme review |
| **Impact indicator** | Measures the long-term impact of the programme, at the beneficiary or community level | XX number of beneficiary families remained formally employed even after the end of the programme | Final evaluation |

### CRITERIA

For the purposes of this guide, **criteria** are any principles or standards against which something may be judged or decided. In an analysis, this could be either just a baseline or threshold against which to compare the value of an indicator or something more complicated, such as the profile a household must have in order to be considered economically insecure. Criteria are normally adopted, adapted or developed as part of the analysis design in order to fit to the analytical framework and provide a basis on which the data will be analyzed;, thereby ensuring that all required data are not only collected, but also collected in a way that enables appropriate analysis. Below are some of the most commonly used criteria in humanitarian work:

| CRITERIA | DEFINITION | EXAMPLE |
|---|---|---|
| **Baselines** | Defined points (measures) regarded as the 'basis' for comparison. Often the measure before, during normal times, on average, etc. | An assessment of Lebanese returnees from Syria found that 48% of were employed as low-skilled wage earners, compared to 13% of the same group before their displacement.[25] |

24 Examples taken from ICRC Colombia *Access to Employment Programme*, 2013

25 IOM, November 2014.

**EXAMPLE MULTIPLE CRITERIA**

The following model was used as a basis for defining 'vulnerable' returnee households in Lebanon. The aim was to have a standard and transparent mechanism for identifying households vulnerable to economic insecurity, and enough flexibility for a review of borderline cases.

The model is broken up into nine domains that are considered to contribute to overall household economic security. Each domain is measured by context-specific indicators and baselines/thresholds identified during a consultation process. The model uses both qualitative and quantitative indicators to feed a mathematical formula that produces an individual score for each domain and a global vulnerability score

| CRITERIA | DEFINITION | EXAMPLE |
|---|---|---|
| **Thresholds** | Predetermined points (measures) that must be crossed to produce a given effect or elicit a response. A threshold must be constant. | The ICRC's EcoSec team in the Democratic Republic of the Congo considers 'four' to be the threshold for the household dietary diversity score (HDDS); households with an average of less than 'four' are considered to be vulnerable. |
| **Categories** | Either text or numbers, but limited to a range of specific options or categories. They are discrete in nature. Databases often refer to category lists as "domains". | The Comprehensive Food Security Survey in Yemen, in early 2014, categorized heads of households into four groups: married with several spouses, married with one spouse, divorced/separated, widowed/single.[26] |
| **Scales** | A technique used to provide order to data as a reference to which they can be compared and/or measured (e.g. first, second, third place). Qualitative scales are in effect a type of category that has a specific order. | A structural damage assessment following the Gaza war in 2014 classified the level of damage on a generalized scale: moderate, severe and destroyed.[27] |

Some analyses combine a number of domains and associated indicators, measures and criteria to create a multiple criteria matrix or framework. For example, determining the vulnerability of an individual may require looking at a number of domains and these may or may not have a number of indicators of which each may or may not be be assigned equal weight (contribution to the level of vulnerability, contribution to the level of risk, etc.).

26 WFP, CSO, UNICEF, 2014.

27 UNOSAT, September 2014.

| VULNERABILITY CRITERIA | | | | | | |
|---|---|---|---|---|---|---|
| **COMPONENT** | **INDICATOR(S) MEASURED** | **1 = LESS** | **2** | **3** | **4** | **5 = MOST** |
| 1 Family composition | Depedency ratio[1] and head of family status | Dependency ratio[1] <=200% | Dependency ratio 201-300% | Dependency ratio 301-400% | Dependency ratio 401-500% | Dependency ratio >=501% or single parent family |
| 2 Income | Access to income generating activities | Permanent IGA | | Temporary or seasonal IGA | | No IGA |
| | Monthly expenditures against Expenditure Basket per capita[2] | 30% > MEB ($105/pers/month) | 16-30% > MEB | 1-15% > MEB | <=MEB and >Survival MEB (SMEB=$88/pers/ month) | Equal to or less than the Survival MEB |
| | Level of debt | 0 debt | 1-150 US$ in debt | 151-300 US$ in debt | 301-450 US$ in debt | >450 US$ in debt |
| 3 Living conditions | Structure type | Whole house or apartment | | Room in a house/ apartment | | Collective centre, informal settlements or homeless / squatting |
| | Structure condition | Good/Acceptable and working heating system | | Need fixing doors/ windows or partially functioning heat system | | Need roof fixing or not waterproof/flooded or no heat system |
| | Crowding of living space | ">3.5 sqm/pers OR 1-2 pers/room" | | "3.5 sqm/pers OR 3-6 pers/room" | | "<3.5 sqm/pers OR > 6 pers/room" |
| | Access to heating and cooking energy | Able to cover costs of both heat and cooking | | Partially able to cover costs of one and able to cover costs of other | | Unable to cover costs of either |
| | Access to clean water | Full access | | | | Not full access |
| | Type of sanitation facility | Flush | Traditional | | Improved | Open air |
| | Number of people sharing toilet | <3 pers/shared toilet | <5 pers/shared toilet | 5-7 pers/shared toilet | 7-10 pers/shared toilet | >10 people/shared toilet |
| | | Comfortable household or livelihood assets and essential assets | | Essential assets | | Not all essential assets |
| 4 Food consumption | Meals per day (any member) | Three | | Two | | One |
| | Expenditures on food | <45% | 45-54% | 55-64% | 65-74% | >74% |

| VULNERABILITY CRITERIA | | | | | | |
|---|---|---|---|---|---|---|
| **COMPONENT** | **INDICATOR(S) MEASURED** | **1 = LESS** | **2** | **3** | **4** | **5 = MOST** |
| 5 Health | Access to health care | Full access | | Partial access (either due to lack of services available or capacity to cover costs) | | No access (either due to lack of services available or capacity to cover costs) |
| | Family member health needs | No family member with special needs related to health | | Pregnant or lactating woman family member | | Family member with disability, chronic illness or serious medical condition |
| 6 Coping mechanisms | Coping strategies framework[3] - coping mechanisms effect on household | No mechanism or reversable coping mechanism (1) | | Borderline coping mechanism (2) | | Irreversable coping mechanism (3) |
| 7 Protection | Unaccompanied children or elderly | No | | | | Yes |
| | Victim of violence | No | | | | Yes |
| | Time since and place of arrival in Lebanon | >2 years | | 1-2 years | | <1 year |
| 8 Assistance | Access to assistance | On-going assistance | | Had received at one point | | Never received |

| SELECTION CRITERIA | | | |
|---|---|---|---|
| Vulnerability Score[4] | <15 | 15-20 | >20 |
| Vulnerability Level | Could be most vulnerable in no more than one domain | Could be most vulnerable in two to three domains | Could be most vulnerable in more than three domains |
| Selection status | Not selected | Reviewed | Selected |

**Indicator definitions**

1 Dependency ratio - (# of children below 16 + # elderly above 59 + # adults unable to eat, wash or go to toilet on their own)/(# of adults able to eat, wash and go to toilet on their own). Adapted PMT.

2 Minimum Expenditure Basket (MEB) - $571/hh/month ($105/pers/month). Defined by Cash Working Group, LBN.

2 Survival Minimum Expenditure Basket (SMEB) - $435/hh/month ($88/pers/month). Defined by Cash Working Group, LBN.

3 Coping Strategies Framework - as per defined by ICRC EcoSec BEY, LBN Oct 2014.

3 Coping Strategies Framework - as per defined by ICRC EcoSec BEY, LBN Oct 2014.

A **framework** is a matrix of elements based on relevant assumptions, theories, variables and indicators and/or criteria. There are innumerable types of framework, and they are often given names other than 'framework' (model, process diagram, etc.). The frameworks most commonly used in humanitarian work are 'conceptual' and 'logical'.

### CONCEPTUAL FRAMEWORK

**Conceptual frameworks** are analysis tools for explaining actual processes or phenomena in an abstract or generalized manner. Conceptual frameworks are developed from historical evidence and knowledge of a given topic, and can be used as guidance for data collection during an exercise. They can also ensure comprehensiveness of data collection and analysis in a particular exercise, and uniformity in exercises analysing similar phenomena. They may be used in their entirety (e.g. all relationships in the framework are considered) or only partially (e.g. one relationship within the framework is considered).

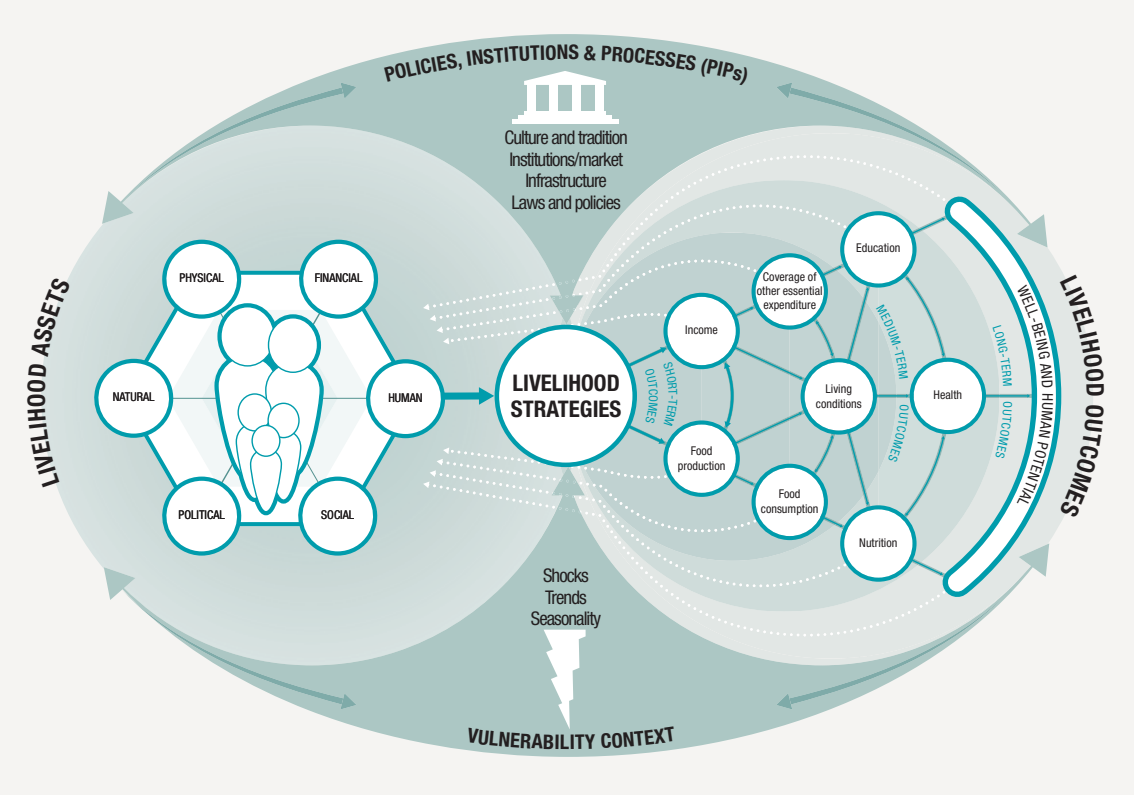



**Figure 5** - **Economic Security Conceptual Framework**

The Economic Security Conceptual Framework (adopted by EcoSec and adapted from the DFID *Sustainable Livelihoods Framework*, 1999) describes, in a simplified way, the interaction between livelihood assets, strategies and outcomes, and how they are affected by and influence policies, institutions and processes (PIPs) and the 'vulnerability context'. A more detailed description of the framework and its field application can be found in the EcoSec handbook *Assessing Economic Security* (ICRC, 2016).

### LOGICAL FRAMEWORK

A **logical framework** demonstrates the causal relationship between the elements that are to be measured. The **results framework** is one type of logical framework, and used in monitoring and evaluation. Results frameworks outline the results expected from a monitoring and evaluation exercise, which then serve as a guide to what must be monitored (process, outcome and/or impact indicators). Results frameworks are a key tool in the results-based management approach taken by the ICRC. See the ICRC handbook *EcoSec Project Design, Monitoring and Evaluation* (ICRC, 2016) for more information.

| | | | |
|---|---|---|---|
| **DESIRED IMPACT** | Most vulnerable households living in conflict-affected areas close to the line of contact and international border with Armenia are **able to cover their essential needs and unavoidable expenses** | 100 households unable to cover their essential needs and unavoidable expenses | 100 households able to cover their essential needs and unavoidable expenses without further ICRC assistance |
| **MEDIUM-TERM OUTCOME** | Most vulnerable households **increase income** | Average monthly household income is 250 Azerbaijani manat | 80% of households benefiting from the microeconomic initiative (MEI) increase income by 30% within 12 months of the launch of the MEI |
| **SHORT-TERM OUTCOMES** | Most vulnerable households have **access to microcredit** | No access to microcredit | 90% or more of MEI beneficiary households with sustained access to microcredit for further business development |
| | Most vulnerable households have **means to repay loans** for project-supported MEI | 100 households with loans for MEI | 95% or more of MEI beneficiary households repay loans (plus interest) in accordance with contractual obligations |
| | Most vulnerable households have access to income-generating activity (IGA) | 100 households without regular IGA | 100 households approved for microeconomic project |

**Table 3 - Extract from results monitoring framework for ICRC Azerbaijan's Operational Plan 2015-2016**

## ANALYSIS PLAN

An **analysis plan** sets out, for a specific exercise, the data that have to be collected for a specific exercise, and from where and/or whom; it also describes how the data need to be combined to enable the drawing of conclusions, and the types of analysis to be used.[28] Analysis plans combine all the indicators and criteria, and the framework that are to be used, and give them a structure. They are normally developed for specific data collection and analysis exercises.

There are various ways of drafting a analysis plan: in combination with another tool, such as a log frame or indicator tracking table, as a formal structured document or informally (by agreed upon it) in a planning meeting.

Whatever its form, the analysis plan is essential for creating a bridge between the design and implementation of data collection and the results derived from it. It guides the choice of data-collection methodology or combination of methodologies and the elaboration of data-collection and analysis tools (questionnaires, guides, etc. for the former; data entry and analysis files, visual templates, etc. for the latter). An analysis plan ensures efficiency,

28 Analysis plans are given other names as well: data-collection plans, indicator-planning matrices, monitoring and evaluation planning tables, etc. These names will depend on their overall purpose and format and the users, who will vary with every exercise and the scope of the plan.

thoroughness and cross-checking of the methods to be used and helps to determine the feasibility of the methods identified. It can include any of the following components:

- Information needs
- Information on the context
- Indicator(s) and criteria for analysis (baseline, threshold, etc.)
- Data required
- Data collection method(s)
- Type of analysis

**Table 4 - Extract from analysis plan for ICRC Mali's Household Economy Assessment, 2014**

| INFORMATION NEEDS | INDICATOR | BENCHMARK | DATA SOURCE | ANALYSIS TYPE[1] |
|---|---|---|---|---|
| **Question 1:** Who are the most vulnerable households in terms of economic security? | | | | |
| Food consumption | HDDS | 4 | Household questionnaire (random sample) | ▪ Descriptive statistics by location<br>▪ Cross-tabulation with household status, main livelihood activity, poverty level, head-of-household status |
| | Number of meals/day | 2 | | |
| | Sources of food | Stable against non-stable | | |
| | Level of expenditure on food | >70%, 60-70, 50-60, 40-50, <40 | | |
| Food production | Access to pasture for animals | Unhindered access, together with productive inputs required | Focus-group discussion (group of livestock farmers chosen specifically for the purpose) | ▪ Qualitative and quantitative description |
| | Access to land for farming | Unhindered access, together with productive inputs required | Focus-group discussion (group of agriculturalists chosen specifically for the purpose) | ▪ Qualitative and quantitative description |
| … | … | … | … | … |

Decisions about information needs should always specify the unit(s) of analysis – such as by geographical region, livelihood zone, etc. – and/or any disaggregation that needs to be done – such as sex of head of household, displacement status, etc. This information will identify the need for stratified samples (or not) and/or may imply additional data-collection needs (household questionnaire to include sex of head of household, displacement status, etc.).

### CHOOSING THE RIGHT DESIGN TOOL

All this talk of models, criteria, frameworks and analysis plans can be a bit confusing. The idea is to understand each, and to use whatever best suits a particular situation. One or more of these analytical tools may be used: by itself (or themselves), together at the same time but separately, or in harmony (in combination).

| TOOL | USES | EXAMPLES |
|---|---|---|
| **CRITERIA** | ▪ Assessments<br>▪ Situation monitoring<br>▪ Case management (e.g. registration)<br>▪ Monitoring and evaluation | ▪ Vulnerability criteria<br>▪ Selection criteria<br>▪ Risk criteria<br>▪ Referral criteria |
| **CONCEPTUAL FRAMEWORK** | ▪ Assumptions and theories<br>▪ Assessments<br>▪ Situation monitoring<br>▪ Monitoring and evaluation | ▪ EcoSec Conceptual Framework<br>▪ Conceptual framework for the causes of malnutrition |
| **LOGICAL FRAMEWORK** | ▪ Situation monitoring<br>▪ Monitoring and evaluation | ▪ Results-monitoring framework |
| **ANALYSIS PLAN** | ▪ Assessments<br>▪ Situation monitoring<br>▪ Monitoring and evaluation | ▪ Assessment analysis plan<br>▪ Monitoring analysis plan<br>▪ Evaluation analysis plan |

**Table 5 - Uses for analytical design tools in humanitarian work**

## ERROR AND BIAS

The accuracy of the results of an exercise can be compromised by a variety of factors: unavailability of background data, insufficient amounts of data collected, misinterpretation of the exercise or subject matter at hand, varying quality and quantity of responses in primary-data collection, sampling errors, bias, etc.

Potential sources of error and bias should be taken into account in the design phase; mitigating bias and error to the greatest extent possible is crucial because they can lead to inconclusive and/or incorrect results, and therefore also to the misuse or rejection of information.

### MITIGATING ERROR

In data and analysis exercises, errors normally take the form of measurement errors during data collection and errors in analysis and reporting.

To mitigate measurement error, all tools should be tested and data collectors trained before data are collected. Testing and training should be done not only in the office, but also via a role-playing exercise or by pilot-testing the process before it is undertaken.

Data analysis and reporting should be peer-reviewed, in order to double-check work before it is shared. Collaborative analysis and interpretation sessions can also help to detect errors.

### MITIGATING BIAS

Bias occurs when the accuracy and precision of a measurement is threatened by the particular experiences, perceptions and assumptions of the researcher, or by the approaches, methods and tools used for measurement and analysis.[29] The box below lists various kinds of bias.

29 IFRC, 2013

| INTENTIONAL | UNINTENTIONAL |
|---|---|
| Example - Analysts collect and consider data from only one side of the conflict because their organization has contacts among the people on that side. | Example - A sample is stratified by the principal livelihoods in a region (e.g. agro-pastoralists, fishermen and small-business owners); however once in the field, the researchers realize that there is a small mining population that has no voice and was not taken into account in the sampling method. |
| **AVOIDABLE** | **UNAVOIDABLE** |
| Example – Proper secondary data analysis does not identify all possible vulnerable groups when choosing an appropriate sample method and size before leaving for the field. | Example – A region is not included owing to security or infrastructure constraints. |

Unintentional and unavoidable bias should always be taken into account during analysis, and reported as limitations of the data, information and conclusions.

## DATA PROTECTION AND ETHICS

Ethical issues and data-protection measures must be taken into consideration before data are collected or collated.

### ETHICAL CONSIDERATIONS

Ethical considerations in data collection and analysis for humanitarian work should follow the guiding principles for ICRC assistance:[30] most specifically, respect for cultural usage and customs, the principle of 'do no harm' and accountability to those we seek to assist. Some common considerations are listed below.

1. Primary-data collection exacts a cost, in terms of time, from both the respondent and the data collector. While it is important to take enough time to acquire information of quality, on which sound decisions can be based, **excessive and/or repeated interviews or other data-collection exercises** can be tiring for all parties involved, take time away from other priorities, and lead to a flood of data.
2. **Questions of a sensitive nature, or on certain subjects,** may be difficult for respondents to answer, or a source of stress, or may cause a disturbance in the community.
3. **Remote digital means of data acquisition,** through SMS or telephone, may be biased towards those with access to a phone and digital network; another point to remember is that the identity of the respondent may be difficult to verify.
4. **The use of social media, through remote digital means,** may not yield sufficient amounts of accurate data and may be biased towards those with access to and using these technologies (digital discrimination); and that will lead to oversights or to assumptions that are incorrect. In some cases, public sharing or sharing over an insecure network may put an informant at risk
5. **Automated analytical methods with no human oversight** may exclude vulnerable populations from the analysis and/or from receiving humanitarian assistance.

Ethical issues must be taken into consideration during the design phase. The simple checklist given below can be used for each piece of data and for every data collection and analytical method. The list is not exhaustive. Ethical considerations are likely to vary from one context to the next, and the list should be reviewed in each context for relevance.

30 See ICRC *Assistance Policy*, Section 3, "Guiding Principles"; and *The Code of Conduct for the International Red Cross and Red Crescent Movement and Non-Governmental Organizations in Disaster Relief*.

**THE ETHICS OF DATA COLLECTION AND ANALYSIS: A CHECKLIST**

- *Why are we collecting this data? Do they already exist?*
- *What will we do with this data? Do we need it all? Do we need this depth of detail?*
- *Can asking these questions upset or agitate the respondent?*
- *Can our collection of these data put the respondent at risk?*
- *Can our collection of these data cause a disturbance in the community?*
- *Are the methods used inclusive and representative?*
- *Do the respondents have the capacity to respond? Are they aware of any risks associated with their participation in this exercise?*

Primary-data collection usually entails a certain level of interaction with informants and/or the community where the data are being collected. These informants and communities must, *at the very least,* be made aware of what data are being collected, and why, and how they will be used. If this cannot be done, then the need to collect the data should be re-evaluated. Communication with informants and communities is discussed in Chapter 4 Primary-data collection in the section titled "Communication and consent".

## DATA PROTECTION

**Data protection** is the collective term for the set of basic principles, rights of data subjects, data controllers' obligations (including 'data security' and 'data integrity') and enforcement measures required to prevent data loss, misuse of data or the breaching of personal rights to data protection and privacy.

**Data security** refers to the technological and organizational measures required to provide adequate protection for data from any risks to which they may be exposed. In this guide, **data integrity** refers to maintaining and ensuring the accuracy and consistency of data over their entire life cycle.[31]

In addition, the rights of data subjects must be kept in mind while collecting and processing personal data. Information on this and on the basic principles of such data collection are available in the *ICRC Rules on Personal Data Protection* (ICRC, 2016) and in even greater detail in an internal ICRC document, "ICRC Reference Framework for Personal Data Protection Handbook" (ICRC, 2015). It is important to note that personal data are not limited to details that can directly identify someone (names, phone numbers, GPS coordinates, addresses, etc.), but can also be data that, when combined, can effectively identify someone. For example, let us take a hypothetical dataset of household interviews that includes information on the village where the interviews are being conducted, the marital status of heads of households, the number of members in each household and the main source of household income. How difficult is it likely to be to identify a widow with three children, whose main source of income is her tailoring business? How many households in that village will have those characteristics?

### DATA PROTECTION BY DESIGN

All aspects of data protection must be considered during the design phase of an exercise. The extent of the data-protection measures required will depend on the nature of the exercise. For example, in every primary-data collection exercise, the data controller (in this case, the ICRC) must, at the very least, meet all their obligations concerning data security and data integrity.

31 Wikipedia definition, accessed in April 2015.

**Data minimization** is one of the first measures that can be taken for ensuring data protection. It is both one of the basic principles of data protection (see "Article 4: Adequate and relevant data" in the ICRC's "Reference Framework on Personal Data Protection Handbook"[32]) and a preventive measure, because data do not have to be protected if they do not exist.

**EXAMPLE**

In an activity monitoring exercise, a selected group of beneficiaries is interviewed for the purpose of monitoring process and outcome indicators. Names of heads of households, phone numbers and GPS coordinates are collected; however the data are never used. In this case, a separate form could be made available for specific beneficiaries to fill out – for instance, for people who raise other issues that have to be followed up (and therefore whose contact information is required).

Furthermore, everyone involved in collecting and analysing data should consider taking key data security and integrity measures such as the following:

- ensure that paper forms and reports are safely handled and not left unattended or unlocked;
- ensure that data shared electronically are conveyed over a secure network with appropriate encryption;
- include minimum metadata about the data, such as: data-collection methods, data source, date of the data and restrictions, if any, on the use of the data;
- include methods of data collection and analysis in any reports so that it is used correctly, and if repeated analysis will be performed, the analysis is done using the same methods or comparative analysis considers any differences in the methods;
- determine the period for which data need to be preserved and stored, and any protection measures that may be needed for the duration of the data's lifecycle; and
- when sharing data, determine the terms of use for third parties.

For more detailed information, see the ICRC's "Reference Framework on Personal Data Protection Handbook" (ICRC, 2015). Programme staff cannot be expected to know every aspect of data protection; there are specialists who can help them to ensure that data are protected. When using new methods for data collection and analysis or technological methods new to the ICRC, or when working with personal data, it would be prudent to consult the pertinent data-protection specialists.

32 ICRC, 2015.

# REFERENCES

Creswell, John W. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 2003. Sage Publications, Inc., 3rd ed. Available from: http://isites.harvard.edu/fs/docs/icb.topic1334586.files/2003_Creswell_A%20Framework%20for%20Design.pdf.

Beebe, James, *A Field Guide to Team-Based Assessment,* 2nd ed., 2014.

DFID, *Sustainable Livelihoods Guidance Sheets*, April 1999.
Available at: http://www.eldis.org/.

Flick, Uwe, *An Introduction to Qualitative Research, 4th ed.,* SAGE Publications, London, 2009.

ICRC, Assessing Economic Security, 2016.

ICRC, Assistance Policy, April 2004.
Available at: https://www.icrc.org/eng/assets/files/other/irrc_855_policy_ang.pdf.

ICRC, "Assistance Reference Frameworks", July 2014. Internal document. ICRC Colombia, "EcoSec Executive Brief: Colombia - Access to Employment Programme 2013", May 2014.

ICRC, "Economic Security Conceptual Framework", October 2014. Internal document.

ICRC, "Guiding Principles on Assessments", July 2013. Internal document.

ICRC, "Lebanon: Cash-Transfer Programming Vulnerability Assessment Criteria", November 2014.

ICRC, *EcoSec Project Design, Monitoring and Evaluation,* 2016.

ICRC, "Reference Framework on Personal Data Protection Handbook", December 2015. Internal document.

ICRC, *ICRC Rules on Personal Data Protection*, ICRC, Geneva, January 2016.
Available at: https://shop.icrc.org/publications/international-humanitarian-law/icrc-rules-on-personal-data-protection.html.

ICRC Colombia, "EcoSec Executive Brief: Colombia - Access to Employment Programme 2013", May 2014.

IFRC/ICRC, *The Code of Conduct for the International Red Cross and Red Crescent Movement and Non-Governmental Organizations in Disaster Relief*, 2004.
Available at: http://www.ifrc.org/en/publications-and-reports/code-of-conduct/.

IFRC, *Project/Programme Monitoring and Evaluation (M&E) guide*, IFRC, Geneva, 2013.
Available at: http://www.ifrc.org/en/who-we-are/performance-and-accountability/monitoring-and-evaluation/.

International Organization for Migration (IOM), *Refugees at Home: A Livelihoods Assessment of Lebanese Returnees from Syria*, November 2014. Available at: https://www.iom.int/files/live/sites/iom/files/Country/docs/Syria-Factsheet-Refugees-at-Home-IOM-Nov-2014.pdf.

Saldaña, Johnny, *Thinking Qualitatively: Methods of Mind*, SAGE Publications, London, 2015.

Tashakkori & Teddlie, *Handbook of Mixed Methods in Social and Behavioral Research*, SAGE Publications, London, 2003. Available at: http://www.sagepub.com/sites/default/files/upm-binaries/19291_Chapter_7.pdf.

UNOSAT, *Damage to Agricultural Areas and Greenhouses, Gaza Strip – Occupied Palestinian Territory*, September 2014. Available at: http://www.unitar.org/impact-2014-conflict-gaza-strip-unosat-satellite-derived-geospatial-analysis.

WFP, CSO, UNICEF. *Yemen: Comprehensive Food Security Survey*, November 2014. Available at: http://www.wfp.org/content/yemen-comprehensive-food-security-survey-november-2014.

# CHAPTER 4
# PRIMARY-DATA COLLECTION

The goal of data collection is to gather and/or measure as much data as needed on variables of interest, in order to enable you to answer the key questions stated alongside the objectives of the analysis. The choice of data to collect is guided by the analysis plan (see the section titled "Chapter 3 **Analysis design**").

In surveys, teams often start by developing the questionnaire, skipping the essential analysis design phase and the elaboration of an analysis plan. This practice increases the risk of the data collected not fitting with the indicators that have to be measured in order to answer the key questions that necessitated the exercise; it also increases the chances of too much or too little data being collected.

This chapter of the guide focuses on primary-data collection. It does not cover secondary-data or desk reviews. The sections on analysis are, however, applicable to both primary and secondary data.

## COMMUNICATION AND CONSENT

Primary-data collection usually entails a certain level of interaction with the informants and/or the community where the data are being collected. These informants and communities must, *at the very least*, be made aware of what data are being collected, and why, and how they will be used. If this cannot be done, then the need to collect the data should be re-evaluated.

> *All our dealings with donors and beneficiaries shall reflect an attitude of openness and transparency.*
>
> Code of Conduct for the International Red Cross and Red Crescent Movement (IFRC/ICRC, 1994).

### COMMUNICATION

Communication with informants and communities that are the subject of a data-collection and analysis exercise may involve:

- **Engaging the community in the process,** either in certain steps in the process (designing the data-collection plan, drawing up beneficiary lists, etc.) or in the entire process (e.g. if a participatory approach is used to analyse the situation, develop future scenarios and humanitarian response options). The level of engagement should be made clear in the design phase.
- Letting them know in advance about the **logistics of a data-collection** exercise, including when and where it will take place, who may be involved and why.
- Communicating clearly **how data will be used, why they are necessary and who may or may not have access to them.**
- Clear communication on **how data will be processed**, and on the measures that will be taken to ensure their protection and integrity.

An open line of communication and full transparency between humanitarian personnel and the communities affected will help to strengthen the quality of both the data and the analysis. In humanitarian work, most data-collection exercises are undertaken to identify damage, vulnerability, needs and the appropriateness and effectiveness of humanitarian action; and the best sources of information are those who are actually affected. Communicating with them, and preferably engaging them in the process, is the least a humanitarian worker can do to demonstrate his or her accountability[33] to them.

Explaining the reasons for data collection, and the uses of the data, to respondents is an opportunity to engage them in the exercise and secure their *participation* in collecting data and making findings – activities that, in the end, are meant to serve them.

33 "Accountability to people affected by conflict or other situations of violence is a way of conducting activities and using resources in full respect of people's priorities, and based on their needs. As such, it is a process and an attitude and not a distinct activity" (ICRC, 2014).

## CONSENT

Data protection rules require the establishment of a 'legitimate basis' for processing data. Legitimate basis could be any of the following: the consent of the data subject, the 'vital interests' of the data-subject or of another person, the performance of a contract, or the fact that the ICRC's activities are carried out on the 'important grounds of public interest'. The last example may be relevant in such cases as the distribution of relief: here, it may not be practicable to obtain the consent of all possible beneficiaries, and the life, security, dignity and integrity of the data subjects or of other people are not likely to be at stake (in which case, 'vital interest' may be the most appropriate basis for processing).

If the staff member in charge decides to process data based on consent, a certain amount of information must be provided to the data subject. This is because consent can be said to be 'informed' only when the data subject is fully aware of the risks/implications of processing, and takes full responsibility for consenting to it.

A certain minimum amount of information about processing must be provided to data subjects, such as:

(i) its purpose (already being done regularly, as part of pre-programme dissemination/ briefings);
(ii) whether the data are likely to be shared with other organizations;
(iii) how long data will be kept before being destroyed or archived;
(iv) the fact that if they want additional information or explanation about the handling of data, and their rights in relation to it, they will find it in the notice on data protection that is being prepared for the programme in question, and/or they may ask the staff member in charge of the data-protection office.

The ICRC's rules for data protection indicate that the organization would prefer consent to be the legitimate basis for processing data. However, in the vast majority of contexts where the ICRC works – given the requirements listed above – consent is not the most feasible basis for processing. This is for a number of reasons:

1. Individuals are likely to be too vulnerable, and their circumstances too precarious, for them to be able to make a genuinely free decision.
2. Individuals may not be in a position to fully appreciate the consequences of data being processed in the manner proposed, either because of their vulnerability or because of the complexity of the processing operation in question (particularly if it involves cloud solutions, new technologies or external processors).
3. Because of the circumstances in which the ICRC works, overburdening data subjects with complex information phrased in legal jargon, on the first encounter, may be counter-productive and not conducive to the establishment of a relationship of trust.

This is why the ICRC's data protection rules state that the most suitable basis for processing is, in most cases, not consent, but some other legitimate basis or a combination of other bases, such as the following:

- vital interests of the data subject or of another person
- important grounds of public interest
- performance of a contract.

Consult with Data Protection Office in Geneva for further guidance.

## PRIMARY-DATA-COLLECTION METHODS

Data-collection methods include the tools used to collect the data, the means for recording them and the manner in which the data are physically gathered or measured. The choice of methods will depend primarily on the following:

- degree of specificity of the objectives (clearly defined information needs, exploratory needs, etc.)
- type of data (primary or secondary, quantitative or qualitative, measurement or observation, etc.)
- type and extent of access to the data sources (access to individuals, telephone access to individuals, access to informants who can speak for the individuals, etc.)
- the amount of data that need to be collected (number of informants, depth of data, etc.)

Data-collection methods have five key elements, which are described in the box below.

| | |
|---|---|
| **SOURCE** | Will the data come from a primary or secondary source? Who/What is that source? How much triangulation (more than one source) will be required? |
| **SCALE** | Does information need to be reported at the individual, local, regional or country level? |
| **TOOL** | For primary data, will you use a structured or semi-structured survey? A participatory tool? An observation checklist? A GPS device? Or perhaps just a checklist of key points? |
| **TIMING** | Will you collect the data during a certain time of year? On a recurrent basis? |
| **SAMPLE** | Do you need to take a sample? If yes, what sample will you use? How will sampling units be selected? |
| **PRECISION** | How precise do the data need to be to perform the analysis? Do you need precise measures or rough estimates? |

### DATA SOURCE, SCALE AND TOOL

Humanitarian work uses sources of primary and of secondary data. Emphasis may be laid on secondary data when they are readily available and reliable and/or when time is limited. Primary data may be emphasized when little is known about a given population or when specific questions need to be answered. The scale of the data will depend on the information requirements, and will have a direct impact on the data source and, in the case of primary-data collection, the sampling method.

EcoSec collects most of its primary data by these means: direct observation, group discussions, key informants, household or individual interviews, and individual or structured monitoring. These are some of the sources of secondary data and information that EcoSec uses: government statistics offices, academic and research institutions, reports from UN agencies and non-governmental organizations, past ICRC reports/data, etc. There are various data-collection tools. That subject will be discussed in more detail in Chapter 5.

**Table 6 - Types and objectives of primary-data sources and typical data-collection tools**

| | | |
|---|---|---|
| **DIRECT OBSERVATION** | ▪ To gather crucial qualitative information that is difficult to collect through discussions or interviews<br>▪ ComplementTo supplement/triangulate quantitative information with objective-based observations | ▪ Free-hand notes<br>▪ Checklist/Note-taking guide<br>▪ Semi-structured form<br>▪ Camera/Recording device |
| **GROUP** | ▪ To collect qualitative and quantitative information in consensus with a general group (community group discussion) or from a specific group or on a specific theme (focus-group discussion)<br>▪ To supplement/triangulate data collected at individual/household level | ▪ Free-hand notes<br>▪ Diagram<br>▪ Participatory tools<br>▪ Checklist/Discussion guide<br>▪ Semi-structured form<br>▪ Camera/Recording device |
| **KEY INFORMANT** | ▪ To collect qualitative and quantitative information on a specific group or theme from an informant speaking for the group<br>▪ To supplement/triangulate data collected at group level | ▪ Free-hand notes<br>▪ Diagram<br>▪ Participatory tools<br>▪ Checklist/Discussion guide<br>▪ Semi-structured form<br>▪ Structured form<br>▪ Communication device |
| **HOUSEHOLD, INDIVIDUAL OR STRUCTURE** | ▪ To collect reliable quantitative and qualitative data to inform indicators in a standardized way, enabling statistical analysis<br>▪ To supplement/triangulate data collected at group level | ▪ Semi-structured form<br>▪ Structured form<br>▪ Communication device |

It is not always necessary to triangulate data and information, but it is often useful. Sometimes, when there is only one source, it is even more important to review the reliability of data before using it; if such data are used, their reliability and any constraints to their use, must be reported. For examples, see the section titled "Reliability of the information" in Chapter 4 of the ICRC handbook *Assessing Economic Security* (ICRC EcoSec, 2016).

## TIMING

Analysis design should factor in any variations in data influenced by a specific period of time or of data collection, or any interval between data-collection exercises (for monitoring or comparison).

Data may be influenced by the time of day (members of households at work, stock liquidation, etc.), the day of the week (market day, public holiday, etc.), the time of the month (payment of State salaries, etc.), the time of the year (cultural holidays, seasonal changes, hunger gap, etc.). Consequently, data will be indicative of the time of their collection.

**EXAMPLE**

Data on household expenditure collected during an EcoSec assessment in northern Mali in July 2014 showed, on average, expenses for social activities that were more than what might have been expected. July was the month of Ramadan, when households spent significantly more money on food and ceremonial clothes. The data on expenditure were therefore not indicative of average monthly spending.

The changeability of variables measured over time may depend on the length of time between measurements. For example, while monitoring market prices, you will have to consider if prices are liable to change on a weekly or monthly basis in order to decide how often to collect data to identify trends.

This sensitivity over time may also influence any comparisons that you wish to make (before and after, in a given season, etc.). For example, the percentage of household production that is consumed by the households themselves: this may differ just before and after a harvest.

These factors – sensitivity over time and intervals between measurements – should be taken into account during the design and planning phase, and also during analysis, to understand their consequences for the results.

## SAMPLING

Sampling is the process of selecting units (people, households, organizations, villages, sites, etc.) from a population of interest for surveying and/or studying; the results of this survey and/or study will then be generalized back to the population from which the sampling units were chosen. Sampling is different from a census, in which every person or entity in the population of interest is included in the survey or study. Carrying out a census is often not feasible; and it will not always add to the credibility of the data collected. In humanitarian work, sampling is normally used in assessment, monitoring and evaluation exercises.

**EXAMPLE**

The EcoSec team in Gaza distributed food and essential household items to 23,491 households after the war in 2014. The team then undertook a monitoring exercise using household surveys; the aim was to learn more about the quality of the items distributed and their use by the households concerned, and also to identify households that were still vulnerable and in need, perhaps, of further assistance. A random sample of 384 households of the 23,491 was surveyed and the data reported as representative of all households (95% confidence and 5% margin of error). The sample was proportionately stratified by the municipality to ensure geographic diversity in the sample.

There are two main sampling methods: probability and non-probability. Probability sampling uses some form of random selection: in this every individual has an equal chance (probability) of selection. The advantage of probability sampling is that bias is reduced and data may be extrapolated back to the entire population of interest with a quantifiable level of precision and confidence.

Non-probability sampling is different from probability sampling in that it does not use random selection throughout the process (however, it could at certain stages); every individual does not, therefore, have an equal chance of being selected. The advantage of non-probability sampling is that it may give analysts more control over the sample and enable them to focus on key areas (assuming that they have sufficient background knowledge), thereby lessening (but not always) requirements for resources. Sampling is discussed in more detail in Chapter 5 Sampling.

## PRECISION

Precision, in the context of this guide, is the level of detail required to analyse a variable. For example, when collecting data about young children for use in a nutrition survey, the age of every child must be recorded very precisely, to the day; but when collecting data for, say, a perception survey, a rough estimate of the age of heads of households may suffice.

### CHOSING THE RIGHT METHOD

There is no single method for collecting a particular variable. For example, if you need to know the average size of households in five villages in Côte d'Ivoire, you could do any of the following: collect the size of every household and take the average; collect sizes from a sample of households and take the average; or ask key informants the average household size and take the average. Methods are determined by feasibility and analytical requirements, and by the homogeneity or heterogeneity of the population or subject of interest. Analysts should ask themselves the following questions:

- What decisions have to be made on the basis of the data?
- Will this be the primary source of information, or will it be used only for triangulation?
- How sure do we need to be? And how accurate?
- Will these data be challenged by decision-makers?

Furthermore, some data and indicators impose specific methodological requirements (e.g. malnutrition cases will require a specific way of measuring malnutrition, a minimum sample size, etc.) or have obvious scales and units (e.g. market prices will be collected in terms of quantities purchased and currency used locally), making it easier to identify data sources and methods. However, others may be measured with a variety of methods; the choice of method will be determined by the analytical approach. The conclusions reached by using different methods, even if they are based on the same data and indicators, will vary; and this must be taken into account when designing the exercise.

## LEVEL OF STRUCTURE AND FLEXIBILITY

The number, level and/or flexibility of the controls for data collection will depend on the information requirements and the degree of interoperability with other data or information that is needed.

**Data collected in a structured manner** may be qualitative or quantitative data; the method is usually employed in rigorous exercises where the information requirements are fully understood and established. Structured data collection is often an element in registrations, surveys or monitoring exercises, where the resulting analysis includes descriptive or inferential statistics and reporting. Effective collection of data in a structured manner requires attributes and methods that are clearly defined in advance, and consistency among and between data collectors and providers. Data analysis tends to be deductive, confirming or finding evidence to support ideas.

**Data collected in a more open or flexible manner** may also be qualitative or quantitative data; this method is normally used when information requirements are difficult to pinpoint or the situation not fully understood. Adequate 'space' is left in the conversation to increase the opportunity for the informant to reveal information that the analyst could not pre-identify. Additionally, the data-collection forms need to be flexible to ensure that data revealed by the subject of interest are captured. Open-ended data collection is often an aspect of initial or very assessments, and of long-term ethnographic studies, where the resulting analysis includes descriptive reporting and stories about experiences, perceptions and/or forecasts. Effective collection of data in an open-ended manner requires experienced data collectors with adequate knowledge of the subject and the information requirements; it also entails extensive note-taking and 'memoing'. Data analysis tends to be inductive and exploratory; it seeks to find patterns and concepts in the data that have been discovered, so to speak.

**Data collected in a semi-structured manner** falls somewhere between the two previous methods. A set rubric may be used for key issues to be addressed with some variables collected in a very structured manner and other matters dealt with in an open-ended way. Semi-structured data collection is one of the most commonly used methods for in-depth and rapid assessments, monitoring and evaluation exercises, and situation monitoring. Data analysis may involve a combination of deductive and inductive reasoning.

## INTEROPERABILITY

**Interoperability** is the ability to make two or more things operate together. The concept is usually associated with electronics and databases. For example, a Swiss electric plug does not fit into a British socket without the aid of a converter. Alone, the Swiss plug and British socket do not interoperate.

Here we extend the term 'interoperability' to the context of data, referring to those qualities of two or more pieces of data that enable them to be combined or compared for the purposes of in-depth analysis across space and time. In other words: *Do the data speak the same language?* Consider a hypothetical economic security assessment in which data on household income were being collected: once they were back in the office, researchers realized that some data were collected in the local currency and some in US dollars, and that the data collected did not specify the unit of measurement. The data could not be rectified without this 'attribute' – the unit of measurement – and making educated guesses would have compromised the quality of the data. In the end, the data were unusable.

To make data interoperable, it is essential to do the following: understand existing secondary data; impose controls on the collection of new primary data; and define analysis frameworks, indicators, data attributes and data-collection methods clearly. This will ensure that data are comparable in terms of their sources, and across space and time.

**INTEROPERABILITY AT MORE THAN ONE LEVEL**

Two different organizations may use the Food Consumption Score, an indicator developed by the WFP, to collect and analyse data on household food consumption. The two organizations may report their findings – categorized as 'poor', 'borderline' and 'good' food consumption – in terms of percentages of households. These data will be comparable only if the same thresholds were used to categorize household food consumption as 'poor', 'borderline' or 'good'" That being said, if the data were collected using the same time frame (seven-day recall) and sampling method, the original data may be combined to create a new analysis.

### CONTROLLING ATTRIBUTES

During the collection or sharing of data – primary and secondary – the attributes of the thing being measured should always be included. During the data-collection phase this is essential for ensuring that data are collected in the same manner. During the data sharing phase it is essential for understanding the data and how they can be used/compared. Numbers and certain words mean nothing without further description. Some examples of attributes are listed below:

- time frame (hourly, daily, 30-day, etc.)
- unit of observation (household, white sorghum, children under the age of five, etc.)
- unit of measurement (kg, cm, unit of currency, etc.).

When data are shared, attributes must also include information on the geographic location and time frame (when the data were collected, the geographic area they cover, etc.).

### CONTROLLING PROCEDURES

Data are particularly susceptible to error when they are not collected in a consistent manner and when predetermined procedures and/or tools are not used. This is particularly the case in the collection of structured data in a mass. For example, market prices may vary with the day of the week and even the time of the day. The procedures used to collect the data may not be reported in the final report, particularly in briefs, and users take them for granted; however these procedures are critical for ensuring data consistency and accuracy, particularly when comparing data across space and time.

The procedures might include:

- when data are collected (hour, day of week, month, season, etc.)
- where data are collected (from what source, one source or more, etc.)
- how data are collected (types of tool used, protocols for rounding fractions, etc.).
- details on the data collector(s) and informant(s).

**Table 7 - The example below shows the following: in the left-hand column, the results of analysis in four separate exercises; and in the two columns on the right, the attributes and procedures that *could have been* defined before data collection. This is an informed guess, so to speak, and just for presentation purposes, as complete details on whether these were actually controlled in the real study were not available. The data could have been collected without defining *all* of these attributes and methods; some are clearly defined as stated by the indicator itself or in the report.**

| ANALYSIS RESULTS | PREDEFINED ATTRIBUTE | PREDEFINED METHOD |
|---|---|---|
| The average monthly household expenditure is 3,100 KSH. | ■ Time (one month or 30-day average)<br>■ Unit of analysis (community)<br>■ Unit of observation (household)<br>■ Unit of measurement (Kenyan shilling) | ■ Households to be surveyed informed one day in advance to ensure that families are at home<br>■ Itemized list of daily cash expenditure over the last month<br>■ Itemized list of irregular cash expenditure over the last six months |
| The nominal retail price of 1 kg of white sorghum is 22,500 SOS. | ■ Unit of analysis (market)<br>■ Unit of observation (white sorghum)<br>■ Unit of measurement (1 kg, Somali shilling) | ■ Data collected on Saturday, main market day<br>■ Data collected between 8 and 10 a.m.<br>■ Data collected from same three traders and average taken |
| Each detainee has access to 300 g of porridge for breakfast. | ■ Unit of analysis (group of detainees)<br>■ Unit of observation (porridge/detainee)<br>■ Unit of measurement (g) | ■ Total volume of prepared food (cooked food) consumed in the morning during a regular meal |
| The main benefits of cattle ownership include milk production (34%), marriage payment (25%), manure (10%), compensation (9%), sales/income (7%), meat (6%), butter (3%), ploughing (3%), hides/skins (2%) and use in ceremonies (1%).[34] | ■ Unit of analysis (cattle owners in a given community)<br>■ Unit of observation (focus group of cattle owners) | ■ Data collected for a focus-group discussion with cattle owners<br>■ Proportional piling was used with 10 community groups |

34 Feinstein Group, 2014.

**LINKING DATA WITH P-CODES**

P-codes can be very important for ensuring interoperability between geographic, demographic and thematic data.

'P-code' is the abbreviated form of 'place code': a code name given to a unique geographic feature, such as a populated place (village, town, city, etc.) or an administrative unit (province, prefecture, state, commune, etc.). P-codes are normally designated by the government geographic or statistics office; if they are not readily available or up-to-date, they are sometimes developed by the UN or some other institution. The objective of the p-code is to provide a unique identifier for the geographic feature. You may ask, "Why not just use the name of the feature?" There are two main reasons:

1. The name is not always unique. Did you know there was a Paris, Kentucky, in the USA? A London in the province of Ontario in Canada?
2. A name may be spelt in many different ways, which can cause confusion when you are trying to harmonize lists of data that correspond to these locations; for instance, it can lead to something like this: "My data show 5,000 people living in Timbuktu and 4,000 in Tombouctou."

There is no global standard for p-codes; however they are normally made up of a series of letters and numbers. Below is an example from Sudan.

| State p-code | State name | Locality p-code | Locality name |
|---|---|---|---|
| SU06 | Jonglei | SU0601 | Altar |
| SU06 | Jonglei | SU0602 | Ayod |
| SU06 | Jonglei | SU0603 | Diror |
| … | … | … | … |

P-codes are often found in lists of population statistics lists or attached to GIS data on points of interest or administrative boundaries. If you don't know about the p-codes in your country, start by consulting with your local GIS officer or looking in the Humanitarian Data Exchange (https://data.hdx.rwlabs.org/), and follow up with the local government statistics or UN office if they are not readily available.

## PRIMARY-DATA-COLLECTION TOOLS

The choice of data-collection tool will depend on the type of variable, the depth of detail necessary, the data source and the data collector. All these factors are interdependent. For example, if the variable is plot size and measurement has to be precise (so that it can be used to write up a deed to a plot of land), then the piece of land (plot) in the deed will be the data source and the data collector will need to have the appropriate equipment and will also have to know how to accurately measure the size of the plot.

There are six types of tool that are commonly used to collect structured, semi-structured or open-ended primary data for humanitarian analysis, decision-making and programming.

**Table 8 - Tools commonly used for collecting primary data for humanitarian work.**

| TOOL | DESCRIPTION | EXAMPLES | S | SS | O |
|---|---|---|---|---|---|
| Registration | Baseline information, normally demographic, and/or minimal number of descriptive variables | ▪ Population census<br>▪ Beneficiary registration<br>▪ Hospital register | x | | |
| Self-administered questionnaire | Questionnaire completed by a respondent and returned | ▪ Feedback survey<br>▪ Polling survey | x | x | |
| Interview | Questionnaire that is completed during an interview with one or more respondents | ▪ Household/Individual questionnaire<br>▪ Key informant questionnaire | x | x | x |
| Group discussion | Discussion led by data collector with a group of relevant respondents | ▪ Community group discussion<br>▪ Focus-group discussion<br>▪ Participatory tools | x | x | x |
| Direct observation | Data gathered through observation of an element or phenomenon | ▪ Observation of damage<br>▪ Observation of type and status of household items<br>▪ … | | x | x |
| Measurement | Technical unit or instrument used to measure | ▪ Middle-upper arm circumference<br>▪ Plot size/Acreage<br>▪ … | x | | |
| Reporting | Analytical units report back regularly | ▪ Crowd-sourcing/seeding<br>▪ Eyewitness account<br>▪ Feedback hotline | x | x | x |

**S** = structured **SS** = semi-structured **O** = open-ended

## NOTE-TAKING AND MEMOING

Note-taking and memoing are methods used to record data collected in an open-ended manner: here, the data collector is himself or herself the tool. These methods are used in qualitative or mixed-method approaches to analysis, where the organization of data is defined by the data collector and/or analyst.

### CONTENT

Notes should always include detailed information on the informants and on the manner in which they were selected; they should also contain any other attributes that have to be taken into account and any limitations of the information provided.

The various sources of data should be clearly differentiated: informants, quotations, eyewitness accounts, direct observation and the data collector himself or herself. This is necessary in order to sort through the data later, and to use it correctly and to maximum advantage. For example, quotations can be a very powerful tool in reporting, and we need to be able to find them in our notes. There are a number of ways of distinguishing these sources from each other: through typography (all capital letters for key words, double inverted commas for direct quotations, etc.), by means of page layout (e.g. thoughts in margins) or by marking each data element (e.g. the letter 'C' can be placed before a comment to distinguish comments from recounts)

### ANALYTIC MEMOS

"Analytic memos are somewhat comparable to researcher journal entries or blogs – a place to 'dump your brain' about the participants, phenomenon, or process under investigation by thinking and thus writing and thus thinking even more about them (Saldaña, 2009)."

**Memoing** is a process for recording data collectors' observations and thoughts as they evolve over the course of the study. Memos capture data collectors' reflections on phenomena, processes, etc. as they occur, and can be incorporated later in the broader analysis of data. They can take the form of extensive marginal notes or comments, and may also serve to track the chain of thought that develops during the analysis of data.[35]

### ANALYTIC MEMOS: TOPICS FOR REFLECTION

The following is a list of topics to reflect on while collecting qualitative data (Saldaña, *Fundamentals of Qualitative Research*, 2011, p. 102[36]):

- how you personally relate to the participants and/or phenomenon
- your study's questions
- your code choices and their operational definitions
- the emergent patterns, categories, themes and concepts
- the possible networks (links, connections, overlaps, flows) among the codes, categories, themes and concepts
- an emergent or related existent theory
- any problems with the study
- any personal or ethical dilemmas about the study
- future directions for the study
- the analytic memos generated thus far
- the final report for the study.

## CODING

You can start developing codes in notes and memos (words or short phrases to classify qualitative data, with a view to discovering and reducing data to their minimum elements – including only the pieces of information needed for analysis). See Chapter 9 for a more detailed description of coding.

### EXAMPLE

Below is an example of a note-taking form. The structure helps the note-taker to differentiate between questions, responses and the data collector's observations (including analytic memo, initial codes, etc.).

The form might additionally include a space to differentiate between who the responses came from (small business owner, school teacher, etc.). This type of form could be used together with a checklist or discussion guide that highlights the topics to address in the discussion.

**GROUP DISCUSSION NOTE-TAKER FORM**

**A. BACKGROUND INFORMATION**

| | | | | | |
|---|---|---|---|---|---|
| 1 | Day \|__\|__\| Month \|__\|__\| Year \|__\|__\| | | 2 | Discussion no. | \|__\| |
| 3 | Moderator's name | | 4 | Note-taker's name | |
| 5 | Administrative unit | | 6 | City/town/village | |
| 7 | Community representative's name | | 8 | Community representative's contact information | |
| 9 | Number of participants | | \|__\|__\| men + \|__\|__\| women = \|__\|__\| total | | |
| 10 | Description of participants | | | | |
| 11 | How were the participants selected? | | | | |
| 12 | Start time | | 13 | End time | |

**B. NOTES**

| Question | Responses | Observations |
|---|---|---|
| | | |
| | | |

35 WFP, 2009.

36 By permission of Oxford University Press (www.oup.com). The material is restricted to viewing only and does not come under a Creative Commons license, or any other open access licence, that would allow reuse without requiring permission from OUP. For permissions to reuse, please contact academic.permissions@oup.com

## GUIDES AND CHECKLISTS

Discussion guides and checklists can take many forms. The main goal usually, is to outline the subjects to be addressed. Generally speaking, a discussion guide will list only the key topics to be addressed. It may also provide more specific guidance on the flow of discussion and pertinent details to address if they do not come up naturally in the conversation; in some instances it may provide structured tables for recording key figures if they are made available, such as people affected.

As with open-ended notes, guides are extremely useful for collecting qualitative information. Being a bit more structured, they should be explicit enough that the required information is captured however open enough as to not 'lead' the respondents in an assumed direction. Discussion guides should have enough space for note-taking and memoing, either in the margins or on separate pages. The method chosen should facilitate distinction between the reports from the informant and the data collector's analysis and thoughts and between the pieces of data provided by the various informants (in the case of group discussions), and should potentially already group data into different categories for analysis.

## STRUCTURED AND SEMI-STRUCTURED FORMS

This section looks at structured and semi-structured forms: self-administered questionnaires, for instance, and others – such as those used in registration, measurement, interviews and group discussions. They are grouped together, as many of the elements and concepts are the same. A form such as a questionnaire seeks to do two things primarily: maximize the number of responses and obtain accurate information that is analysable and interpretable. Well-designed forms can both optimize cost-effectiveness and ensure the quality and accuracy of data; but the subject of form design seldom gets as much attention as sample design or the development of data analysis software.

Each form is unique, and characterized mainly by the medium used and the unit of observation, and by whether it emphasizes structured or unstructured information. This section goes over some key concepts in form design. It focuses on semi-structured and structured methods – usually surveys - for collecting primary data. Certain parts of this section will be applicable only to certain types of form or certain mediums (e.g. a form on paper might be different from its electronic equivalent).

This chapter takes many examples from the Economic Security Assessment and Monitoring Data Collection Tools.[37] These tools are available on the "Data and Analysis" page at the EcoSec Resource Centre on the ICRC intranet (http://intranet.gva.icrc.priv/ecosec).

### CONTENT

The analysis plan is a starting point for determining the content of a questionnaire. Each element (piece of data or question) in the questionnaire should be of pertinence and capable of collecting accurate information. Those for whom the questionnaire is intended should be able and willing to respond and/or provide the necessary information.[38] The following table highlights some key points to **double check** when preparing and reviewing the content of a questionnaire. It should be comparatively uncomplicated to prepare a questionnaire on the basis of a well-designed analysis plan; and reviewing its contents in the light of the considerations listed below should be fairly routine.

37 ICRC EcoSec, April 2015.

38 Iarossi, 2006.

| WHEN DEVELOPING A QUESTIONNAIRE, THINK ABOUT… | | |
|---|---|---|
| Relevance | ▪ Is the element (piece of data or question) necessary and useful? Is it in my analysis plan? | ▪ *Do I need the age of each child or just the number of children under the age of 16?* |
| Accuracy | ▪ Is the element "double-barrelled"? Should it be broken up into two, or three, or four questions?<br>▪ Does the element need to be more specific?<br>▪ Does the element need to be more general? | ▪ *If I ask them their household income, will they include information on remittances and gifts?*<br>▪ *If I ask them their level of satisfaction with the food and non-food aid, will I know if they were satisfied with one and not the other?*<br>▪ *Do I need the number of children and adults in addition to the overall number of household members?*<br>▪ *If I ask about remittances over the last 30 days, will I have enough detail to understand their contribution to the overall household economy?* |
| Willingness and ability of respondents to talk | ▪ Can the respondents provide the information requested?<br>▪ Will the respondent answer truthfully?<br>▪ Is it a sensitive subject? If so, how should I approach it? | ▪ *Can members of the household recall how many sheep they owned before the crisis?*<br>▪ *Will they tell me how many heads of cattle they own?*<br>▪ *How do I ask them why they decided to leave their home? Will they talk about any challenges or discomfort they experience in the host community?* |
| Repetitiveness | ▪ Have I already asked this question in another way? Do I need to try and collect it in two different ways to ensure that I get it right, or will that become tediously repetitive?<br>▪ Are other organizations collecting the same data? | ▪ *If I ask them whether they received any remittances and if so how much, do I need to collect this data again under monthly income?*<br>▪ *Have other organizations collected data on remittance patterns?* |

## LENGTH

There is no consensus among social scientists about the optimal length of questionnaires, which are often criticized for their influence on response rate and data accuracy. Evidence on the effects of questionnaires on the rate of response is inconclusive; there is, however reason to believe that the length of questionnaires and the time taken to complete them influence the accuracy of data. Longer questionnaires put a burden on respondents, who become tired and perhaps bored. They may not refuse to answer questions, but may begin to give quick or random replies to get the interview over with. There is a general consensus that face-to-face interviews should last no longer than 45 to 60 minutes, and that interviews over the telephone and internet should be even briefer.[39]

Be alert to the possibility of flagging motivation and fatigue among respondents. With that in mind, review lengthy questionnaires for repetitiveness (*Am I asking the same question three different times? If yes, does it help in triangulation or contribute to confusion from data overload?)* and relevance (*Do I need all these details? Do I need them from this source?*).

39 Iarossi, 2006.

## WORDING

Studies show that the wording of questions has a direct influence on the responses to them. The designer should have some sense of the respondents' capacities and when framing the questions, try to imagine that he or she was the typical respondent, or the least educated among the respondents, who would have to reply to them. Knowledge of the context and of local languages is key. The objective is to have the answers reflect the actual situation or someone's interpretation of it as asked for; respondents should not have to *interpret* the questions put to them.

There are four criteria for framing questions that must be met. Questions should be brief, objective, simple and specific (BOSS).

| **FOUR CRITERIA FOR WORDING QUESTIONS** (*EXPLANATIONS TAKEN FROM IAROSSI, 2006).* | | |
|---|---|---|
| **BRIEF** | ▪ Keep questions as short as possible (without comprising the depth of detail in the response) and to the point<br>▪ Ask one question at a time | ▪ **Don't**: How much is your current household debt?<br>▪ **Do**: Does your household currently have any debt? If yes, how large is it? |
| **OBJECTIVE** | ▪ Avoid leading questions<br>▪ Avoid loaded questions with emotionally-charged words<br>▪ Be wary of making assumptions<br>▪ Be careful with pre-set multiple- choice lists – are they inclusive? | ▪ **Don't**: How many times in the last week did your household have to beg in order to have enough food?<br>▪ **Do**: In the last week, was your household in a situation where you did not have enough food or money to buy food? |
| **SIMPLE** | ▪ Use simple and direct words familiar to all<br>▪ Avoid technical jargon<br>▪ Use the same terminology throughout the questionnaire<br>▪ Avoid negative questions – replace with positive affirmation | ▪ **Don't**: What percentage of households in the community does not own a mobile phone?<br>▪ **Do**: What percentage of households in the community owns a mobile phone? |
| **SPECIFIC** | ▪ Be careful with general words (often, occasionally, etc.) and descriptions<br>▪ Avoid abbreviations<br>▪ Avoid compound or "double-barrelled" questions<br>▪ Ensure that the respondent won't have to work too hard to be as specific as requested<br>▪ Avoid hypothetical questions | ▪ **Don't**: How often does your household eat meat?<br>▪ **Do**: In the last seven days, on how many days did your household eat meat? Is this more or less than, or the same as, in an average week before the crisis? |

## SEQUENCE

The sequence and flow of questions should make it easy to complete the questionnaire and to keep the respondent engaged and comfortable. Special consideration should be given to the beginning of the questionnaire, the flow of subjects and questions and the placement of potentially sensitive questions.

### OPENING/INTRODUCTION

First impressions are extremely important. The introduction and first few questions shouldn't be too difficult for the respondents; in interviews, they can even be used to break the ice and set the stage for a good conversation. The data collector should identify himself or herself and the organization properly, state the objectives of the interview, and explain how the data will be used and how much time it will take to complete the questionnaire. The data collector may also provide a consent form (for particularly sensitive or personal data), and describe any expectations he or she may have of the respondent (in case a follow-up is expected).

### FLOW

The flow of subjects and questions should be from the general to the specific, and from the easy to the difficult. Respondents may feel more comfortable if the first few questions are open-ended and easy to answer; they may feel cornered if they are confronted right away by a series of questions requiring 'yes-or-no' answers. Information should be gathered in a way that seems natural and has a certain psychological logic; this will prevent data collectors and respondents from getting lost. One topic should be completed before starting another; shuttling between one subject and another should be avoided as far as possible.

**EXAMPLE**

Data on sources of food could be collected together with or immediately after data on food consumption, as this will be a natural or easy change of subject for respondents.

'Skip logic' is a useful tool for passing over irrelevant questions. For example, rather than asking, "How much is your current household debt" (a double-barrelled question), one could ask, "Does your household currently have any debt?" If the respondent says "Yes", then a follow-up question is asked, such as, "How much is your current household debt?" If the respondent says "No", then interviewers can skip to the next question.

### SENSITIVE QUESTIONS

Sensitive questions should be asked only if the data are absolutely necessary for analysis. They should be asked only after the respondent's faith – in the interviewer and in the objective of the interview – has been won. They should be introduced gradually, and preceded by a series of general questions on the topic.

**EXAMPLE**

When collecting data on IDP movement at the household level, the interviewer may wish to start with simple questions about where the family came from (usual place of residence) and when they arrived at their current location, before discussing why they had to leave their homes and if they plan to return.

## LAYOUT

### INSTRUCTIONS AND GUIDANCE

Every form should have explicit instructions in its use, at the beginning and throughout the questionnaire. The complexity of the questionnaire will determine the amount of instructional material required. The following should be covered:

- **introduction:** introducing oneself, describing the objectives of the exercise are, explaining how the information will be used, etc.;
- **selecting respondents:** choosing sites, selecting sampling units such as households or individuals, criteria for key informants, composition of focus groups, etc.;
- **using the form**: following the order of questions exactly as listed, follow-up questions (depending on the answers), allowing the interview to follow its own course as long as all topics are touched upon, etc.

**Figure 6 - The following example from a household registration exercise clearly defines a 'family' to ensure that all data collectors use the same definition. This is critical in cases where some data collectors consider extended family members and others do not.**

| REGISTRATION APPLICATION FORM | | | |
|---|---|---|---|
| **1. GENERAL INFORMATION** | | | |
| 1.01 | Data Collector Name | | |
| 1.02 | Registration Site | Raion: \|_____\| | Location: \|_____\| |
| 1.03 | Registration Date | Day \|__\|__\| Month\|__\|__\| Year \|__\|__\| | |
| **2. FAMILY PROFILE** | | | |
| *This section is about the **immediate family**. A family, for purposes of this form, is a husband and spouse (if there is a spouse) and all children under 18 years old.* | | | |

Instructions for the questions might cover the following:

- **asking questions:** where the questions are listed as multiple-choice, it might be better sometimes to put them in an open-ended way, and to leave it up to the data collector to simply select the most appropriate response; in some cases the questions are not meant to be put to the respondent, but to be answered by the data collector via direct observation;
- **the period of reference for an answer:** household income over the last six months, rice sales during the 2013-2014 season, or before the crisis in 2005, etc.;
- **units of measurement** for any quantitative data where that is not obvious (household income in Sudanese pounds, the cost of one litre of milk, rice production in kg, etc.;
- **precise definition of categories:** because these may be open to interpretation (who is considered a member of the household, what age range is used for children, what type of 'group' is being referred to, etc.

**Figure 7 - The following example explains to the data collector that he or she should only count those articles that are in working condition; thus the overall data set will include only working items. This fits the objective of the question: to know what the household owns and can use as a proxy for their living conditions.**

| 1.0 | **Fill the table with the number of items the household has of each article.** *Do not count articles which do not work or are in such poor shape that they are not/should not be used.* | | | |
|---|---|---|---|---|
| | **Article** | **Number** | **Article** | **Number** |
| | **01 – Blankets** | \|__\|__\| | **07 – Wash basin** | \|__\|__\| |
| | **02 – Clothing sets** | \|__\|__\| sets | **08 – 10-litre jerrycan** | \|__\|__\| |
| | **03 – Mosquito nets** | \|__\|__\| | **09 – 20-litre jerrycan** | \|__\|__\| |
| | **04 – Plastic sheeting** | \|__\|__\| | **10 – Mobile phone** | \|__\|__\| |
| | **05 – Cooking pots** | \|__\|__\| | **11 – Radio set** | \|__\|__\| |
| | **06 – Eating utensils** | \|__\|__\| sets | **12 – TV set** | \|__\|__\| |

## SECTIONS AND NUMBERING

Questions may be organized into sections, each with its own set of instructions if relevant (e.g. the questions in this section should be put only to livestock-farming households). Questions and pages should be numbered so that the data collectors can easily follow where they are in the process.

**Figure 8 - Example of questionnaire with clearly marked sections and question numbers.**

1. GENERAL INFORMATION

| 1.1 | **Interviewer's Name** | | | |
|---|---|---|---|---|
| 1.2 | **Village/Town** | \|_____\| | **1.3 Date** | Day \|__\|__\| Month\|__\|__\| Year \|__\|__\| |

2. HOUSEHOLD DEMOGRAPHICS

| 2.1 | **What is the household's status?** *Select one.* | | |
|---|---|---|---|
| | **IDP** \|__\| | **Returnee** \|__\| | **Resident** \|__\| |
| 2.2 | **What is the size of the household (number of household members)?** | | \|__\|__\| |
| 2.3 | **How many household members were working/employed before the shock?** | | \|__\|__\| |
| 2.4 | **How many household members are working/employed now?** | | \|__\|__\| |

## VISUALS

Visuals, such as bolded fonts, underlining and 'shapes', will be of great help to interviewers as they make their way through the form. For questions that refer to specific periods of time, units, etc. bolding or underlining the font can help to remind the data collector not to forget to refer to the given period, unit, etc.. Arrows are commonly used to direct 'skip logic'.

## SPACE

Sufficient space should be available for writing down comments, notes, etc. Multiple-choice questions should always include an extra line for 'other', with enough space for writing down the name of the category that was not pre-listed.

### UNIQUE ID

A unique ID is a unique identity for each form: this can be a number or a character, or a combination of the two. When data are collected and entered electronically, unique IDs are usually created automatically by the application being used. If paper or Excel is the medium, a unique ID should be created by the user. The only requirement is that each form should have an ID unique to it. For example, if you have 100 structured questionnaires, you should have 100 unique IDs.

IDs should not contain 'special' characters – +,",*,ç,%,&,,",*,ç,%,#, etc. – as they can be misread by data-processing software. IDs can include letters, numbers, dashes ( –) and underscores ( _ ).

Unique IDs can:

1. establish a link between the data-collection form and the database (critical in paper data collection);
2. facilitate archiving and reference back to paper forms (much like using a catalogue of the books in a library or sorting through the records of prisoners of war in the ICRC's archives in Geneva); and
3. be used as in analyses to count the number of unique records and calculate response rates.

**SAMPLE UNIQUE ID METHOD**

The best practice is to have a unique code on each questionnaire. If that is not feasible when the questionnaires are printed, one easy solution is to take a number of elements already collected in the questionnaire and combine them in a way that makes each questionnaire different from all the others.

For example, let us assume that the introductory section contains information on a) the data collector's initials, b) the name of the site and c) the number (first, second, third, fourth, etc.) of the questionnaire collected by the data collector at that site; the unique ID could be developed by using those three elements, as their combination could never be the same (because there can be only one instance of that data collector collecting that xth questionnaire at that site).

See the two examples below.

| Data collector 1 (SM) at site Ningerum | Data collector 2 (TP) at site Ningerum |
|---|---|
| Ningerum1SM | Ningerum1TP |
| Ningerum2SM | Ningerum2TP |
| Ningerum3SM | Ningerum3TP |
| … | … |

## TYPES OF QUESTION

### BACKGROUND

The introductory section of a form usually contains background information that can be completed before the interview starts. It should include information on the data collector, data, and location of the interview and the respondents.

**Figure 9 - Example of an introductory section that can be filled out before the interview.**

*This section should be completed before the interview to save time.*

| | | | | | |
|---|---|---|---|---|---|
| **1.10** | Day \|__\|__\| Month\|__\|__\| Year \|__\|__\| | | **1.20** | **Interviewer's name** | |
| **1.30** | **State** | | **1.40** | **Payam** | |
| **1.50** | **Boma** | | **1.60** | **Village** | |
| **1.70** | **Team number** | \|__\|__\|__\|__\| | **1.80** | **Household number** | \|__\|__\|__\|__\| |

**TIP:** You are strongly urged to keep track of data collectors and their data-collection forms (that is, you should know, in every case, who collected the data in a particular form). This will be useful at the data processing and analysis stage, because it will be much easier then to contact the right person in case any questions arise in connection with the material in the forms.

## DEMOGRAPHICS

Demographic data are data on a given population: population statistics, gender, age, ethnicity, etc. A section on demographic data may have to be included in the data-collection form or questionnaire when:

- **population figures or estimates need to be collected** (e.g. community questionnaire or group discussion) and possibly disaggregated by type (ethnic group, gender, age, etc.); or
- **data need to be disaggregated** (household or individual questionnaire, registration form, monitoring form, etc.) by gender, age, displacement status, number of people in the household, sex of head of household, age, level of education, etc.

**Figure 10 - Example of disaggregating figures on household members.**

| 1.0 | **Members of the household** *A household, for the purposes of this exercise, consists of a group of people living and eating together and sharing the same resources. If the household is hosting another family and they are sharing the same resources, count them here. Do not count members who are not currently in the city/town/village.* | | | |
|---|---|---|---|---|
| **Total** | **1 - Children** (<5) | **2 – Children** (5-17) | **3 – Adults** (18-65) | **4 – Elderly** (>65) |
| \|__\|__\| | \|__\| | \|__\| | \|__\| | \|__\| |

The analysis plan should indicate if this information is required.

## MULTIPLE-CHOICE QUESTIONS

Multiple-choice questions are questions with two or more possible answers, and may be limited to either one response among the possible responses or multiple responses (i.e. more than one response) within the possible responses.

- **The choice of answers should be made very carefully, in consultation** with experienced field staff, staff members with local knowledge, and in line with secondary data, both past and present (in order to permit cross-comparison).
- **Each choice should be unique**, and clearly expressed in the local language, with little or no room for misinterpretation.
- Questions for which there may be responses other than those listed on the form should always come with at least one (and sometimes two) **"other" option(s)"** for responses.
- The instructions should say whether the interviewer **should read the list of options to the interviewee or whether he or she should ask the question in an open-ended way** and then choose the most appropriate choice himself or herself.

**Figure 11 - Question 1.0 below is a multiple-choice question with more than one possible response; question 1.1 has only one possible response**

| 1.0 | **Through direct observation (do not ask): Does the structure have any apparent damage or significant wear/tear to any of the following?** *(more than one response possible)* | \|_\| Roof<br>\|_\| Walls<br>\|_\| Flooring<br>\|_\| Doors<br>\|_\| Windows<br>\|_\| No damage | 1.1 | **Through direct observation (do not ask): What is the type of structure?** | \|__\| | A = Mud house, grass roof<br>B = Mud house, tin roof<br>C = Brick house<br>D = Cement house<br>E = Other, specify<br>_____ |
|---|---|---|---|---|---|---|

**SHOULD I 'CODE' MY CHOICES?**

Coding, here, is a method of replacing a verbal response with a 'code', usually either a number or a letter. In paper-based data-collection forms, codes are extremely useful when response options are many and wordy; they save time and spare data collectors from having to write out an entire response. However, this is an option only for data collectors who are at ease with using codes. In mobile data-collection tools, codes (or identifiers) are built in so that the data collector never sees them. Using codes in the final database can be essential for wordy/verbose responses, where it can be a challenge when writing formulas to search for certain responses – in Excel, for example, when using SEARCH or MATCH.

## NUMBER QUESTIONS

Number questions are questions that seek numerical answers that can be recorded, as integers or decimals. The number in the question can be a specific unit of measurement or a percentage. Numbers can be reported individually or as part of a larger series, often represented in a table. Number questions and numerical tables should always specify the unit of measurement. It may be useful to employ a data-entry format in which each digit in the number has its own space. This might make it easier to read the various data collectors' handwriting in paper questionnaires.

**Figure 12 - Example of dedicated spaces for digits.**

| | | | | |
|---|---|---|---|---|
| 1.0 | **Estimate the weekly household expenditure over the last one month in South Sudanese pounds.** *If some kind of expenditure are not included in the list, add them under points 10-12 until the estimation for total monthly expenditure is complete. Include only daily/weekly expenditure. Longer term expenditure will be captured under question 2.0.* | | | |
| | **Item** | **Expenditure** (SSP) | **Item** | **Expenditure** (SSP) |
| | **01 - Food** (including oil, sugar and salt) | \|__\|__\|__\|__\| | **07 - Firewood/Charcoal/Fuel for cooking** | \|__\|__\|__\|__\| |
| | **02 - Coffee/Tea** | \|__\|__\|__\|__\| | **08 - Communication** | \|__\|__\|__\|__\| |
| | **03 – Drinks** (water, soda) | \|__\|__\|__\|__\| | **09 - Transportation** | \|__\|__\|__\|__\| |
| | **04 – Soap** | \|__\|__\|__\|__\| | **10 -** _____ | \|__\|__\|__\|__\| |
| | **05 - Clothes/Shoes** | \|__\|__\|__\|__\| | **11 -** _____ | \|__\|__\|__\|__\| |
| | **06 - Milling and grinding** | \|__\|__\|__\|__\| | **12 -** _____ | \|__\|__\|__\|__\| |
| 1.1 | **Total weekly expenditure last one month** | | \|__\|__\|__\|__\|__\|__\| SSP | |

Digital data-collection forms (such as those used on laptops, handheld mobile device or over the Web) can incorporate controls for numbers (limit minimum and maximum values, limit numbers to integers or decimals, etc.). This can be extremely useful in quality control of data.

## MEASUREMENT QUESTIONS

Scale and measurement questions may use ordinal, interval or ratio measurement levels. **Ordinal measurements** rank values: 1st, 2nd, 3rd, 4th; high, medium, low; and so on. They are often used in questions about levels of satisfaction, preferences, sources (primary or secondary), etc.

**Figure 13 - Example of an ordinal scale to capture level of satisfaction.**

| | |
|---|---|
| 7.0 | **What was the households overall satisfaction with the FOOD delivered to them?** |
| | \|_\| Very satisfied \|_\| Satisfied \|_\| Somewhat satisfied \|_\| Not satisfied \|_\| Disappointed |

**Interval measurements** are used to classify information in specific ranges. They are a useful way to record figures that are difficult to express with precision or figures that all the respondents may not be able to provide with confidence. For example, interval measurements of age may be written like this: between 0 and 10, 11 and 20, 21 and 30, 31 and 40, 41 and 50.

**Figure 14 - Example of an interval scale; age range.**

| | | | |
|---|---|---|---|
| 4.0 | **Head of household's age** | \|__\| | **A** = <=17 yrs.<br>**B** = 18-60 yrs.<br>**C** = >60 yrs. |

**DESCRIPTIVE WORDS FOR MEASUREMENT** (majority, a lot, very, etc.) run the risk of being interpreted differently by the data collector and the respondent. The use of descriptive language should be given careful thought while drafting the analysis plan, when designing the questionnaire, and during the analysis of the results. For example, let us assume four categories for measuring the extent of damage to a home: 'none', 'moderate', 'severe' and 'total destruction'. One could differentiate between 'moderate' and 'severe' by saying that the former meant damage to windows and doors and that the latter would include damage to walls and the roof. This should be recorded in the questionnaire, highlighted during the data collectors' training, and tested before the questionnaire is used.

## SKIP LOGIC

Skip logic can be used to guide the interviewer to ask only questions that are pertinent, i.e. questions that apply to a given respondent. For example, if the respondent answers 'yes' to a question, there may be a follow-up question; otherwise, the interviewer will move on to the next question.

**Figure 15** - **Example of skip logic.**

| | | | | |
|---|---|---|---|---|
| 1.0 | **Does the household currently have any debt?** | | \|_\| Y \|_\| N *If N, skip this section.* | |
| | 1.1 | **If yes, what is the total debt?** | \|__\|__\|__\|__\|__\|__\| SSP | |
| | 1.2 | **…from who?** *(More than one source possible)* | \|_\| Friends/Family<br>\|_\| Hum./Dev. organisation<br>\|_\| Local shop/boutique<br>\|_\| Cooperative | \|_\| Private lender<br>\|_\| Bank /Credit union<br>\|_\| No response<br>\|_\| Other, specify |
| | 1.3 | **…and why?** *(More than one reason possible)* | \|_\| Buy food<br>\|_\| Buy household items<br>\|_\| Pay health fees<br>\|_\| Pay school fees<br>\|_\| Inputs for IGA | \|_\| Pay for labour<br>\|_\| Pay rent/Utilities<br>\|_\| For ceremony costs<br>\|_\| No response<br>\|_\| Other, specify |

Another example of skip logic would be a series of questions that are asked only if the respondent falls into a certain category: IDPs, returnees, refugees, residents, livestock owners, etc.

Skip logic is extremely useful for expediting data collection; however, clear instructions on questionnaires (for paper) and well-coded rules (for electronic data collection) are needed to ensure that data collectors use it properly.

## CASCADING SELECTS

'Cascading selects' are unique to electronic data collection/entry. 'Cascading' is a feature that enables researchers or analysts to alter the optional answers to one question based on the response to another. For example, if the first question is about a country, and the choices are Sudan, Ethiopia and Eritrea, and the second is about a city, the choice of cities can be limited to only those cities in the country selected (Sudan, Ethiopia *or* Eritrea).

In order to use cascading selects, proper lists must be prepared in advance and entered into the system.

**Figure 16** - **Example of a format for listing administrative units in Lebanon that can be used for a cascading select question. Columns J and K are a unique list of second administrative units. Column H and I are the first administrative units associated with the second ad administrative unit. The list in this format allows any data entry or database to look to provide response options to one question based on the selectio of a previous question. Say for example, Mount Lebanon was selected as the governorate (first administrative unit); then, the list of optional districts (second administrative unit) would include only Aley, Baabda,Chouf, El Meten, Jbeil and Kesrwane.**

| Adm1_Name_EN_Adm2Linked_List | Adm1_Name_AR_Adm2Linked_List | Adm2_Name_EN_List | Adm2_Name_AR_List |
|---|---|---|---|
| Akkar | عكار | Akkar | عكار |
| Baalbek-El Hermel | بعلبك - الهرمل | Baalbek | بعلبك |
| Baalbek-El Hermel | بعلبك - الهرمل | El Hermel | الهرمل |
| Beirut | بيروت | Beirut | بيروت |
| Bekaa | البقاع | Rachaya | راشيا |
| Bekaa | البقاع | West Bekaa | البقاع الغربي |
| Bekaa | البقاع | Zahle | زحلة |
| El Nabatieh | النبطية | Bent Jbeil | بنت جبيل |
| El Nabatieh | النبطية | El Nabatieh | النبطية |
| El Nabatieh | النبطية | Hasbaya | حاصبيا |
| El Nabatieh | النبطية | Marjeyoun | مرجعيون |
| Mount Lebanon | جبل لبنان | Aley | عاليه |
| Mount Lebanon | جبل لبنان | Baabda | بعبدا |
| Mount Lebanon | جبل لبنان | Chouf | الشوف |
| Mount Lebanon | جبل لبنان | El Meten | المتن |
| Mount Lebanon | جبل لبنان | Jbeil | جبيل |
| Mount Lebanon | جبل لبنان | Kesrwane | كسروان |

### OPEN-ENDED QUESTIONS

Open-ended questions usually return responses in the form of text; however, sometimes responses may be a combination of text, numbers, images, etc. Open-ended questions are crucial for collecting qualitative information. They should be explicit enough to ensure the collection of the required information, but also open-ended enough to not move the respondents in some predetermined direction.

As with multiple-choice questions, interviewers should be given instructions on posing questions and guiding discussions. In most cases, interviewers **should not read out questions in exactly the way they appear in the questionnaire;** they must ask questions or guide the discussion in a manner that is most suitable, in the circumstances, for collecting the information necessary (they should have discussed this and had some practice in it before going to the field).

Open-ended questions should be broad enough to enable data collectors to collect all the information they need, and should at the same time ensure that the data collector's notes are comprehensive, and specific enough to meet the information requirements. The objective is to minimize the risk of vague and therefore useless responses.

**Figure 17 - Example of semi-structured open-ended question.**

| 2.0 | Would the household refuse any type of aid delivered to them? | | \|_\| Y \|_\| N |
|---|---|---|---|
| | 2.1 | IF YES, what? | …and why? |
| | | | |

**OPEN OR CLOSED-ENDED QUESTIONS?**

Open-ended questions have the advantage of not limiting responses, thereby leaving open the possibility of a deeper than expected understanding of the issues; but, they may also provoke responses that are ambiguous and uninterpretable.

Closed-ended questions (such as multiple-choice questions), on the other hand, help to return information that is useful and can be analysed; but, the responses to them may not be as rich and may be of limited value when the choices for respondents are not well formulated.

These considerations must be weighed when developing the questionnaire. It may be useful to work with open-ended questions at first, and then to develop suitable closed-ended questions based on the responses to the original questions. Closed-ended questions should be reviewed from time to time. In face-to-face interviews, the options for responses need not be read out in every instance; in fact, they should not be on most occasions. Questions may be asked in an open-ended manner, after which it is up to the data collector to select the most suitable option or to indicate "other" when a response is not among the predetermined options listed on the form. This is an area in which data collectors must be trained.

### COMPARISON BEFORE AND AFTER

When baseline data are lacking, unreliable or in a format not usable on the scale of the primary-data collection exercise (too broad, too specific, covering a different sample, imcomplete, etc.), it may be necessary to collect them during the exercise. In emergency assessments and programme monitoring, even if baseline data do exist, it may be useful to collect before-and-after data to gauge from respondents whether or how things had changed after a shock or an ICRC project/programme.

This can be done quantitatively (exact figures or estimates in the form of numbers or interval measurements) for quantitative data, or qualitatively (ordinal measurements or by means of a less-structured qualitative method) for quantitative and qualitative data.

**Figure 18** - Quantitative comparison

| 1 - Food group | 2- Food items | # of households out of 100 with access to food item | | 5 - If changed, reasons for change |
|---|---|---|---|---|
| | | 3- Before shock | 4 - Now | |
| Cereals | | / 100 | / 100 | |
| | | / 100 | / 100 | |
| | | / 100 | / 100 | |
| | | / 100 | / 100 | |
| **Pulses** (legumes, nuts and seeds) | | / 100 | / 100 | |
| | | / 100 | / 100 | |

**Figure 19** - Qualitative comparison

| | | | |
|---|---|---|---|
| 1.0 | Did the household income change after the shock? | | \|_\| Y \|_\| N |
| | 1.1 | IF YES, how? | \|_\| Increased<br>\|_\| Decreased |

Quantitative measurements can be more precise when comparing quantitative data, but they can also be less reliable for a number of reasons: the circumstances, the availability of historical records and the sheer difficulty for respondents of remembering details with the required accuracy. Qualitative information is less precise and inherently subjective, but when participants are unable to remember details precisely it may, in fact, be more reliable. This should be thought through while planning the analysis.

## RECORDING OBSERVATIONS

Observations are the result of seeing, listening and smelling – each thing by itself or many things together – and of questions asked and answered; what they do is to create an impressionistic account of a situation. Observations can be structured (observing the number, and the physical state, of animals bought at market over a given period of time) or unstructured (e.g. watching and listening and constantly).

## DESCRIBING DATA QUALITY

Notes on data quality are useful when collecting data that may be inaccurate, unreliable or biased. The quality of data can be judged by either objective or subjective means.

Objective descriptions can be used when collecting secondary measurements of quantitative data: population figures, production output, etc. This entails collecting information to establish whether data are based on precise measurements or on estimates, which can then be taken into account when using the secondary data.

**Figure 20** - Objective assessment of data quality

| | | | | |
|---|---|---|---|---|
| 1.0 | **How many people are currently living in the community?** *Include as much detail as possible. Do not count people/households who are not currently in the city/town/village.* | | | |
| | Type | | Individuals | Households |
| | 01 | Residents | \|__\|,\|__\|__\|__\|,\|__\|__\|__\| | \|__\|,\|__\|__\|__\|,\|__\|__\|__\| |
| | 02 | Returnees | \|__\|,\|__\|__\|__\|,\|__\|__\|__\| | \|__\|,\|__\|__\|__\|,\|__\|__\|__\| |
| | 03 | IDPs | \|__\|,\|__\|__\|__\|,\|__\|__\|__\| | \|__\|,\|__\|__\|__\|,\|__\|__\|__\| |
| | 04 | Refugees | \|__\|,\|__\|__\|__\|,\|__\|__\|__\| | \|__\|,\|__\|__\|__\|,\|__\|__\|__\| |
| | 05 | _____ | \|__\|,\|__\|__\|__\|,\|__\|__\|__\| | \|__\|,\|__\|__\|__\|,\|__\|__\|__\| |
| 1.1 | | Total population | \|__\|,\|__\|__\|__\|,\|__\|__\|__\| | \|__\|,\|__\|__\|__\|,\|__\|__\|__\| |
| 1.2 | | These figures are based on… | | \|_\| estimates \|_\| census \|_\| don't know |

Subjective descriptions are used by data collectors to report their impression of the accuracy or reliability of the information. They may be particularly useful for dealing with large data sets, where it is difficult to remember each and every case. That said, it is still best practice for data-collection teams and analysts to go over such impressions at the end of each day, while they are still fresh in the minds of the data collectors.

**Figure 21 - Subjective assessment of data**

| 6.14 | For the data collector, how confident are you about the data on expenditure? |
|---|---|
| | \|_\| Very confident \|_\| Confident \|_\| Somewhat confident \|_\| Not confident \|_\| Don't know |

These impressions can be influenced by various factors:

- how the respondents answered the questions (Did they have to think about their answers a long time? Did they hem and haw?);
- social customs or norms (e.g. Is it customary for people in the area to share this type of information?); and
- the difficulty of remembering certain pieces of information – productive output, income, etc. – without the help of documents or records.

### PARTICIPATORY TOOLS

Participatory tools are an effective method of gathering information, particularly when respondents are not accustomed to structured surveys. Commonly used tools include:

- drawings of timelines, calendars and maps
- proportional piling
- ranking and scoring
- the why-why tree
- stakeholder analysis matrix.

These tools may not fit every situation or every community, and should be used only after careful thought. Each tool or technique will require a particular method for data gathering (interviews or group discussions); and data may be recorded on its own form or part of a questionnaire, discussion guide or checklist. Examples of these tools are available in the EcoSec handbook, *Assessing Economic Security* (ICRC, 2016).

## PRIMARY-DATA-COLLECTION MEDIUMS

Primary data can be collected on various mediums, the most common of which are paper forms, laptop computers and handheld mobile devices; they can also be collected by telephone, via short-message service (SMS) and through the internet.

### PAPER

Paper is the most commonly used medium, because it is relatively inexpensive and flexible – notes can be added, questions reordered, etc. – and because, unlike electronic data-collection tools, it does not require technical skills to develop the questionnaire itself. That said, electronic data-collection tools are getting easier and easier to use. Paper forms are useful in situations where data requirements are not fully defined (initial rapid assessments), where a lot of qualitative data are required or where it may be inappropriate or impractical to carry laptop computers or handheld devices. Paper forms can also be used as a backup or in addition to other mediums.

**USES FOR PAPER -** Registration, self-administered questionnaires, interviews, observation, measurement and reporting

**PROS**
- Speedy data collection
- Flexible
- Requires few technical skills

**CONS**
- Slow and costly data entry/processing
- Greater risk of error in data entry/processing, because of lack of measures to approve or verify accuracy of responses, ensure legibility or make the asking of certain questions compulsory
- Greater risk of data going astray if paper archiving system is not in place
- Need for additional tools – GPS, audio recording devices, cameras, etc.

## LAPTOP COMPUTERS

Laptop computers can be used for direct electronic data entry, by means of word-processing documents or spreadsheets, or via forms developed by data-processing software such as MS Access, CSPro, Sphinx, and SPSS.

**USES FOR LAPTOPS -** Registration, interviews, observation and measurement

**PROS**
- Data collection and entry take place together, enabling speedy processing
- Computers are sturdy and able to house a variety of software
- Synchronization of data on the fly when network available

**CONS**
- Less portable than paper or handheld device
- May need designated station or, frequently, power source (for charging)
- May be inappropriate in dangerous or poverty-stricken areas
- Depending on the laptop, may need additional tools to enable GPS, take pictures, etc.

## HANDHELD MOBILE DEVICES

Handheld mobile devices – smartphones, tablets etc. – can be used in conjunction with data- collection software such as Open Data Kit (ODK), KoBo Toolbox and Device Magic. Mobile devices are useful in data collection on the go – for example, for data collectors going from house to house.

**USES FOR HANDHELD MOBILE DEVICES -** Registration, self-administered questionnaires, interviews, observation, measurement and reporting

**PROS**
- Data collection and entry phase take place together, enabling speedy processing
- Data quality can be enforced through the use of skip logic, validation rules and mandatory questions
- Data quality can be monitored through time stamps and automated GPS coordinates
- Can carry a number of different applications useful in field work (text, photo, GPS, etc.)
- Synchronization of data on the fly when network available

**CONS**
- Unsuitable for handling large amounts of qualitative or open-ended information
- Need power source frequently, for charging
- May be inappropriate in dangerous or poverty-stricken areas
- Some areas may have restrictions on the use of cameras or GPS devices

## TELEPHONE

Data can be collected over the telephone in certain circumstances: where an unusually large number of people have to be interviewed for data collection or when access is limited, or for an initial survey screening. Whatever the case, the data collected would then have to be recorded on a data-collection form via paper, laptop computer or electronic mobile device.

**USES FOR TELEPHONES -** Self-administered questionnaires, interviews and reporting

**PROS**

- Speedy data collection, because field movement not required

**CONS**

- Informants required to have functioning telephones
- Difficulty of verifying data and data source
- Interactions may be limited (10-15 minutes)
- Response rates can be lower than for in-person interviews

## SHORT-MESSAGING SERVICE (SMS)

SMS can be used not only to communicate information (one-way), but also to establish dialogue with informants and collect information (two-way). It is particularly suitable for collecting small amounts of data; data aggregators such as FrontlineSMS and RapidPro, and crowdsourcing platforms such as Ushahidi, can then be used to facilitate efficient data processing and analysis.

**USES FOR SMS -** Self-administered questionnaires, reporting and crowdsourcing/crowdseeding

**PROS**

- Data collection is instantaneous
- Medium expands coverage of information providers

**CONS**

- Informants required to have mobile devices capable of sending SMS messages
- Difficulty of verifying data and data sources
- Because not all network providers encrypt SMS messages when they are sent, data security can be an issue
- Data limited to 160 characters, and number of interactions (in case of two-way) may also be limited
- Response rates can be lower than for in-person interviews

## INTERNET

Internet surveys such as SurveyMonkey are commonly used by people with access to the internet, and who process and centralize data on the fly. Surveys are normally carried out via email or publicly, on a website, for instance.

**USES FOR THE INTERNET -** Self-administered questionnaires and reporting

**PROS**

- Data collection is instantaneous
- Medium expands coverage of information providers

**CONS**

- Informants need to have access to computers and the internet
- Difficulty of verifying data and data sources

### RECORDING DEVICES

Recording devices such as tape recorders and video cameras are used to ensure the integrity of data, which, collected in this way, can be transcribed later for review. They are particularly useful when members of the analysis team are unable to attend interviews, and when they need full transparent access to discussions.

The consent of participants is required for recording; and a means of transcription should be taken into account while budgeting time and resources.

| **USES FOR RECORDING DEVICES** – Interviews and group discussions | |
|---|---|
| **PROS** | **CONS** |
| ▪ Medium insures data integrity | ▪ Some participants may not wish to be recorded<br>▪ Transcription is time-consuming and costs money |

| **CHOOSING THE RIGHT MEDIUM:** SOME CONSIDERATIONS: | |
|---|---|
| **Data** | *What is the most appropriate method for collecting the type of data required? Compare the medium with the type of form required. Can a smartphone be used to enter data from 1 0 open-ended questions? Is a paper form suitable for recording rapid quantitative measurements?* |
| **Perceptions** | *How might the sources, and people in the community, feel about the medium being considered?* |
| **Data collectors and informants** | *What are data collectors or informants most likely to feel comfortable with? Would that also be most effective medium? What is required to change that?* |
| **Cost** | *How much does the medium cost? Can we afford it?* |
| **Time** | *How quickly can data be collected and processed with the medium, given the time we have (or wish to spend on this)?* |
| **Quality** | *How effective is the medium in on-the-fly quality control and prevention of error?* |
| **Flexibility** | *How flexible is the medium? Can it cope with changing circumstances? Do we need a backup?* |

## PUTTING IT INTO PRACTICE

### DATA-COLLECTION TEAM

Data collectors should be carefully selected. Only these people should be chosen: those whose neutrality, in connection with the situation and with incoming information, can be relied upon; who have sufficient local knowledge; who know enough about the information to be collected; and who are motivated. Gender and ethnicity should be given careful consideration.

Ideally, one person should conduct the interview or facilitate the group discussion, and another person (or two) should enter data and/or take notes (but this may not always be feasible). The interviewer or facilitator will then be able to focus on the informants and the subject matter, maintaining eye contact and projecting ease and friendliness; this increases the chances of him or her being perceived as a human presence and not as someone who there solely to collect data. The note-taker(s) can concentrate on recording information correctly and jotting down memos.

## TRAINING AND PRE-TEST

Data-collection methods, mediums and tools should always be tested and reviewed before there are put fully into practice. Data collectors conducting interviews and group discussions should be trained and their training should, ideally, cover the following areas:

- ✓ describing the objectives of the study
- ✓ describing the organization(s) behind the study (design, implementation and funding)
- ✓ communicating the expectations of the data collectors
- ✓ sampling logic and selection of respondents
- ✓ walk through of the tools with explanation of the objective and of each question
- ✓ a role-playing exercise that simulates the interview
- ✓ mitigating bias and dealing with non-response
- ✓ communicating with the respondent and obtaining his or her consent (when applicable).

## CONDUCTING INTERVIEWS

Conducting interviews is a skill, and acquired through experience. The main goals of an interview are to engage the participant, elicit a response and collect accurate data. The following section touches lightly on some key concepts as they relate to the ICRC.[40]

### INTRODUCTION

Any interview should always begin with an introduction: the data collector should introduce himself or herself and the study and its objectives. Any agreement on data use, including permission for or formal consent to the use of data in analysis, may also be made at this stage; in addition, the interviewee should be informed about his or her rights with regard to the ICRC, which is the controller of the data to be collected. The interviewer should then set the stage for the interview. How that is to be done will depend on the type of data to be collected (measurement, specific information that has to be recollected, etc.) and/or on the respondents (children, highly educated people, less educated people, etc.). The circumstances and the setting should, of course, also be taken into account: cultural usage, local customs, shocks recently incurred, etc.

"At the beginning of any interview, the respondent is not sure what role he or she is supposed to play. The respondent is not clear if he or she has to play a task-oriented role in which he or she is required to respond adequately and accurately, or a conversational role in which he or she tries to relate to the interviewer by conforming to the interviewer's apparent opinions or by attempting to make a good impression on the interviewer. During the brief social interaction of an interview, the style employed by the interviewer has an impact on how the respondent perceives his or her role and, consequently, this has an impact on the accuracy of the data collected" (Iarossi, 2006).

If the interviewer's manner is formal and impersonal, the respondent is likely to realize that he or she is required to provide accurate information without too much commentary. A more formal interviewing style may also be appropriate for educated or highly educated respondents.

A less formal and more personal style might encourage respondents to provide additional information and comments, respond with their own questions, or engage in an extended discussion. This interviewing style may yield more qualitative data and/or a higher response rate. It may also be more suitable for a less educated respondent. There is however a risk of laxity: misled by the interviewer's manner, the respondent may neglect to provide accurate details.

40 Iarossi, 2006.

Research has shown the importance of participants' interest in the subject and of their motivation to participate. "Many forces motivate people to participate in a survey: an interest in the topic, a desire to be helpful, a belief of [sic] the importance of the survey, a feeling of duty (…) Other forces influence people to refuse: difficulty in understanding the questions, fear of strangers, the feeling of one's time being wasted, difficulty in recalling information and embarrassment at personal questions." (Plateck, Pierre-Pierre and Stevens, 1985.)

The choice of style or combination of styles will depend on the circumstances. The interviewer must act as a 'neutral communicator', but he or she also has to convince the respondent to participate. Our recommendation is to employ a personal style at the outset, to secure the respondent's participation; and to explain to informants/interviewees if or when the proceedings are likely to become more formal.

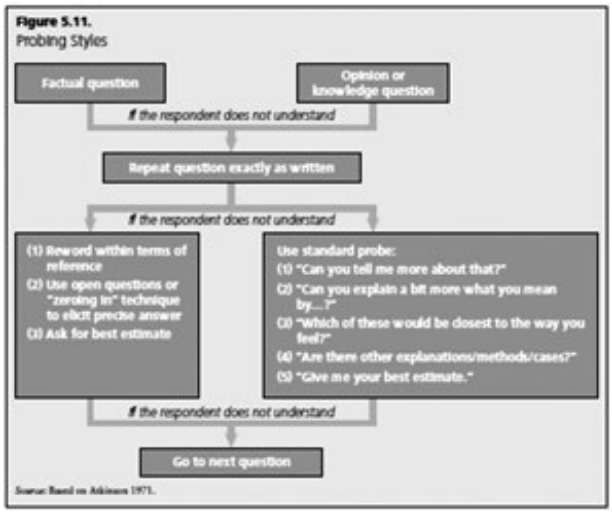
## ASKING QUESTIONS

The approach taken can be either rigid (e.g. questions are read out exactly as they appear on the form) or flexible (e.g. questions are worded according to the data-collector's sense of their suitability). How questions are to be asked will depend, in every case, on the data-collection methods and tools employed, and on the context. For example, an open-ended interview is likely to entail a flexible style of questioning; a structured questionnaire may require the interviewer to adopt a more rigid manner.

Techniques can be considered and discussed in-country before data collection. In the case of surveys, this can be part of training and pre-testing. Participants with experience in the context should share what they have learnt. When the questions are reviewed, care should be taken to ensure that all of them can be understood and interpreted in the same way by both interviewers and respondents, and perhaps alternatives suggested to lessen the possibility of misinterpretation.

## PROBING

When respondents do not understand a question or give an incomplete, irrelevant or obviously inaccurate response, interviewers may want to follow up or reformulate the question. 'Probing' means asking a respondent to provide more information or to clarify their response. If this is beyond the respondent's ability, probing may be irrelevant. The difficulty is in intuiting when probing may be relevant (e.g. when responses are inaccurate or incomplete); interviewers must also take care not to be too direct or pose leading questions when they probe, and that, too, can be difficult. The table below[41] shows two styles of probing, associated with factual questions and opinion questions.




41 Source: Atkinston, 1971. Taken from Iarossi, 2006.

### PROMPTING

**Prompting,** which means suggesting possible answers (e.g. those in a list of responses to multiple- choice questions), is a technique used to guide respondents in a particular direction. Prompting should be avoided unless specifically required by the data-collection method.

| GUIDELINES FOR INTERVIEWERS | |
|---|---|
| **DOS** | **DON'TS** |
| ■ Establish eye contact with the respondent<br>■ Stick to the questions as they are written, and use pre-formulated alternatives if needed*<br>■ Follow the order of questions set out in the instructions*,**<br>■ Ask every question (even if you think you know the response)<br>■ Be a good listener, and show patience<br>■ Be gentle: probe, encourage, elaborate or ask for clarification if the respondent is unable to understand a question or provide a sufficiently detailed<br>■ Use tools (tape measure, camera, maps, drawing pads, etc.) and props (seed, balls, etc.) to collect information | ■ Spend the whole time reading from the form<br>■ Change questions<br>■ Shuttle back and forth between sections (you might miss questions)<br>■ Skip questions<br>■ Finish the respondent's sentences<br>■ Force answers if the respondent is not forthcoming |
| *Not applicable to open-ended interviews<br>**Less applicable to semi-structured interviews | |

## OUTCOMES AND LESSONS LEARNT

Because humanitarian work often takes place in the same locations, tools used and lessons learnt should be shared, archived and readily accessible for future use; teams coming afterwards will then have a wealth of previous experiences to draw on.

**Data storage** here refers to the place where data are consolidated for eventual treatment and analysis. For structured data (collected in a structured manner or later given a particular structure) that is stored digitally, that place is usually a database. Unstructured data may be stored electronically in word documents, or on paper in file cabinets or archives. Data should be stored only for as long as they are needed.[42]

This section focuses on developing very simple electronic databases for structured data collected on paper forms with few relational features. MS Excel 2010 is used in all the examples shown, as the objective is to provide a simple solution when more advanced technologies are not available. This is not relevant when data are collected electronically on mobile data-collection forms, Web-based forms or more advanced databases. The main concepts can, however, be applied to other data and statistical software packages. Digital archiving and sharing are covered in Chapter 6: **Data treatment**.

42 See Article 6, "Retention, destruction, and archiving of data that are no longer needed", in *ICRC Rules on Personal Data Protection*, January 2016.

## DATABASE

A **database** is an organized collection of data.[43] Databases can be either relational or non-relational.

**Non-relational databases** are two-dimensional arrays of data (normally in the form of a single table) that can be developed with most data-manipulation tools like MS Excel, or statistical software packages such as SAS, SPSS or Sphinx.

**Relational databases** are databases housing a collection of tables and elements, each having some relation to the other through a constant, a variable or a set of criteria. Basic relational databases can be developed in MS Excel and statistical software packages, in which relationships are for the most part managed manually. More complex relational databases, with many tables, relationships and types of relationship, may require more robust software, such as MS Access, MS SQL or Oracle, which have advanced functionality for defining, storing and managing relationships.

**KEY PRINCIPLES IN STRUCTURED DATA STORAGE[44]**

- Organize your data to support your analysis
- Minimize the extent to which you have to analyse across tables (disconnected data)
- Store pieces of information together when they have the same data structure

## DATA MODEL

The first step in building a database is to construct the data model, or the plan of the house where all the data will be stored. It is a model of the structure and elements of the database including tables, rows, columns, data values and documentation about the database. There are four key steps in building the data model:

1. defining tables, including rows and columns
2. defining data values
3. optimizing the model for data entry
4. developing documentation.

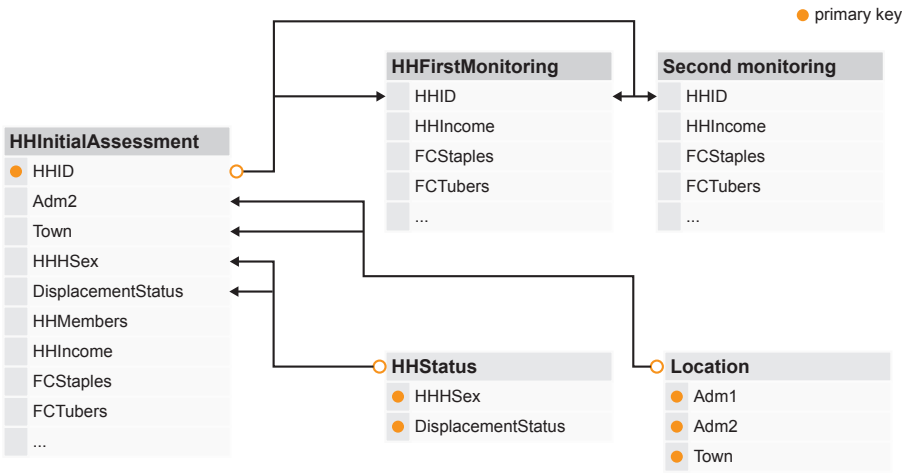
Tables are the various layers in which data will be stored. The first step is to determine what kind of tables, and how many, will be needed. When using Excel, it may be prudent to keep the number of tables to a minimum; this will help to prevent errors associated with manual management of relationships between tables.

A simple database normally includes a data table and one or more domain tables, which can be linked together through a unique Name (as it is referred to in Excel). The graphic below shows three data tables (HHInitialAssessment, HHFirstMonitoring, HHSecondMonitoring) linked through a common, unique household identifier (HHID) and two domain tables (HHStatus, Location), which are baseline tables that feed directly into the data tables. Each table consists of a list of variables.

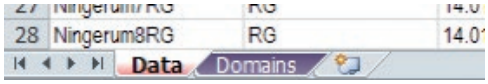
43 Wikipedia, Wikipedia entry on "Database", accessed in April 2015.

44 ACAPS, August 2013.

**Figure 22 - Simplified example of a database schema and linkages between data and domain tables.**

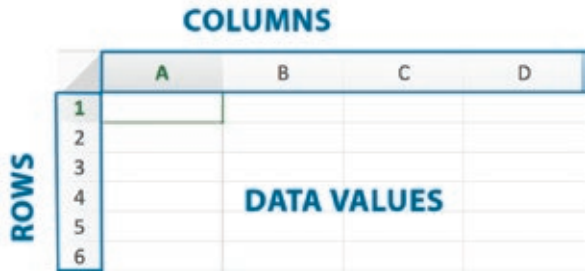



In Excel, each data table is represented by a sheet within a workbook; domain tables may be stored together in one or more sheets, depending on the number of domain lists.




## DATA TABLES

A **data table** is a table in which data records are stored. Each data table is defined by its rows, columns and data values.




*How many tables should I use?* Ideally, in simple databases – such as those developed in MS Excel – there should be one table for all data, because that will limit errors associated with managing relationships and linking data and records manually. Relational databases, such as those developed in MS Access or other database software, may have many tables. Each table is then linked through a common variable or unique ID, and through a defined relationship.

### ROWS

**Each row after the column header row(s) represents one record**, corresponding to one response gathered by your data-collection tool. For example, if your tool was a household survey, and you surveyed 200 households, each row will correspond to one household. The first row, or sometimes the first few rows, will represent the column headers.

### COLUMNS

**Each column represents one database field**, corresponding to one discrete unit of information, such as a constant or variable. In some cases, one column corresponds to the response to one question in your data-collection form (e.g. sex of head of household); other cases will require more than one column (e.g. multiple-choice questions with several possible responses).

### RECORD IDENTIFICATION

The first column of any data table should contain the unique ID (unique identifier or name for each record, which can be a number or character, or a combination of the two), enabling cross-referencing between records (rows) in the database and the original data-collection forms, and making it possible for you to trace information back to the original form.

The unique ID may have been created as part of the data-collection tool (which will be discussed in Chapter 5: Sampling). If that was not done, the unique ID should be created at this stage. In the case of paper data-collection forms, the unique ID should be added to the data-collection form to ensure the link.

### COLUMN HEADERS

The 'column headers' are the unique names for each variable or constant. Each column header should be unique, as that will facilitate cross-reference with the data-collection form, and cross-tabulation at the analysis stage. Depending on the complexity of the data-collection form, column headers may take up to four rows for the following:

- **Section name** – The name of that section of the data-collection form where the question can be found (Introduction, Demographics, Food production, etc.). A section name may span several columns, to include all the variables in a given section.
- **Variable long name** – A longer description of the variable, one that is easily understood (e.g. Sex of head of household).
- **Variable short name** – A short name for the variable, one that may cause puzzlement without the accompanying long name (e.g. if the short name for sex of head of household is HHHSex), but one that is useful at the analysis phase in cross-tabulation and table/graphic development. Short names should not contain any spaces or special characters – +,",*,ç,%,&,,",*,ç,%,#, etc. – as they can be misread by data-processing software. They can be made up of letters, numbers, dashes ( - ) and underscores ( _ ).
- **Data value type** – This can be useful for indicating the type of data (number, text, date, domain, etc.) that are found in a given column, and for guiding data entry and database users. A few types of data value are listed below[45]:

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1. INTRODUCTION | | | | | | 2. DEMOGRAPHICS | | | | | 3. LIVELIHOOD STRATE |
| 2 | UniqueID | Enumerator | Date | District | Place | Questionnaire # | Head of HH Sex | Head of HH Marital Status | HH Status | Adult Members | Child Members | Agriculture |
| 3 | UniqueID | Enumerator | Date | District | Place | Quest | HHHSex | HHHMarital | HHStatus | MemAdult | MemChild | LSAgriculture |
| 4 | Ningerum1ND | ND | 14.01.14 | North Fly | Ningerum | 1 | Female | Widow | Resident | 1 | 2 | x |
| 5 | Ningerum2ND | ND | 14.01.14 | North Fly | Ningerum | 2 | Male | Married | Resident | 2 | 4 | |
| 6 | Ningerum3ND | ND | 14.01.14 | North Fly | Ningerum | 3 | Male | Married | Resident | 2 | 2 | |
| 7 | Ningerum4ND | ND | 14.01.14 | North Fly | Ningerum | 4 | Male | Married | Resident | 2 | 3 | |
| 8 | Ningerum5ND | ND | 14.01.14 | North Fly | Ningerum | 5 | Male | Widow | Returnee | 1 | 2 | |
| 9 | Ningerum6ND | ND | 14.01.14 | North Fly | Ningerum | 6 | Male | Married | Resident | 3 | 4 | |
| 10 | Ningerum7ND | ND | 14.01.14 | North Fly | Ningerum | 7 | Male | Married | Resident | 2 | 2 | |

45 Note that in statistical packages such as SPSS, the data type is built into the database.

**NOTE ON MULTIPLE-CHOICE QUESTIONS**

Multiple choice questions can either be restricted to one response or allow for a response to be selected from multiple options. Single responses need only one column in the database; however, multiple responses will require one column for each possible response.

For example, consider the question below:

| 3.1 | What is the household's principal livelihood activity? | *Check all that apply* |
|---|---|---|
| | Agriculture ☐ Livestock/poultry ☐ Fishing ☐ Mining ☐ Commerce ☐ | |

A column should be created for each possible response (Agriculture, Livestock/Poultry, etc.) as shown below:

| 3. LIVELIHOOD STRATEGIES | | | | |
|---|---|---|---|---|
| Agriculture | Livestock/poultry | Fishing | Mining | Commerce |
| Dichotomous | Dichotomous | Dichotomous | Dichotomous | Dichotomous |
| LSAgriculture | LSLivestock | LSFish | LSMining | LSCommerce |
| 1 | | | | 1 |
| | | | 1 | |
| | | | 1 | 1 |
| | | | 1 | |
| | | | 1 | |

## DATA VALUES

Data will be entered in the blank spaces. Controls should be used to limit and ease data entry. Controls are implemented by column (variable) and may include limits on the range of numbers, number of characters, drop-down lists, date formats, etc. These will be addressed in the next two sections: Domains and Data values.

## DOMAINS

A **domain** in the database sense is a defined and discrete list of possible responses. It is essentially a standard set of predefined responses, such as those used for multiple-choice questions/drop-down menus. Examples of domains include lists of geographic locations, household type and sex.

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | PLACEList | | HHHSexList | | HHHMarStatusList | | HHStatusList |
| 2 | Koroba | | Male | | Married | | Resident |
| 3 | Tari | | Female | | Single | | Displaced |
| 4 | Telefomin | | | | Divorced | | Returnee |
| 5 | Kiunga | | | | Widow | | |
| 6 | Ningerum | | | | | | |
| 7 | Tabubil | | | | | | |
| 8 | | | | | | | |

| DOMAINS SHOULD BE… | WHY? |
|---|---|
| Managed separately from the data table | ▪ To avoid confusion with the data<br>▪ To enable use of one domain by more than one variable |
| Easily accessible and discoverable | ▪ To facilitate their management by non-developers<br>▪ For easy reference back during analysis |
| Organized to support data entry and analysis | ▪ To minimize error by maximizing efficiency<br>▪ To enable use of advanced relationships (cascading drop-down lists, FIND and MATCH functions, etc.) |

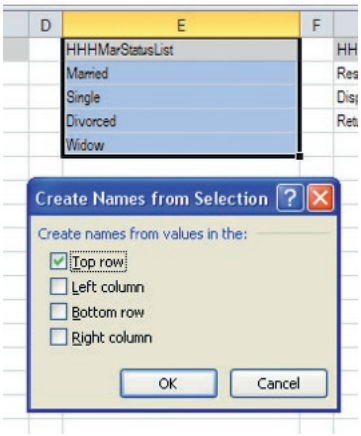
## CREATING A DOMAIN LIST

Defined Names is a function in Excel for managing unique cells and ranges of cells that need to be 'called' frequently by other functions – from simple calculations (SUM, AVERAGE, etc.) to more advanced tasks (MATCH, Data Validation, etc.). It simply replaces the standard name for the cell (e.g. A4) or range of cells (e.g. A4:A12) with a pertinent 'name'.

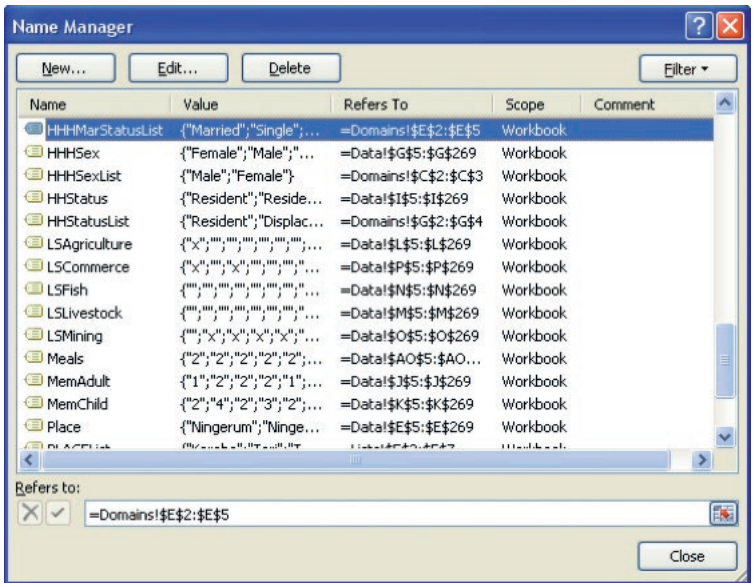
The advantage of working with Named Ranges in Excel is that it facilitates database upkeep: for instance, changes to any formula in the list that refers to that 'Name' are automatically updated.

To create a 'Name' for a list:

- navigate to the list that you would like to 'name';
- at the top of the list, insert a row with the name that will be used for the list;
- highlight the entire row, including the title and the list of elements;
- navigate to the tab Formulas on the main menu panel, and under Defined Names, select Create from Selection;
- check 'Top row' and select OK; and



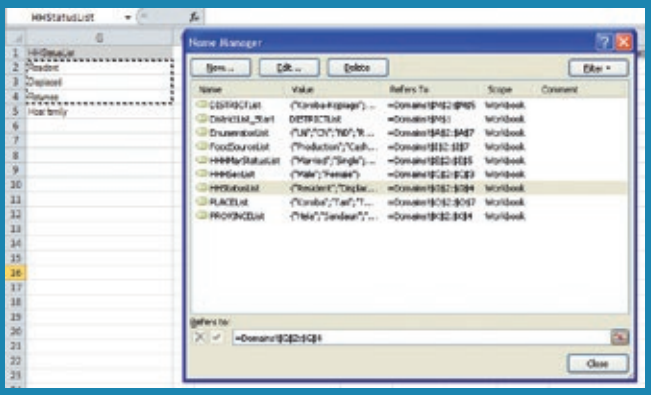

- check your work by navigating to Formulas > Names Manager and confirming that the list of names that were created and the associated ranges are in the Excel Workbook.



*If I change an item in my Named Range list, are my drop-downs that refer to this list automatically upated?* Yes, if an item is changed in the list, it is automatically reflected in the drop-downs. For example, if the list stated 'Displaced' and you change 'Displaced' in the list to, say, 'IDP', the drop-down list will automatically changed. However, a data-entry record would not be changed automatically. That needs to be updated manually.

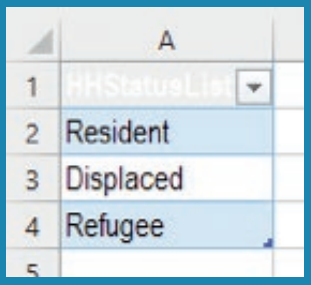
*If I add an item to my list, is the 'Name' list automatically updated?* No, the list is not automatically updated. You need to tell Excel that there are more elements in the list. For example, say you have a list of household status that contains the terms 'Resident', 'Displaced' and 'Returnee', and you want to add 'Host family':

- first, you add 'Host family' to the bottom of the list;
- then, navigate to Formulas > Names Manager;
- then highlight the household status list; and
- under 'Refers to' you can then change the reference to include another row.

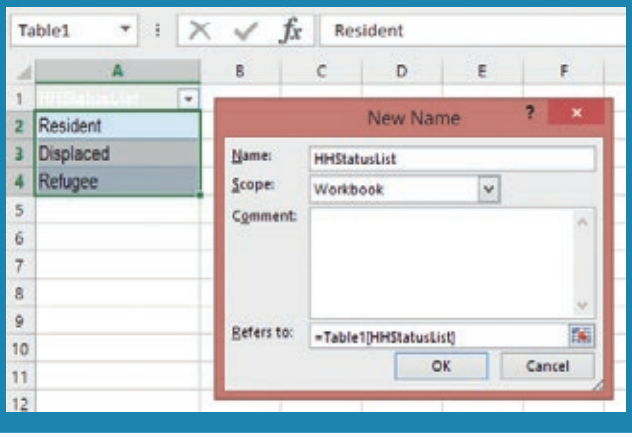


A better option for lists that will need to be updated frequently is to use a self-expanding table, called a Table list in Excel, and feed it into the 'name'. That can be done in the following steps:

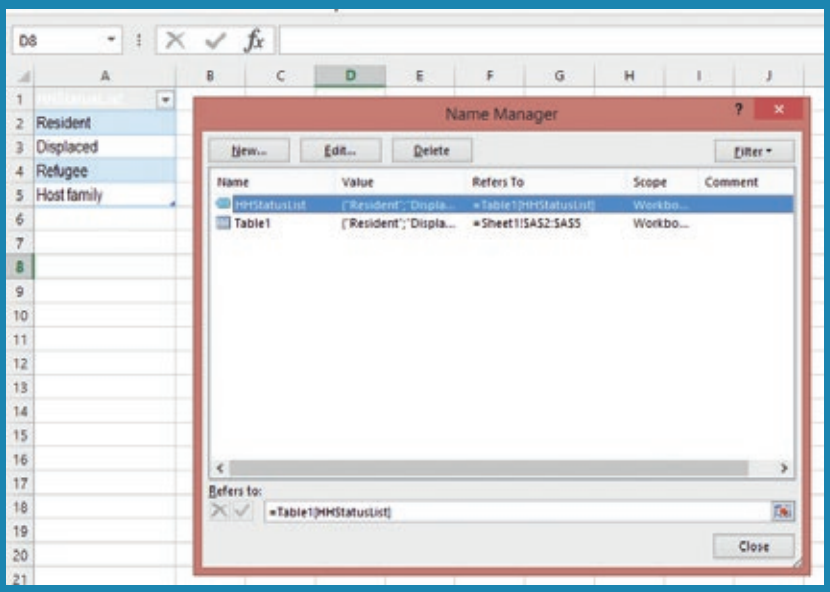
- highlight the entire list;
- then navigate to Insert > Table;
- select 'My table has headers' if the table has a title;
- now you will see your table highlighted in alternating colours;




- now you need to create a 'Name' for your table by navigating to Formulas > Defined Names > Define Name;
- create a 'Name' for your list under "Name:" and insert reference to all items in the list under "Refers to:";



- now the table and associated 'Name' will automatically update any additions to the list;
- if you select Formulas > Name Manager, you will see your table and the 'name' in the list. You can call the name in any formula required (e.g. data validation for drop-down list).




**TO CODE OR NOT TO CODE?**

In the context of domains, coding – it should not to be confused with .NET, HTML or other coding languages – is a method of replacing a text response with a "code", usually either a number or a letter.

Whether or not to code will be determined by the users of the database (will it facilitate their work or create more errors?) and the analytical requirements of each variable. There are two situations in which coding may be considered:

1. It can be a useful replacement for wordy options. This is a common practice in advanced relational databases. In manual databases, however, such as those built in MS Excel, translating text to code, and then back to text from code, can be time-consuming and creates room for error.

2. It can be useful to replace dichotomous variables (yes/no) with the code '1' for positive (or yes) and '0' for negative (or no). A simple count function can then be used during the analysis phase to calculate the total positive (yes) and negative (no) responses.

## DATA VALUES

### CREATE UNIQUE IDENTIFIER

One easy way to create a unique ID for each record is to join together elements in the data set that, when combined, make them unique, and unlike all the other elements. In Excel, the CONCATENATE or & function can be used to do this automatically, if it was not done at the data-collection phase.

- In the example below, the first column is used to create the unique ID by taking elements from column D (Place name), E (Questionnaire number) and B (Data collector initials).
- Starting from A4, the following formula is used: =CONCATENATE(F4;G4;B4) or =F4&G4&B4.[46]
- The formula joins values from column D (Place name), E (Questionnaire number) and B (Data collector initials) and returns the results in column A.

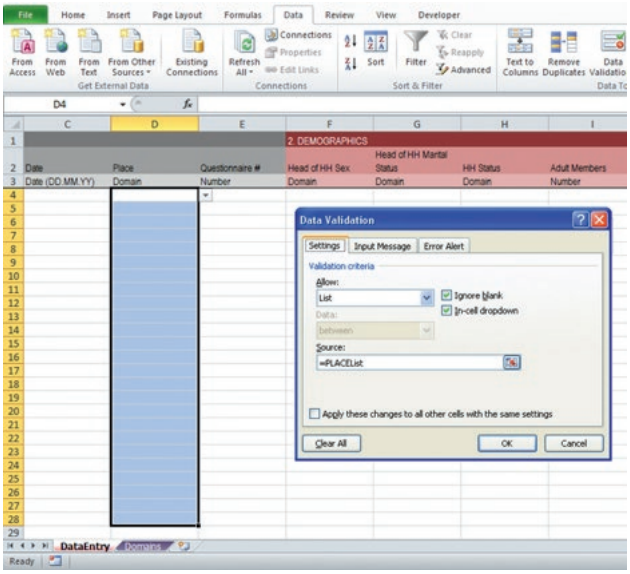
46 Depending on the regional setting of Excel, formulas may use a semicolon (;) as shown in this guidance or a comma (,). Check which 'list separator' is used under the Regional and Language options of the PC.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | 1. INTRODUCTION | | | | |
| 2 | UniqueID | Enumerator | Date | Place | Questionnaire # |
| 3 | Automatic | Domain | Date (DD.MM.YY) | Domain | Number |
| 4 | Kimbe1CN | CN | 02.02.14 | Kimbe | 1 |
| 5 | Kimbe2CN | CN | 02.02.14 | Kimbe | 2 |
| 6 | Kimbe3CN | CN | 02.02.14 | Kimbe | 3 |
| 7 | Kimbe4CN | CN | 02.02.14 | Kimbe | 4 |
| 8 | Kimbe5CN | CN | 02.02.14 | Kimbe | 5 |
| 9 | | | | | |

## DROP-DOWN LISTS

Drop-down lists support data entry by controlling and in turn validating data coming into the database. Creating drop-down lists involves calling the range of options in the relevant domain table and returning them as options in the data field. In Excel, this can be done by using the Data Validation function.

- Highlight the cell(s) that should have a specific drop-down list.
- Navigate to Data > Data Validation and under Allow: select List, and under Source: type "=" and the 'Name' of the list created when setting up the database.




## CASCADING DROP-DOWN LISTS

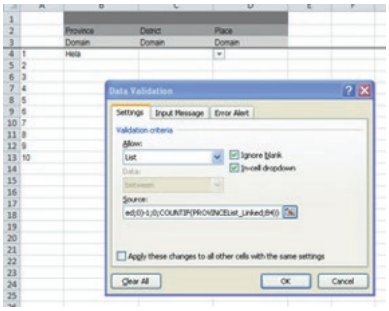
Cascading drop-down lists can be used to limit drop-down lists in one cell to the value in another cell. For example, if you want the data-entry staff to select a village from a long list of villages, you may first want them to select the region, after which the drop-down list will show only the list of villages in that region. This can be done in Excel by using the OFFSET, MATCH and COUNTIF functions together with Data Validation. The following are the steps to be taken to create a cascading drop-down list for geographic locations (provinces, districts and locations):

- First, you need to set up the lists and give them 'names'. In the first column of an empty spreadsheet, create a list called PROVINCEList, and give the unique list of provinces. Create a Name for the list by highlighting the list and navigating to Formulas > Defined Names > Create from Selection; use the top row as the name of the list.
- In column D, create a list with the unique list of Districts and create a Name for the list, using the same methodology as in step 1. In column C, create a second list called PROVINCEList_Linked, and indicate in the list the provinces associated with each district in the list in column D. Create a Name for this list.
- Follow the same process for Places and create a linked district list.
- The result should look like the table below:

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | PROVINCEList | | PROVINCEList_Linked | DISTRICTList | | DISTRICTList_Linked | PLACEList |
| 2 | Hela | | Hela | Koroba-Kopiago | | Koroba-Kopiago | Koroba |
| 3 | Sandaun | | Hela | Tari-Pori | | Tari-Pori | Tari |
| 4 | Western | | Sandaun | Telefomin | | Telefomin | Telefomin |
| 5 | | | Western | North Fly | | North Fly | Kiunga |
| 6 | | | | | | North Fly | Ningerum |
| 7 | | | | | | North Fly | Tabubil |

- Create a new tab for data entry, and three columns: Province, District and Place.
- In the first (Province), create a standard drop-down list for Provinces by highlighting the first ten rows in the column and navigating to Data > Data Validation.
- In the data validation wizard, set Allow: to List and under Source: type =PROVINCEList.
- Highlight the first ten rows under District and go to Data > Data Validation. Set Allow: to List and under Source: type in the following formula:

**=OFFSET(DISTRICTList;MATCH(B4;PROVINCEList_Linked;0)-1;0;COUNTIF(PROVINCEList_Linked;B4))**




- Select OK and verify the result by selecting a province in the first column (Province), and then checking the drop-down list in the second column (District).
- Now highlight the first ten rows under Place and go to Data > Data Validation. Set Allow: to List and under Source: type in the following formula:

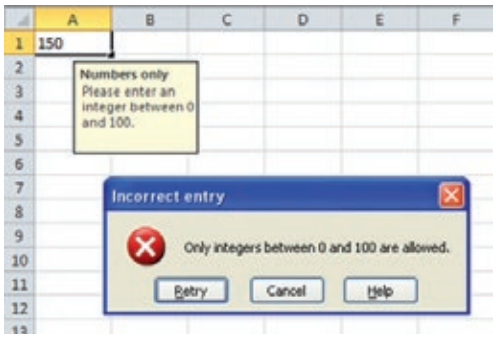
**=OFFSET(PLACEList;MATCH(C4;DISTRICTList_Linked;0)-1;0;COUNTIF(DISTRICTList_Linked;C4))**

## DATA-VALIDATION RULES

Data-validation rules are rules added to a data-entry form to guide users in entering the correct type of data in the correct format, and to prevent other data from being entered. For example, you may set a rule that an entry must be a number; if someone then tries to make an entry in the form of text, it will not be accepted.

Every electronic data-entry system sets up data validation in its own way. In Excel, Data Validation can be used to control the entering of data, and even to add messages when incorrect entries are made. To do this, go to Data > Data Validation and use the Settings tab to create the rule, Input Message tab to add a message to guide data-entry staff in what should be entered in the cell, and Error Alert to establish an alert message when an incorrect entry is made.

**Figure 23 - Example of data-entry message (top left next to cell), and error alert (bottom right) when incorrect data are entered.**




All database users should be made aware of all data-validation rules before they start entering data. This is in order to ensure that they understand why the rules are there, and to ensure also that they know what to do if a rule does not fit a response in the questionnaire.

**'HOLEY' DATA**

Have you ever encountered a data set containing a mass of blank values and zeros, or of strange symbols, such as !%$ and #N/A? And then wondered what it all meant? A good database should establish rules that are known and understood by everyone using it: the developer, the data-entry person, people searching the database, and the analyst. While developing the database, and the data-entry form, make sure that you give thought to how the following will be treated in the database and in the analysis:

| | | |
|---|---|---|
| **Non-response** | Respondent did not respond to the question | Establish a code (-9999, n/r, etc.) or leave blank. Importance of differentiation will depend on the need to distinguish between refusal to respond and poor data-collection practices. |
| **Non-applicable** | Not applicable to the record (individual, household, etc.) | Establish a code to distinguish between blank values (-9998, n/a, etc.) if necessary for analysis. These can also be filtered out by using another variable. |
| **Zero values** | Value of zero (0 children, 0 cars, etc.) | Indicate 0 and make sure that 'non-response' and 'non-applicable' are not recorded as 'zero'; otherwise, the descriptive statistics will be erroneous. |

## PROTECT DATA FORM

Data-entry forms that are to be used by many people, or over a period of time, should establish controls for modifying the form (type of question, wording of question, multiple-entry choices, number of questions, comment boxes, etc.). Any modification that is required should be done by a central user, who will replicate it in all forms in use. This facilitates merging or regrouping of data collected by many different people.

In Excel, a data form may be protected on two levels: at the individual sheet level and at the workbook level. Data-sheet protection can be used to prevent columns or rows from being added, or any changes from being made to the description of the variable. What this means is that it will then be possible only to enter data; formatting changes or other modifications to the sheet cannot be made. The following method can be used to protect a data sheet:

- Select the range of cells where users can enter data, then right click and select Format Cells. Select the Protection tab and click the checkbox to deselect the Locked option.
- Navigate to the Review Tab and select Protect Sheet.
- Check only 'Protect worksheet and contents of locked cells' in the main menu and 'Select unlocked cells' in the sub-menu.
- Type in a password for the sheet and click OK.
- Do the same for the other sheets in the workbook to prevent users from making changes.
- Workbook protection can be used to prevent users from modifying the names of sheets or moving sheets around. To protect a workbook, navigate to the Review Tab and select Protect Workbook; then check Structure to prevent users from rearranging, deleting, renaming, etc. sheets in the workbook.

Always remember to record the password somewhere so that you don't forget it and/or it can be found by future users working on the database that need to unlock it to make modifications.

*How do I deal with verbose qualitative data?* Verbose qualitative data – i.e. wordy responses -such as that collected from open-ended questions, can be difficult to handle with programmes more suited to short qualitative entries or quantitative data (such as Excel and SPSS). Here are a few rules of thumb to ensure that data are fully exploited during the analysis phase:

**DO NOT** break the link between the record and its data. It may be important to compare verbose qualitative information with other data collected from the same person, household, etc. for deeper understanding or triangulation. The unique ID can be used to keep this link even if verbose data are stored separately from other data.

**DO NOT** always assume that verbose data should not be entered into the database, because they can be used later in techniques for analysing qualitative data (discussed in Chapter 9: Combining Analyses and Drawing Conclusions).

**DO** keep only pertinent information. In many cases, it will not be necessary to transfer all details to the data-entry form: for instance, paragraphs and full sentences can be condensed into concise phrases. Memos may be kept separately on word documents. This should, however, be done very carefully to ensure that important information is not thrown away.

Large amounts of unstructured open-ended data can be stored chronologically in a working word document file. Formatting is an extremely useful tool for navigating documents to find records when they are needed.

## DATA DOCUMENTATION

Databases should always include documentation (or metadata), at the very least, on:

- variable definitions (particularly where not evident), through a data dictionary
- modifications to the original database
- sources of data and information
- constraints in data use.

This can either be embedded within the file or stored as a separate 'read me' file.

# REFERENCES

ACAPS, *Technical Brief: How to Approach a Dataset – Part I: Database Design,* August 2013.

Atkinson, Jean, Handbook For Interviewers, HMSO, London, 1971.

Feinstein Group, *Participatory Impact Assessment: A Design Guide*, 2014.
Available at: http://fic.tufts.edu/assets/PIA-guide_revised-2014-3.pdf.

Iarossi, Giuseppe, *The Power of Survey Design: A User's Guide to Managing Surveys, Interpreting Results, and Influencing Respondents*, World Bank, Washington D.C., 2006.
Available at: https://openknowledge.worldbank.org/handle/10986/6975.

ICRC, *Assessing Economic Security*, 2016.

ICRC, *Economic Security Assessment and Monitoring Questionnaire Libraries*, April 2015.

ICRC, EcoSec Executive Brief on Accountability to Affected Populations,
25 September 2014.

ICRC, *Lebanon: Cash-Transfer Programming Vulnerability Assessment Questionnaire*,
November 2014.

ICRC, Rules on Personal Data Protection, ICRC, Geneva, January 2016.
Available at: https://shop.icrc.org/publications/international-humanitarian-law/icrc-rules-on-personal-data-protection.html.

IFRC/ICRC, The Code of Conduct for the International Red Cross and Red Crescent Movement and Non-Governmental Organizations in Disaster Relief, 2004.
Available at: http://www.ifrc.org/en/publications-and-reports/code-of-conduct/.

Plateck, R., Pierre-Pierre, F. K. and Stevens, P. Stevens, *Development and Design of Survey Questionnaires,* Statistics Canada: Census and Household Survey Methods Division, Ottawa, 1985.

Saldaña, Johnny, *The Coding Manual for Qualitative Researchers*, SAGE Publications, London, 2009.

Saldaña, Johnny, *Fundamentals of Qualitative Research,* Oxford University Press, New York, 2011.

WFP, *Emergency Food Security Assessment (EFSAs) Technical Guidance Sheet No. 8: Introduction to Qualitative Data and Methods for Collection and Analysis in Food Security Assessments*, February 2009.

# CHAPTER 5
# SAMPLING

**Sampling** is the process of selecting units (people, households, communities, etc.) from a population of interest for surveying and/or studying; the results of this survey and/or study will then be generalized back to the population from which the sampling units were chosen. Sampling is different from a census, in which every person or entity in the population of interest is included in the survey or study.

## KEY CONCEPTS IN SAMPLING

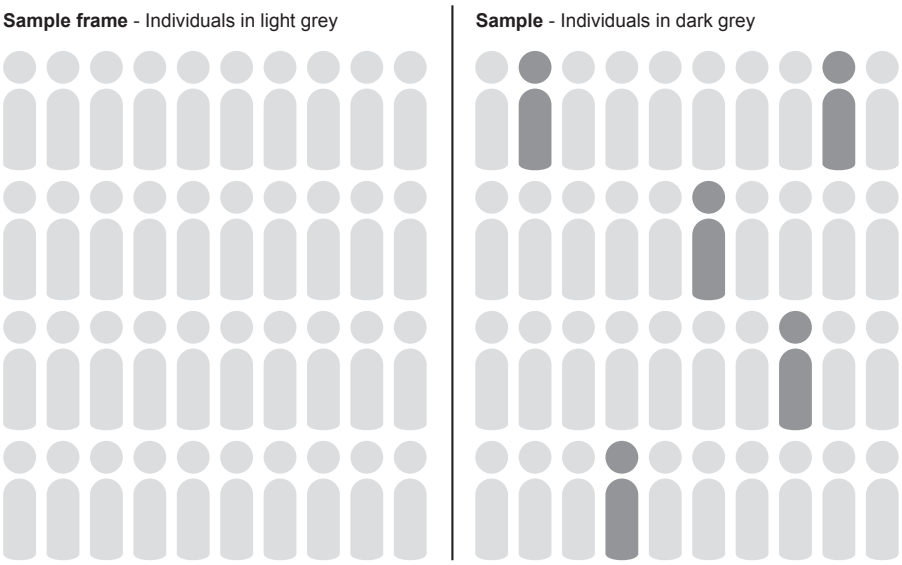
### SAMPLE FRAME

A **sample frame** is a list of potential sampling units (people, households, institutions, etc.). It is, in fact, an exhaustive list of all the sampling units that have a chance or probability of selection for the 'sample'. Units that do not have a chance are not part of the sample frame.

### SAMPLE

A **sample** is that group of people, households, institutions, etc. which is selected from the sample frame for interviewing or studying. There are two main sampling methods: probability and non-probability. A **probability sampling method** is any method of sampling that uses some form of random selection in which every individual or entity has an equal chance (probability) of being selected for the sample; and in which the selection of one individual or entity is independent of the selection of another. A **non-probability sampling method** does not use random selection throughout the process (it may, however, be used at certain stages); therefore, every individual or entity does not have an equal chance of being selected for the sample.

**Figure 24 - Conceptual graphic of sample frame and sample.**




**EXAMPLE**

A team decides to conduct a monitoring exercise following the distribution of fuel sticks to 4,315 households in four camps; the aim is to understand if the fuel sticks were needed and how much they contributed to the household's overall cooking needs. The team interviews 255 randomly selected households. In this case, the sampling frame is the 4,315 households in the four camps, and the sample is the 255 households that were selected for the interviews.

### SAMPLING UNITS

**Sampling units** are units used to draw the sample: households, individuals, children, etc. In two-stage sampling, there is a **primary sampling unit** and an **ultimate sampling unit**. The primary unit is the one used in the first stage. the ultimate unit is used in the second stage of sampling. For example, in two-stage cluster sampling, the primary unit might be villages – where only a sample of all the villages in the sample frame is selected - and the ultimate sampling unit might be households within the sampled villages. In multi-stage sampling, a primary and a secondary unit (and more) may precede the ultimate sampling unit.

## SAMPLING METHODS

### PROBABILITY SAMPLING

A **probability sampling method** is any method of sampling that uses some form of random selection in which every individual or entity has an equal chance (probability) of being selected for the sample; and in which the selection of one individual or entity is independent of the selection of another. The advantage of probability sampling is that results can be extrapolated to cover the entire population with quantifiable precision and confidence. There is much less sampling bias than in non-probability sampling, owing to the entirely random selection of sampling units.

### NON-PROBABILITY SAMPLING

A **non-probability sampling method** does not use random selection throughout the process (it may, however, be used at certain stages); therefore every individual or entity does not have an equal chance of being selected for the sample.

### STRATIFICATION

**Stratification** is the process of dividing the population of interest into non-overlapping sub-groups (called strata) that share certain characteristics of pertinence to the objectives and the potential outcomes of the study. Two common examples of stratification in economic security surveys are: IDPs, host families and non-host families (three strata); and beneficiaries and non-beneficiaries (two strata).

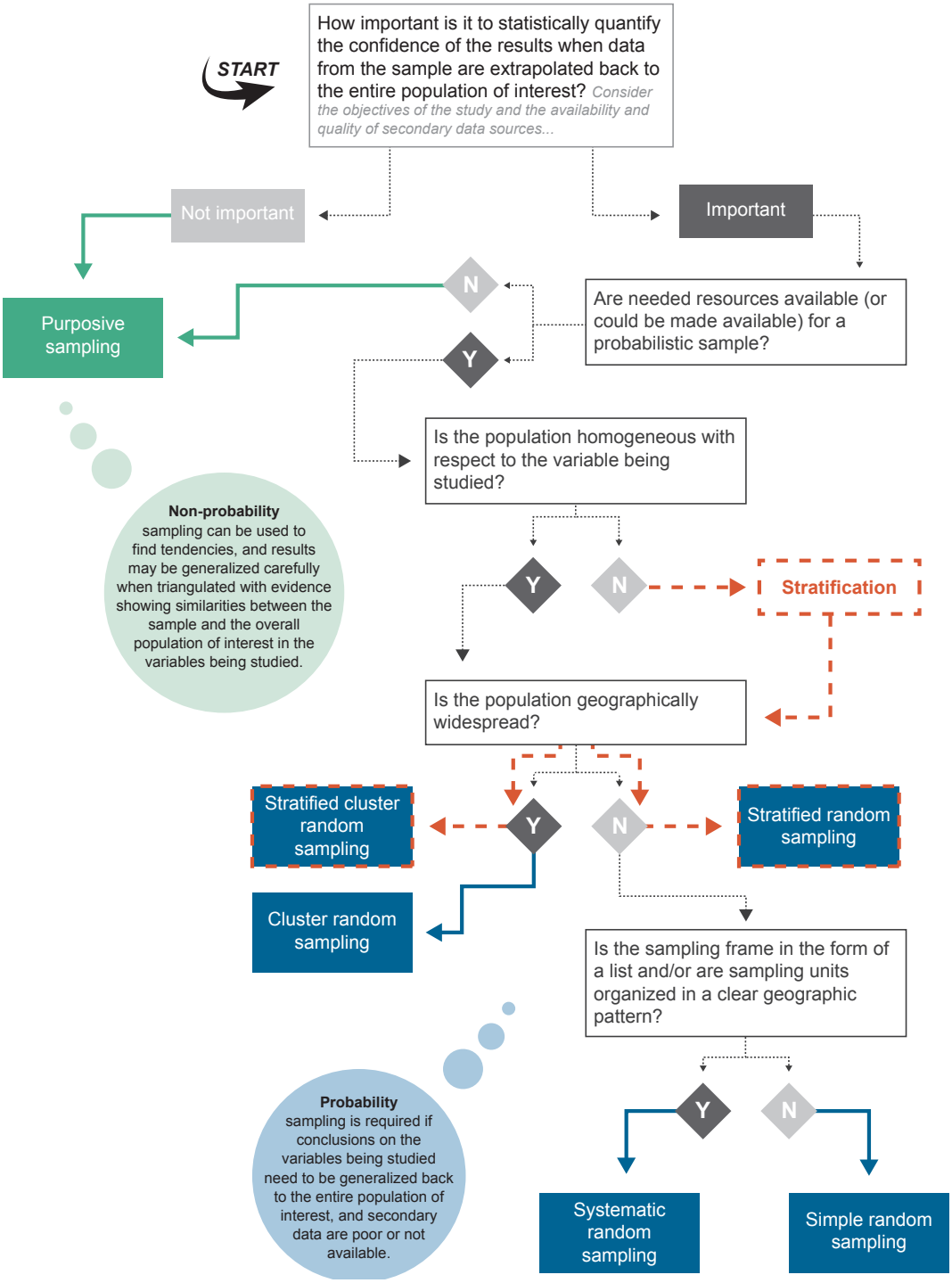
**EXAMPLE**

A team would like to monitor – through indicators on dietary diversity – access for displaced households to a diversified diet. Displaced households in the region of interest are living either in camps or with host families. Their living situation is assumed to influence their access to income and food. The analyst stratifies the sample into two strata: displaced households living in camps and displaced households living with host families.

### DETERMINING THE RIGHT METHOD

Choosing the right sampling method is not an exact science, because it depends on a number of factors: the objectives of the study, the time and resources available, and the size and diversity of the population of interest. There are two main categories of sampling – probability and non-probability sampling – within which there are numerous sampling methods.

**Figure 25 - The 'decision tree' below is a guide for choosing the most suitable sampling method. It is by no means exhaustive; it does, however, cover those methods most commonly used in assessment, monitoring and evaluation exercises undertaken by the ICRC. Other methods should be considered as appropriate.**



## SAMPLE SIZE

'Sample size' is defined as 'the total number of observations in a sample'. For example, in a household assessment, it is the total number of households visited in all sites selected. The size of the sample will be dependent primarily on the following considerations:

- **main indicator measured** (Household Dietary Diversity Score, anthropomorphic measurements, income levels, coping strategies, etc.);
- expected **variations in response** in the population of interest (e.g. will the responses be similar, different, how different?) ;
- **analytical requirements** (average number of days that food distribution served beneficiary households' dietary needs, average percentage of expenditure on food in urban and rural areas, pattern of food consumption before and after project/ programme, etc.);
- **resources available** (time, human resources, etc.);
- **level of precision and accuracy** required (in probability sampling, there is a formula for calculating this; in non-probably sampling, it is conceptualized); and
- **sampling method** (simple random sampling, cluster sampling, stratification for comparison, purposive sampling, etc.).

The idea is that the sample should be large enough to enable discovery of key tendencies or trends in the data, but not so large that the data become redundant. To find out how to calculate the size of a sample, see the sections on 'sample size calculation' under 'Probability sampling' and 'Non-probability sampling' in this chapter.

**CHOOSING AN INDICATOR TO DETERMINE SAMPLE SIZE**

Surveys measuring one indicator use the expected variation in responses (variance) of that particular indicator to determine the sampling methodology and sample size. However, in economic security exercises, for example, several indicators are measured within the same survey. In such cases, all decisions about the sample should be driven by the indicator that is the most demanding in terms of sample size. Other indicators' sample requirements will be fulfilled in turn.[47] For example, a population of interest may be homogeneous in terms of household water storage; their food consumption patterns are, however, expected to vary. In this case, the food consumption indicators used in the study should determine the sampling methodology (simple random, stratification, etc.) and sample size required.

## PROBABILITY SAMPLING

A probability sampling method is any method of sampling that uses some form of random selection in which every individual or entity has an equal chance of being selected for the sample; and in which the selection of one individual or entity is independent of the selection of another. The sample must be taken from a population that is as homogenous as possible with respect to the characteristics being studied. The advantage of this method is that the results may be extrapolated to the entire population; in addition, sampling bias is drastically minimized (compared to non-probability sampling).

### KEY CONCEPTS IN PROBABILITY SAMPLING

The following are a few of the most important concepts in probability sampling. They are of pertinence only to probability sampling, as the accuracy of the results of non-probability sampling cannot be measured in a way that is statistically relevant.

47 Magnani, 1999.

## PRECISION

The precision of the statistics produced can be measured using the 'sampling distribution' and the 'standard error'. The **sampling distribution** is the distribution of an infinite number of samples of the same size as the sample in the study. The **standard deviation** (σ) is the spread of values around the estimate in a single sample, or the variation of the sample values. The **standard error** (it is called the 'sampling error' in sampling) is the standard deviation of the sample estimate, and is expressed in percentage points (e.g. +/- 5% or +/- 0.05). Simply put, it indicates to what extent differences between the results in the sample and those in the target population are due to simple luck: the standard error, or sampling error, gives an idea of the precision of the statistical estimate. For example, a study found that sampled malnourished children had an average weight gain of 20%, with a standard error of +/- 5%. The average weight gain for the entire population of malnourished children in the sample frame is then between 15 and 25%.

A low standard error means that there is less variability or range in the sampling distribution. The standard error is also related to sample size: the greater the sample size, the smaller the standard error, because the sample is closer to that of the actual population.[48]

## CONFIDENCE INTERVAL

The **confidence interval** is the range within which a certain percentage of the population or the responses fall. It is used to gauge the reliability of an estimate. Confidence intervals are usually measured at 90, 95 and 99% levels, where 95% means that we can be 95% confident that the true value of the characteristic being measured (at population level) falls within the estimates based on the sample population. The confidence interval depends on the sampling distribution and the standard error: the further the deviation from the mean, the lower the confidence interval.[49]

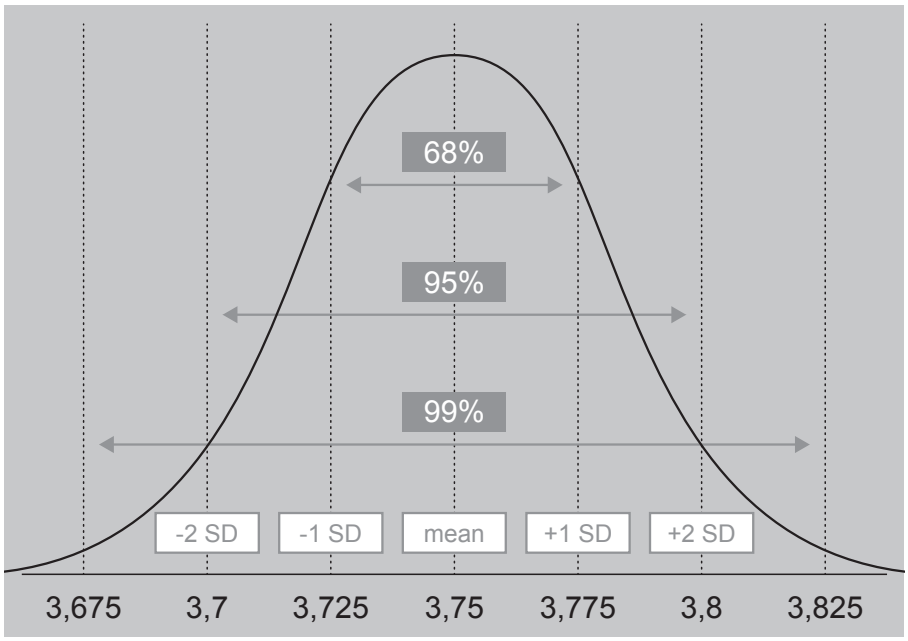



**Figure 26 - The graph above shows how, in a normal distribution, the confidence level increases with the standard deviation from the mean (e.g. accuracy increases as more units are included in the estimate). For example, 95% of cases are expected to fall within 2 standard deviations from the mean (in this case, between 3.7 and 3.8).**

48 Scheuren, 1997.

49 Scheuren, 1997.

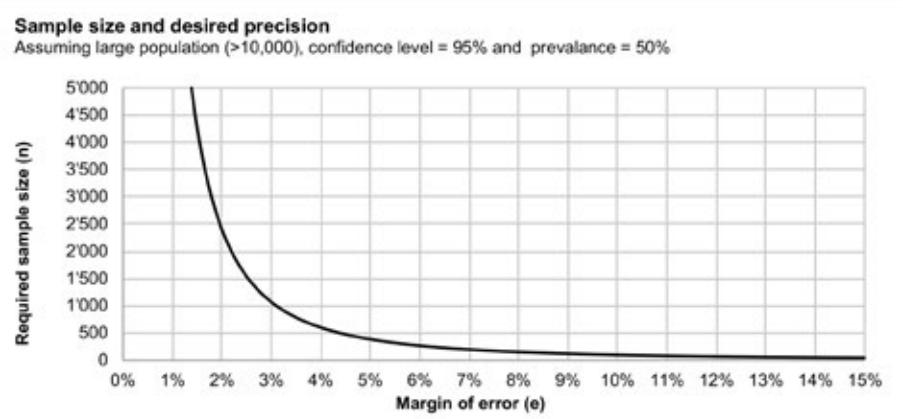



**Figure 27 - The graph above shows the effect of sample size on the margin, assuming 95% confidence and 50% prevalence. The graph demonstrates that the margin of error does not change dramatically with changes to the sample size above 1,000, and stays below 5% with sample sizes of 400 or more.**[50]

## STATISTICAL SIGNIFICANCE AND POWER

**Statistical significance** (α) and **statistical power** (β) are used in comparative studies to determine the risk of false positives and false negatives. Insufficient power (β) may lead researchers or analysts to conclude falsely that there were no changes in indicators over time or differences between groups (false negative). Insufficient significance (α) may lead them to conclude falsely that there were changes or differences, when actually those were observed in the sample only by chance (false positive). In economic security analysis, a minimum value of β of 0.80 and α of 0.90 are generally accepted; if resources permit it, β of 0.90 and α of 0.95 are preferred.[51]

## DESIGN EFFECT

When complex methods such as multi-stage or 'cluster' sampling are used, the sample size needs to be increased to account for the **design effect** (DEFF). It may be thought of as the factor by which the size of a complex sample, such as a cluster sample, would have to be increased in order to produce survey estimates with the same precision as a simple random sample. Ideally, the DEFF from the results of previous similar surveys, which used the same indicators on the same population, should be employed; however, this information is not usually available. A default value of 2.0 (doubling the sample size) is commonly used for cluster samples when cluster sizes are small (30 sampling units or less).[52]

## SAMPLE SIZE CALCULATION

In probability sampling, specific formulas are used to estimate the minimum sample size required to capture expected prevalence rates or measurements with a certain level of precision (margin of error) and confidence (confidence interval). The formulas most commonly employed in general economic security surveys use expected prevalence rates. These formulas were designed for use with categorical variables (proportion of population earning less than 1.25 US dollars per day, proportion of population with HDDS below 4, proportion of population selling cattle as a coping strategy, etc.).

50 WFP, 2010.

51 Magnani, 1999.

52 Magnani, 1999.

To use the formulas, an assumption must be made about the expected prevalence. This can be done by looking at previous surveys in the same context, making educated qualitative estimates or relying on the worst-case scenario of 50%, which will give the largest sample size. The graph below shows how the expected prevalence affects the sample size required to achieve 95% confidence (+/- 5%) for larger populations. Note that sample size increases as prevalence nears 50%.[53]

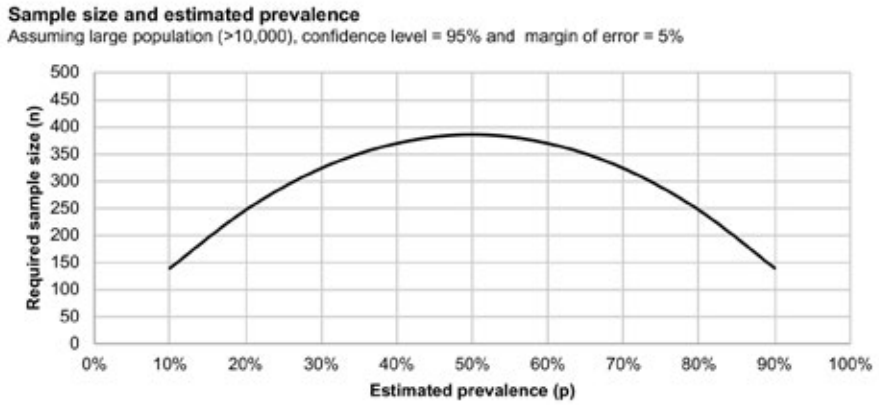



**Figure 28 - The graph above demonstrates the effect of the estimated prevalence (p) on required sample size, assuming a population greater than 10,000, 95% confidence level and 5% margin of error. Sample size is greatest when prevalence is 50%.[54]**

There are also formulas for continuous variables, where the mean or the total is the focus of measurement (average household income, mean daily per capita calorie consumption, coping strategies index, etc.). These formulas are less commonly used – they find favour in general or multi-indicator economic security surveys (apart from surveys of nutrition) – because to produce a 'good' sample size estimate using these formulas, a 'good' estimate of the standard deviation is required, and that is often difficult.[55] The formulas are presented in Annex III.

All the formulas in the following sections are available for automated calculation in the EcoSec sampling calculator, which is available at the EcoSec Resource Centre on the ICRC intranet.

## BASIC FORMULA

One of the most commonly used formulas in social science analyses, such as food security and livelihood analyses, is the basic formula for proportions (proportion of population that ate fish over a 24-hour period, proportion of population engaged in farming, etc.). The formula is designed for use with categorical or qualitative data, and takes into account the desired confidence level, estimated prevalence of the characteristic measured[56] and precision. The following formula is used for populations whose size is unknown:

$$n_0 = \frac{Z^2 \times p \times (1-p)}{e^2}$$

53 WFP, 2010.

54 WFP, 2010.

55 An alternative for continuous quantitative variables is to combine responses into two categories (e.g. above a given threshold and below it, such as the proportion of the population that eats fewer than four food groups in a 24-hour period) and estimate the proportion of the population on either side, together with the basic formula for proportions (Magnani, 1999).

56 This is also known as the **degree of variability, because** a more homogeneous population will have larger proportions with the given characteristic (e.g. 80% of the population with a given characteristic is easier to "catch" than a population where 50% do and 50% don't).

where:

$n_0$ = first estimate of sample size

Z = z-score associated with desired confidence (90 to 95% is most commonly accepted; see table below for associated scores)

p = proportion of the target population with the characteristic being measured (e.g. poor food consumption) or expected prevalence of characteristic (if unknown, 50% or 0.5 is accepted as the most conservative estimate)

e = precision or margin of error (usually 0.05 or 0.10)

**Table 9 - Standard value for Z-score by desired level of confidence (CL) for populations greater than 30.**

| CL | Z |
|---|---|
| 99% | 2.576 |
| 95% | 1.96 |
| 90% | 1.645 |

The formula presented above does not take into account the overall population size. That is because, according to probability theory, population size does not have an effect on the standard error of the sample mean, except in small populations.

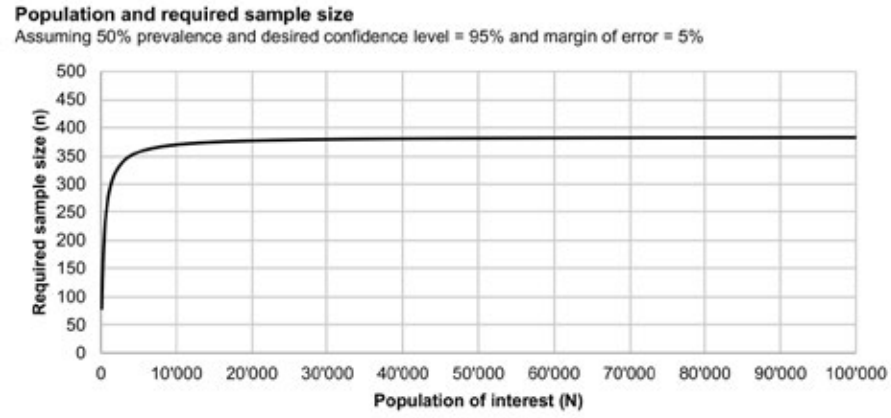



**Figure 29 - The graph above demonstrates how, to gain precision, sample sizes do not have to change significantly for populations greater than 10,000. Sample size requirements vary most for populations less than 500.[57]**

In practice, if the first estimate of the sample size does not exceed 5% of the population (e.g. n/N > 0.05), then the sample size does not need to be adjusted. If the population is small and the first estimate of the sample size is less than 5%, then the sample can be adjusted. The following formula is used:

$$n = \frac{n_0}{1 + (n_0 / N)}$$

where:

n = sample size adjusted to population size

$n_0$ = first estimate of the sample size (as calculated above)

N = Known population size

57 WFP, 2010.

## COMPARISON SURVEY FORMULA

A comparison survey may be designed to analyse the evolution of an indicator (e.g. food consumption levels before and after a project/programme) or to compare groups (food consumption levels of IDPs and residents, project and non-project areas, etc.). If we assume that the levels will be different in each round of the survey or in the group surveyed (e.g. prevalence rate lower or higher), the basic formulas presented above may not provide the optimal sample size estimate, as they take into account only the prevalence (degree of variability) at the time of the current survey or for one group. The following formula from FANTA[58] can be used to estimate an optimal sample size for each round or group.

$$n = \frac{(Z_\alpha + Z_\beta)^2 * (P_1 * (1-P_1) + P2 * (1-P_2))}{(P_2 - P_1)^2}$$

where:

- $n$ = required sample size for each round or group
- $Z_\alpha$ = z-score corresponding to the degree of confidence desired in order to conclude that an observed difference ($P_2$-$P_1$) would not have occurred by chance (α represents statistical significance)[59]
- $Z_\beta$ = z-score corresponding to the degree of confidence desired to be certain of detecting a difference ($P_2$-$P_1$) if one actually occurred (β represents statistical power)[60]
- $P_1$ = estimated proportion of the target population with the characteristics being measured in the first round of the survey or the first group surveyed
- $P_2$ = expected proportion of the target population with the characteristics being measured in the second round or group or in the comparison group, such that the quantity ($P_2$-$P_1$) is the size of the difference that analysts/researchers want to be able to detect

**Table 10** - Standard values of $Z_\alpha$ and $Z_\beta$

| A (ALPHA) | $Z_\alpha$ ONE-SIDED TEST | $Z_\alpha$ TWO-SIDED TEST | β (BETA) | $Z_\beta$ |
|---|---|---|---|---|
| .90 | 1.282 | 1.645 | .80 | 0.840 |
| .95 | 1.645 | 1.960 | .90 | 1.282 |
| .975 | 1.960 | 2.241 | .95 | 1.645 |
| .99 | 2.326 | 2.576 | .975 | 1.960 |
| | | | .999 | 2.320 |

58 Magnani, 1999.

59 Statistical confidence controls for Type I or alpha errors. Alpha is the probability of falsely accepting a difference when there is in fact no difference (e.g. false positive). Value depends on the type of significance test that will be performed. A one-sided test (also known as a 'directional hypothesis') examines the significance of the relationship between variables in one direction (e.g. hypothesis that a sample mean is lower than the population mean). A two-sided test (also known as a 'non-directional hypothesis') examines the significance of the relationship between the variables in either direction (e.g. a sample mean is different from, either lower or higher than, the population mean). If the direction is not known, use the more conservative two-sided test.

60 Statistical power controls for Type II or beta errors. Beta is the probability of falsely accepting no difference when there is in fact a difference (e.g. false negative).

## ADJUSTING FOR THE DESIGN EFFECT (DEFF)

When using complex sampling methods such as multi-stage and cluster sampling, the sample size needs to be adjusted for DEFF. The sample size is simply multiplied by the estimated effect. Default values of 2.0 are commonly used for cluster samples where cluster sizes are generally small (30 sampling units or less).[61]

$$n = n_0 \times D$$

where:

- $n$ = sample size adjusted for DEFF
- $n_0$ = sample size calculated for random sample without DEFF being taken into account
- $D$ = DEFF

## ADJUSTING FOR NON-RESPONSE

When it is anticipated that some samples may not respond or may be unavailable, the sample size can be adapted. The sample size is simply divided by the anticipated rate of non-response. We recommend using the response rates found in previous studies of the same type in the same region (for example, if a previous study of the same type and at the same location had an 80% response rate, the sample size should be divided by 0.8).

$$n = \frac{n_1}{R}$$

where:

- $n$ = sample size adjusted for non-response
- $n_0$ = sample size calculated for random sample without non-response being taken into account
- $R$ = *anticipated* response rate

**TOOLS RECOMMENDED FOR CALCULATING SAMPLE SIZE**

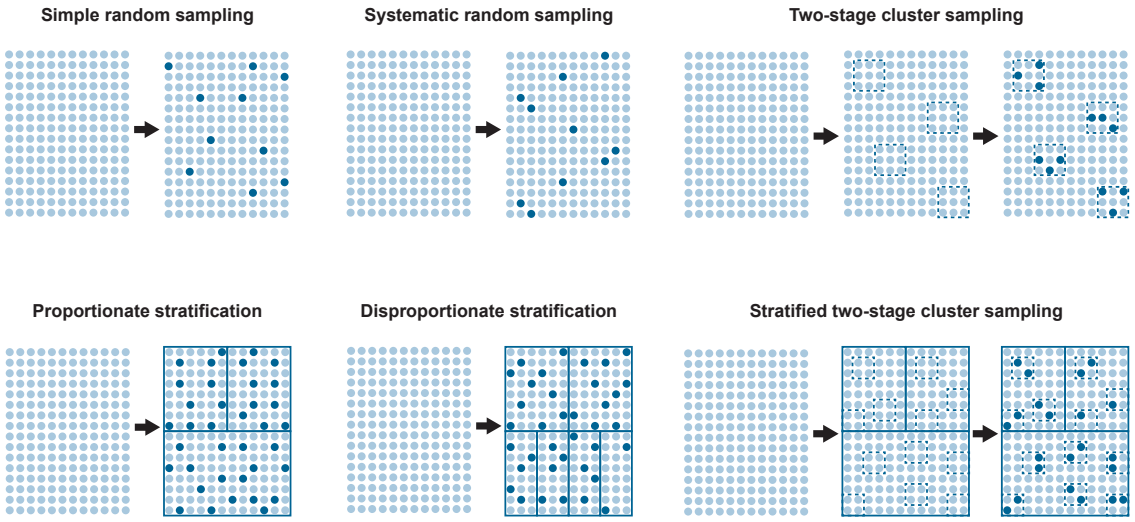
Grasping the formulas is important for some analysts to fully understand the statistics behind calculating sample size. However, there are a number of tools that calculate all the formulas mentioned above, and more. The following are recommended:

| | |
|---|---|
| **EcoSec sample calculator**<br>http://intranet.gva.icrc.priv/ecosec/topics/data-and-analysis/index.jsp | ■ Survey planning – resource requirements and feasible sample size<br>■ Sample size, using basic formula with option for finite population correction<br>■ Sample size, using comparison formula<br>■ Proportionate stratification |
| **OpenEPI**<br>http://www.micronutrient.org/nutritiontoolkit/sampling.htm | ■ Sample size, using basic formula for proportions with option for finite population correction<br>■ Sample size for simple random, comparison and cluster surveys, based on nutrition indicators |
| **Emergency Nutrition Assessment (ENA) Software**<br>http://www.smartmethodology.org/index.php/ | ■ Sample size for simple random and cluster surveys, based on nutrition indicators<br>■ Cluster selection |

61 Magnani, 1999.

## PROBABILITY SAMPLING METHODS

There are six methods that are commonly used in economic security surveys: simple random sampling, systematic random sampling, proportionate stratification, disproportionate stratification, two-stage cluster sampling and stratified cluster sampling.




### SIMPLE RANDOM SAMPLING

In simple random sampling, units from the sample frame are selected at random; in other words, every unit within the sample frame is given an equal opportunity of being selected. Simple random sampling is the purest form of probability sampling, and the most preferred method.

**When should simple random sampling be used?** In humanitarian contexts, simple random sampling can be used for small homogeneous populations when resources permit the drawing of a representative sample or as a second stage of stratified and cluster sampling.

**How should simple random sampling be done?** There are many ways to draw a simple random sample. Two commonly used methods are described below:

**Method 1 – Drawing names from a hat**

In this method, the name of every unit in the sample frame is written on individual pieces of paper and put into a hat. These pieces of paper – the samples – are drawn one by one until the appropriate sample size is reached.

**Method 2 – Using a random number generator**

Every unit is given a number between 1 and N (where N is the total number of units in the sample frame) and listed in a table, and n random numbers (where n is the sample size) are generated using a computer programme such as MS Excel (Using the Sampling tool in the Analysis Toolpak) or an online programme such as www.random.org/integers or ENA Software. The units associated with those numbers are selected for the sample (e.g. if the programme results are 357, 873, 456, 777, etc., then the 357th, 873rd, 456th, 777th, etc. units are selected).

**EXAMPLE**

An assessment aims to understand if 600 women who lost their husbands during a recent conflict have sustainable access to income. The main variables of study are access to income-generating activities and levels of income. The following steps are taken to randomly select the sample:

1. **Determining the sample size** – A sample calculator using the basic formula for proportions is used to determine that 235 women will need to be included in the sample to have a 95% level of confidence and 5% margin of error.
2. **Assigning a unique ID** - If the data do not already have a unique ID, one is assigned.
3. **Random selection** - The Sampling tool in the Excel Analysis Toolpak is used to randomly select 235 of the 600 records. The Input Range is the data set's unique ID and the Sampling Method is Random.
4. **Selecting the sampling units** - The 235 random numbers are re-matched to the women assigned those corresponding numbers, using the VLOOKUP function in Excel.

| | | | | | |
|---|---|---|---|---|---|
| 1 | 11 | 21 | 31 | 41 | 51 |
| 2 | 12 | 22 | 32 | 42 | 52 |
| 3 | 13 | 23 | 33 | | 53 |
| 4 | 14 | 24 | 34 | 44 | 54 |
| 5 | 15 | 25 | | | 55 |
| 6 | 16 | 26 | 36 | 46 | 56 |
| 7 | 17 | 27 | 37 | 47 | |
| 8 | 18 | 28 | 38 | 48 | 58 |
| 9 | 19 | 29 | 39 | 49 | 59 |
| 10 | 20 | 30 | 40 | 50 | … |

## SYSTEMATIC RANDOM SAMPLING

In systematic random sampling, every xth (every third, sixth, etc.) unit from a sample frame is selected until the required sample size is reached.

**When should systematic random sampling be used?** In the same way as simple random sampling: in humanitarian contexts, systematic random sampling can be used for small homogeneous populations when resources permit the drawing of a representative sample or as a second stage of stratified and cluster sampling. Systematic random sampling is useful when the sample frame is in the form of a set list (list of beneficiaries, patient registry, etc.); it can also be used in household surveys of villages or towns where households, or their dwelling-places, are arranged in rows or in some other pattern.

**How should systematic random sampling be done?** The method used for systematic random sampling will depend on whether or not a list of the units in the sample frame is available.

### Method 1 – For individual, household or community-level surveys where lists are available

In this method, the name of every unit in the sample frame is written on individual pieces of paper and put into a hat. These pieces of paper – the samples – are drawn one by one until the appropriate sample size is reached.

1. **Organize the data**: Ensure that the list is not ordered in a way that could introduce bias (according to age, gender, etc.). Assign each unit in the sample frame a number between 1 and N (where N is the total number of units in the sample frame) and list it in a table.
2. **Calculate the sampling interval**: Determine the appropriate sample size and calculate the sampling interval, using the formula SI = N / n, where: SI = sampling interval, N = total population in the sample frame and n = sample size.

3. **Select the starting point**: Draw a random number between 1 and the sampling interval (SI). Select the SIth unit as the first in the sample.
4. **Continue using the same procedure**: Keep selecting every SIth unit in the table until the sample size is reached.

**EXAMPLE**

The ICRC distributed food rations to 5,095 households over a three-month period. To determine whether adjustments have to be made to the programme, the ICRC has to monitor the households' food consumption (the quantity of food they consume and the diversity of their diet). The main indicators used are the household dietary diversity score and meals per day. The following steps are taken to randomly select the sample:

1. **Determining the sample size** – A sample calculator that employs the basic formula for proportions is used to determine that 358 beneficiary households need to be sampled to have a 5% margin of error and 95% level of confidence.
2. **Defining the sampling interval** – 5,095 households / 358 samples = approximately 15.[62]
3. **Organizing the data** – Assign a number to each household between 1 and N (where N is the total number of households).
4. **Selecting the households** – The random number 7 is generated in Excel and used as the first unit; then, every 15th household on the list is selected until 358 households have been selected.

| | | | |
|---|---|---|---|
| 1 | 11 | 21 | 31 |
| 2 | 12 | | 32 |
| 3 | 13 | 23 | 33 |
| 4 | 14 | 24 | 34 |
| 5 | 15 | 25 | 35 |
| 6 | 16 | 26 | 36 |
| | 17 | 27 | |
| 8 | 18 | 28 | 38 |
| 9 | 19 | 29 | 39 |
| 10 | 20 | 30 | 40 |

**Method 2 – For household-level surveys where lists of households are not available**

Lists are often not available for household surveys. When that is the case, and if the households (or their dwelling-places) are organized in a pattern (e.g. as rows), the following method can be used.

1. **Calculate the sampling interval**: Determine the appropriate sample size and calculate the sampling interval, using the formula SI = N / n, where: SI = sampling interval, N = total population in the sample frame and n = sample size. For example, if there are 5,000 households and 250 have to be included in the sample, the sampling interval is 20.
2. **Select the starting and end points**: Select a start and an end point, and a path that covers all the households in the village. Go to the starting point.
3. **Identify the starting direction**: Randomly select the direction to take to choose the first household. This can be done by throwing a pen in the air and basing the choice of direction on where the head points when it lands, spinning a pen or bottle on the ground, etc.
4. **Identify the first household**: Select a random number between 1 and the sampling interval. Walk in the direction selected, counting the households along the way. Stop and survey the household equal to the random number chosen. For example, if the random number between 1 and the sampling interval (in this case 20) was 7, stop at the 7th household.
5. **Identify the subsequent households**: After the first household is surveyed, continue to follow the path chosen and walk SI households (in this case, 20 households) away from the first household and stop and survey the household. Follow this procedure until the entire sample has been covered.

If the households (or their dwelling-places) are not arranged in a pattern, and systematic random sampling is not possible, subsequent household selection can be used (see figure below, which illustrates the WHO EPI method). Note that subsequent household selection is the least preferred method, because of the risk of bias: the selection of a household is not independent of the selection of the previous one; and the use of proximity may cause the entire sample to be drawn from the same area, which increases the likelihood that certain

62 Always round upwards.

characteristics, found only in areas not included in the sample, will be left out.[63] Subsequent household selection may be carried out by taking the following steps:

1. **Select the starting point**: Select a central location in the village/town (near the approximate geographic centre).
2. **Identify the starting direction**: Randomly select the direction to take to choose the first household. This can be done by throwing a pen in the air and basing the choice of direction on where the head points when it lands, spinning a pen or bottle on the ground, etc.
3. **Identify the first household**: Count the number of households that exist along the directional line (in the starting direction identified in step 2 above) from the starting point (identified in step 1). Draw a random number between 1 and the total number of households along the directional line selected. Use this number as the starting household. For example, if the random number is 7, start at the seventh household from the starting point along the starting direction.
4. **Identify the subsequent households**: The second household visited is the household nearest the first (starting) household. The next nearest household to visit is the one whose front door is closest to the front door of the household just visited. Follow this pattern until the required sample size has been reached.

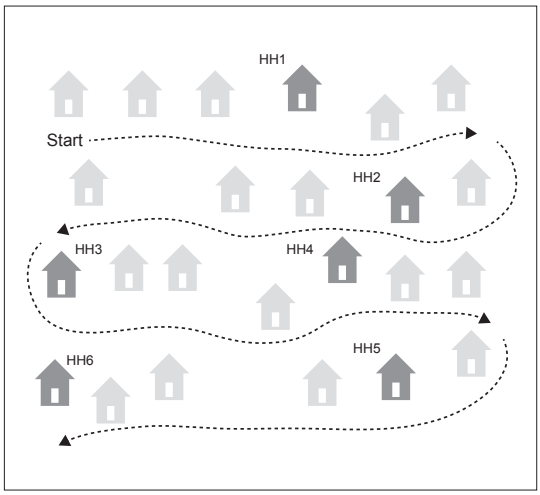



**Figure 30** - Simplified illustration of systematic household selection

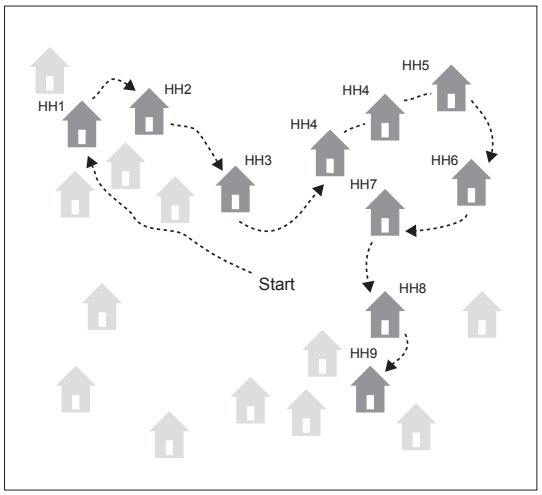



**Figure 31** - Simplified illustration of subsequent household selection[64]

**WHAT IF A HOUSEHOLD OR RESPONDENT IS NOT AVAILABLE?**

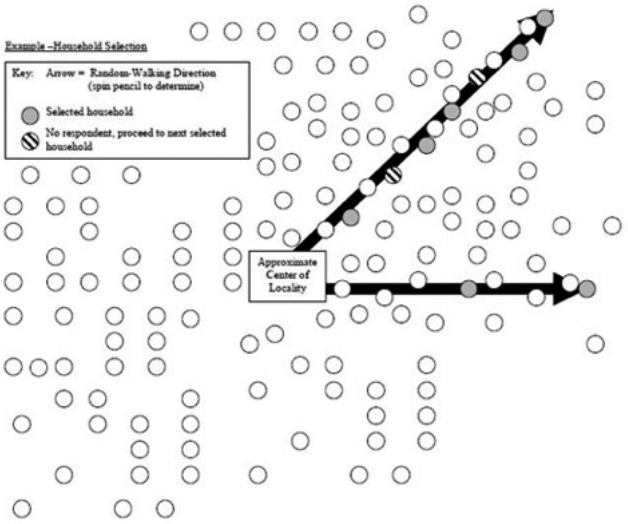
If a respondent is not available, every attempt should be made to get in touch with that particular household, even if it has to be visited later in the day or on subsequent days. If that is not possible, replacement respondents may be considered.

Replacement respondents can be taken into account in the sample plan, figured in as 'non-response', and strategies developed for their selection before setting out to the field.

The WFP used the following method during a survey in South Sudan when two households were unavailable: To select replacement households, the data-collection team returned to the geometric centre of the locality, randomly selected a transect line and divided the total number of households (8) along the line by 2 (the number of replacement households needed). They then visited every 4th household along that line (sampling interval = 8/2).

63 SMART, 2012.

64 WHO, 1991.




## PROPORTIONATE STRATIFICATION

Proportionate stratification is a technique used to increase precision by accounting for the heterogeneity of the variable measured that is related to some key characteristic of the population of interest (geographic location, livelihood group, gender, etc.). In proportionate stratification, the sample frame is divided into non-overlapping sub-groups (called strata) that are, in terms of the variable measured, homogeneous within and heterogeneous between. The sample size of each stratum is proportional to its representation in the overall population of interest.

**When should proportionate stratification be used?** Proportionate stratification may be used to ensure that particular groups within a population are adequately represented in the sample without creating bias. It is designed to increase the precision of an estimate for the whole population, not for an individual stratum.[65] If more precise estimates of individual strata have to be made, disproportionate stratification may be used (see the next section on disproportionate stratification). Proportionate stratification may also be used as a precursor to other sampling methods, such as cluster sampling.

**How should proportionate stratification be done?** In proportionate stratification, the total sample size is calculated, the sample frame divided into strata, and a simple or systematic random sample then taken individually from each stratum.

1. **Determine the sample size (n)** as described in the section on probability sampling.
2. **Divide the population into non-overlapping groups (strata)**, N1, N2, N3, …, Ni, such that N1 + N2 + N3 + … + Ni = N, where N is the size of the total population of interest.
3. **Calculate the proportion of each stratum to the total population in the sample frame** (e.g. for a data set stratified by urban vs. rural, 25% of the sample frame are in the urban stratum and 75% in the rural stratum). Use the following formula to calculate proportions:

$$P1 = N1 / N$$

where:

P1 = proportion of the first stratum to the total population of the sample frame
N1 = population of first stratum
N = total population

65 Tulane University, 2015; accessed in January 2015.

4. **Calculate the sample size in each stratum** by multiplying the total sample size by the proportion for each stratum (e.g. for the example above, multiple the total sample by 0.25 to determine the size of the urban sample and by 0.75 to determine that of the rural sample). Use the following formula to calculate the sample size for the first stratum:

$$n1 = P1 \times n$$

where:

n1 = sample for the first stratum
P1 = proportion of the first stratum to the total population of the sample frame
n = total sample

**EXAMPLE**

After three rounds of food and seed distribution for conflict-affected people in the south-eastern section of the Central African Republic, a team is designing a monitoring exercise to learn more about the economic situation of the households assisted, particularly with regard to food consumption and access to income-generating activities; so that the delegation can be told if further relief activities are necessary and/or whether intensification of livelihood-support activities might be more appropriate. The team would like to ensure that, to increase precision, four towns where the programme was implemented will be adequately represented in the sample, as variation related to location will be included in the sample. Proportionate stratified random sampling will be used in this case, to divide the sample over the four towns. While the study will look at numerous variables, household dietary diversity will be used as the basis for calculating sample size, as it is the most demanding in terms of sample size. The following steps are taken.

1. **Beneficiary figures obtained** – There are a total of 11,848 beneficiary households: 4,896 from Zémio, 2,789 from Obo, 1,847 from Mboki and 2,316 from Rafaï.
2. **Required sample size calculated** - A sample size calculator is used to determine that a sample of 266 households is needed for a 5% margin of error and 90% level of confidence.
3. **Proportion of each stratum to the total population in the sample frame calculated,** and the sample size per stratum. To calculate the proportion of each stratum, the population of each stratum is divided by the total population (e.g. proportion from Zémio = 4,896 / 11,848). Then, to calculate the sample size for each stratum, the proportion is multiplied by the total sample size (e.g. sample size for Zémio = 266 x 0.413). The results are shown in the table below.
4. **Fractions rounded off** – Some estimates end up as fractions. These are rounded off, and the sample size adjusted accordingly.

| Strata | Population | Stratum / Population | Weight (W) | Sample size first estimate (266*W) | Sample size adjusted (after rounding) |
|---|---|---|---|---|---|
| Zémio | 4,896 | 4,896 / 11,848 | 0.413 | 109.9 | 110 |
| Obo | 2,789 | 2,789 / 11,848 | 0.235 | 62.6 | 63 |
| Mboki | 1,847 | 1,847 / 11,848 | 0.156 | 41.5 | 42 |
| Rafaï | 2,316 | 2,316 / 11,848 | 0.195 | 51.9 | 52 |
| Total | 11,848 | - | - | 266 | 267 |

**PROPORTIONATE VERSUS DISPROPORTIONATE STRATIFICATION**

**Proportionate stratification** is designed to increase the precision of an estimate for the whole population, not for individual strata. **Disproportionate stratification** can be used to make reliable estimates for each and every stratum or reliable comparisons between strata, or to optimize costs and/or precision.

A combination of the two methods may be used in surveys. For example, disproportionate stratification may be used to draw samples, and calculate their size, for displaced and non-displaced populations, in order to generalize findings back to each group. Proportionate stratification may be used afterwards to stratify by geographic area, in order to ensure that all regions and variation associated with them are included.

## DISPROPORTIONATE STRATIFICATION

Disproportionate stratification is a method of stratification in which the sample size of each stratum is not determined in proportion to its representation in the overall population of interest. This may be for one of the following reasons:

- **to facilitate within-strata analyses** (extrapolate back to all IDPs in a region, extrapolate back to all residents of a region, etc.);
- **to facilitate between-strata analyses** (food consumption of displaced households compared with that of resident households, living conditions in region A compared with those in region B, etc.); or
- **to focus on optimizing costs** (e.g. more weight given to areas easier to reach) or on precision (e.g. more weight given to more heterogeneous strata).

**When should disproportionate stratification be used?** To make reliable estimates for each and every stratum, to make reliable comparisons between strata or to optimize costs and/or precision when strata differ in terms of data-collection costs or the variation within each stratum.[66] Like proportionate stratification, disproportionate stratification may also be used as a precursor to other sampling methods, such as cluster sampling.

**How should disproportionate stratification be done?** In disproportionate stratification, the sample frame is divided into strata, and individual samples drawn from each stratum; sample sizes will be determined by the level of precision required to draw conclusions or to compare each stratum with others.

1. **Divide the population into non-overlapping groups** (strata), N1, N2, N3, …, Ni, such that N1 + N2 + N3 + … + Ni = N, where N is the size of the total population of interest.

2. **Determine the sample sizes** (n1, n2, n3, …, ni) for each group (stratum) using one of the following methods, the choice of which will depend on the analytical requirements.

   **Method 1 – Individual samples, for facilitating within-strata analyses**

   The idea here is to ensure that the sample size of each stratum has the appropriate variance (according to the expected prevalence or standard deviation) of the characteristic measured to enable conclusions to be drawn with the desired level of precision. In this case, the most accurate method for estimating the required sample size would be to calculate sample sizes for each stratum separately.[67] Knowledge of the expected values (e.g. when maximum prevalence of 50% is used) for each stratum may not be available. In the case of large populations where the same precision and level of confidence is expected, the same sample size may be calculated and used for both strata. This may also be easier to explain to field teams and informants (e.g. the study will sample 250 households in urban and 250 in rural areas). Use the *Basic formula for a random sample* (see "Sample size calculation" on page 103) for each stratum.

   **Method 2 – Equal allocation, for facilitating between-strata analyses**

   This is the most commonly used of the three methods. The idea here is to ensure that the sample size is large enough to detect differences between the strata up to a desired level (detect differences when the true difference is 10%, 15%, etc.). Use the formula for comparison surveys (on page 90) to calculate the required sample size for each stratum separately. If differences in the prevelance – i.e. percentage of population with the given characteristic – are unknown for each stratum, use the maximum 50% prevalence for the first group ($P_1$) and add the desired level of difference to calculate the prevalence for the second group ($P_2$). For example, if it is desired to capture differences between the groups when the real difference is 15% between $P_1$ and $P_2$, and P1 is set to 50%, $P_2$ = .50 +.15 = .65. The sample size for each stratum in this case is 132 (for $\alpha$=.95 and $\beta$=.80).

66 Daniel, 2012.

67 UN DESA, 2005.

**Method 3 – Optimum allocation**
Calculate the appropriate sample size for the total population of interest, and estimate the allocation for each stratum based on analytical and resource requirements (e.g. more weight given to regions that are closer, owing to resource constraints; more weight given to regions with greater variation in the variable under study).

3. **Calculate survey weights**: Aggregated analysis of data from all strata, collected via disproportionate stratification, requires the use of weights. See page 129 for more information on sample weights.

**EXAMPLE**
Using the same scenario as in the section on proportionate stratification above, suppose the beneficiaries include both IDP and resident households. The objective of the exercise remains the same: "to learn more about the economic situation of the households assisted, particularly with regard to food consumption and access to income-generating activities; so that the delegation can be told if further relief activities are necessary and/or whether intensification of livelihood-support activities might be more appropriate." In this case, however, based on their field experience, members of the team are sure that the outcomes for resident and displaced populations will be different. For this reason, they would like to compare the results from the two groups and present the resulting information. In this case, disproportionate stratification will be used with one stratum for IDP households and one for resident households, followed by proportionate stratification to ensure that all four towns are proportionately covered. The following steps are taken:

1. **Stratum determined and beneficiary figures obtained** – There are a total of 11,848 households: 5,176 IDP and 6,672 resident households.
2. **Sample size required for each stratum calculated** – In this case, the comparison formula is used to facilitate between-strata analyses. A sample calculator is used to determine that 221 households are required per group (total sample size of 442) to detect differences when the true difference is 10% (for $\alpha$=.90 and $\beta$=.80). The following result is obtained.

| Strata | Population | Sample size |
|---|---|---|
| IDPs | 5,176 | 221 |
| Residents | 6,672 | 221 |
| Total | 11,848 | 442 |

3. **Proportion of population by type in each town in the sample frame, and the sample size per town, calculated,** using method under proportionate stratification.
4. **Fractions rounded off** – Some estimates end up as fractions. These are rounded off, and the sample size adjusted accordingly.

| | Population | | Sample size | |
|---|---|---|---|---|
| Strata | IDPs | Residents | IDPs | Residents |
| Zémio | 1,952 | 2,944 | 84 | 98 |
| Obo | 1,686 | 1,103 | 72 | 37 |
| Mboki | 408 | 1,439 | 18 | 48 |
| Rafaï | 1,130 | 1,186 | 49 | 40 |
| Total | 5,176 | 6,672 | 223 | 223 |

5. **Random sample taken** – A systematic random sample is taken within each stratum, using the sample sizes defined in step four and methodologies defined for systematic random sampling.
6. **Weights computed for analysis** – As this sample employed disproportionate stratification, weights are required when computing statistics for the overall population (e.g. averages for both IDPs and residents). See page 129 for more information on sample weights.

| RULES IN STRATIFICATION | WHEN DO I NEED TO FOLLOW THE RULES? |
|---|---|
| Each stratum should have at lease one sampling unit within it. | **Always** If no units are selected, then no units in the stratum had a chance of being selected, thus inhibiting the randomness of the sample. |
| Each stratum should be as different as possible in terms of the variable being measured. | **Always as predictable** Stratified sampling is most effective when variability within strata is minimized, variability between strata is maximized, and the variables upon which the population is stratified are strongly correlated with the desired dependent variable (which defines the stratum). |

## TWO-STAGE CLUSTER SAMPLING

Cluster sampling divides the population into homogeneous clusters (geographic boundaries or structures such as camps, collective centres and primary-health-care units), and the clusters are randomly selected through random sampling. In simple one-stage cluster sampling every unit within the selected clusters is part of the sample. Cluster sampling is, however, usually combined with another form of sampling, such as simple random sampling, in which only a sample of units within each cluster is selected. This is called multi-stage sampling. For example, in two-stage sampling a number of units within each selected cluster are randomly selected. This is often the case when villages are used as clusters and a sample of households is taken from each cluster (village).

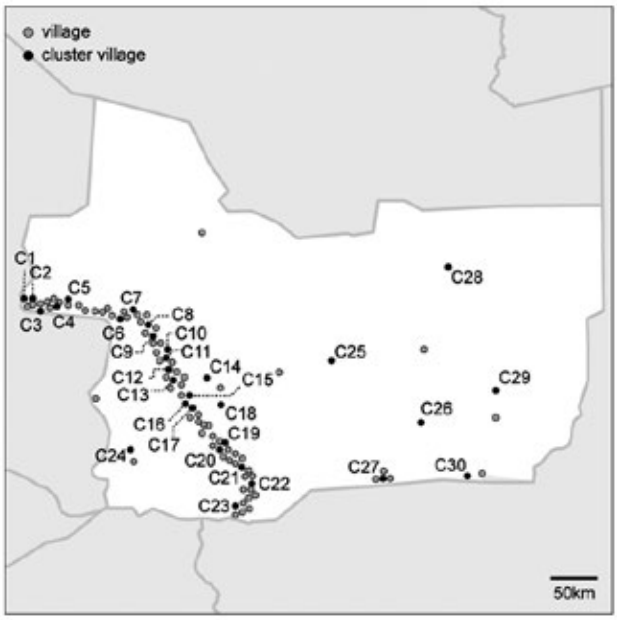



**Figure 32 - The example above shows a region with over 300 villages. In this case, 30 villages are used as 'clusters' in the sample.**

Cluster sampling is different from stratified random sampling in that only a sub-set of clusters is represented in the sample; in stratified sampling, all strata are represented in the sample.

**When should two-stage cluster sampling be used?** When the population of interest is large and/or dispersed, and/or precise population figures are not available, two-stage cluster sampling may be easier, more efficient and more cost-effective than simple, systematic or stratified random sampling; however it is less precise, particularly amongst populations characterized by heterogeneity or responses with heightened variance. In these cases, the design effect (DEFF) needs to be considered, and the sample size increased. Two-stage cluster sampling is frequently used in health and nutrition studies and in large-scale surveys covering a vast area.

**How should two-stage cluster sampling be done?** Two-stage cluster sampling involves the following steps:

1. **Determine primary and ultimate sampling unit**: The primary sampling unit (PSU) is the first unit used in the first stage. In two-stage cluster sampling, it is the 'cluster'. Clusters should be as homogeneous as possible to each other (e.g. the characteristics being measured display the same degree of diversity in every cluster), because not all clusters in the sample frame will be chosen. The ultimate sampling unit (USU) is the unit from which the information will be collected (household, individual, etc.).
2. **Determine the sample size (n)**: Use the most appropriate method, from among those described in the section on sample size on page 100. Remember to take the DEFF into account.
3. **Determine the number of clusters and ultimate sampling units within each cluster**: The total number of clusters will be the total sample size divided by the number of USUs (households, individuals, etc.) that can be visited in each cluster. For example, 30 households (USUs) in 30 villages (PSUs or clusters) amounts to a total sample size of 900 households. It is better to have more clusters with fewer samples than fewer clusters with more samples. The number of USUs in each cluster should be the same in each cluster, to ensure equal probability of selection.
4. **Determine the number of backup clusters**: Sometimes, certain clusters may not be reachable during the survey; the recommendation in such cases is to plan for backup clusters during the design phase. The number will depend on the context: three to five is usually sufficient for larger surveys.
5. **Identify the clusters to visit**: Randomly select the clusters according to the required number, using the probability proportionate to size (PPS) method in MS Excel or ENA Software (see comment box below).
6. **Take a simple or systematic random sample** (stage 2 of multi-stage sampling) of the ultimate sampling unit within each cluster according to the sample sizes per cluster defined under steps two and three.
7. **Calculate survey weights**: If PPS was not used to select clusters, aggregated analysis of data may need weights. See page 129 for more information on sample weights.

**PROBABILITY PROPORTIONATE TO SIZE** (PPS) is a method used in cluster sampling during cluster selection (the first stage in two-stage cluster sampling). It gives larger clusters a greater chance of selection than smaller ones. PPS can be used only if an estimated size of each cluster (e.g. population) is available.. When this method is applied in stage 1 (sie selection) together with the same number of ultimate sampling units selected in each cluster in stage 2 (e.g. 30 households in each cluster), each ultimate sampling unit has the same overall probability of selection. The result is a self-weighted sample, which is an advantage during data analysis.[68] See the section on 'sample weights', on page 112, for more details on self-weighted samples.

68 Magnani, 1999.

**EXAMPLE 1 (USING EXCEL)**

A team is designing an assessment of the economic situation of almost 19,000 internally displaced and resident households, spread over 123 villages in Leer County, South Sudan, who have benefited from numerous rounds of food distribution. As part of the assessment, they will carry out a household-level survey of food consumption, food production, living conditions, access to income and coping mechanisms; the findings will be used to guide medium-term programming. People are constantly on the move, which means that timestamps of IDP locations become outdated and population figures disaggregated by displacement status, unreliable. As the population is large and dispersed, the team decides that two-stage cluster sampling is the best option. Stage one entails randomly selecting the villages to visit and stage two, randomly selecting the households to visit in each village. The team takes the following steps.

1. **Defines the primary and ultimate sampling units** – Villages will be used as the primary sampling unit (or cluster) as it is the smallest unit available. There are 123 villages in the sample frame. The ultimate sampling unit consists of households, of which there are precisely 18,826.
2. **Determines the sample size needed** – By means of a sample size calculator, and the basic formula for proportions, it is determined that for simple random sampling, 376 households have to be sampled to have a 5% margin of error and 95% level of confidence. As cluster sampling will be used, the sample size is multiplied by 2 to account for the DEFF: the sample size arrived at in this way is 752.
3. **Determines the number of clusters** – The team determines that they can visit up to 30 households per village. The team must visit 26 villages in order to reach 752 households (i.e. 752 total households / 30 households per village = 25.1 villages, which is rounded off to 26). The rounding off of villages increases the sample size to 780 (26 villages x 30 households per village).
4. **Determines the number of backup clusters** – It is the rainy season in South Sudan, and the team is sure that some villages will not be accessible. They decide that it would be ideal to have up to 5 backup clusters.
5. **Identifies the clusters to visit, using PPS** – The clusters are chosen at random, using the following method:
   - **The list of villages is randomized** by creating a randomly ordered list of 123 unique numbers with an online tool (https://www.random.org/integer-sets/). These numbers are set alongside the data. The list of villages is then 'sorted' in Excel by the random number associated with the village.
   - **The cumulative sum of the population of the villages is calculated** by adding the first and second population values (for the 2nd village), the 1st + 2nd + 3rd (for the 3rd village), 1st + 2nd + 3rd + 4th (for the 4th village), etc. until the last village (see column 3 in table below).

| 2 | Random_Number | Village | TotalHH | Cumulative_Sum |
|---|---|---|---|---|
| 3 | 0.004552228 | Koat | 121 | 121 |
| 4 | 0.019492254 | Mager | 108 | 229 |
| 5 | 0.024829967 | BulToulony | 61 | 290 |
| 6 | 0.027612708 | Kuiybol/Dhorbol | 61 | 351 |
| 7 | 0.034079772 | Pomdhor | 281 | 632 |
| 8 | 0.046799993 | Geer | 287 | 919 |
| 9 | 0.051794916 | Tuoluong/Chamriak | 189 | 1,108 |
| 10 | 0.054870122 | Dhorreed | 124 | 1,232 |
| 11 | 0.061182656 | Dhornor (Guat) | 97 | 1,329 |
| 12 | 0.065248729 | Dhorbol/Torbamg | 187 | 1,516 |
| 13 | 0.071796971 | NorJoot | 205 | 1,721 |
| 14 | 0.078166582 | Kuath/Thoriak | 85 | 1,806 |

   - **The sampling interval is calculated** with this formula: total population / no. primary clusters + backup clusters = 607
   - A random number is generated to start the random selection by using the RAND function in Excel. Example: For RAND()*607, where 15,000 is the sampling interval, the random start is 517.
   - **Five random numbers between 1 and 31 are generated** to facilitate random selection of the five backup clusters (which will be visited only if other clusters are not reachable). In this case, the following numbers are generated: 1, 11, 14, 23 and 31.

- **Thresholds for determining the clusters to be sampled are chosen:** the 1st is that which has a cumulative value closest to the random start, the 2nd is that which has a value closest to the random start + sampling interval, the 3rd is that which has the closest value to the random start + (sampling interval x 2), etc., until 31 clusters are selected. The 1st, 11th, 14th, 23rd and 31st clusters are identified as backup clusters (BC) and all others as primary clusters (C). See columns five and six below.

| | Random_Number | Village | TotalHH | Cumulative_Sum | Cluster_Threshold | Cluster_Number |
|---|---|---|---|---|---|---|
| 3 | 0.004552228 | Koat | 121 | 121 | | |
| 4 | 0.019492254 | Mager | 108 | 229 | | |
| 5 | 0.024829967 | BulToulony | 61 | 290 | | |
| 6 | 0.027612708 | Kuiyboi/Dhorbol | 61 | 351 | 517 | BC1 |
| 7 | 0.034079772 | Pomdhor | 281 | 632 | | |
| 8 | 0.046799993 | Geer | 287 | 919 | | |
| 9 | 0.051794916 | Tuoluong/Chamriak | 189 | 1,108 | 1,124 | C1 |
| 10 | 0.054870122 | Dhorreed | 124 | 1,232 | | |
| 11 | 0.061182656 | Dhomor (Guat) | 97 | 1,329 | | |
| 12 | 0.065248729 | Dhorboi/Torbamg | 187 | 1,516 | | |
| 13 | 0.071796971 | NorJoot | 205 | 1,721 | 1,731 | C2 |
| 14 | 0.078166583 | Kuoth/Thoriak | 85 | 1,806 | | |
| 15 | 0.087465404 | Sukgidim | 182 | 1,988 | | |
| 16 | 0.092326646 | Dhorkeih | 88 | 2,076 | | |
| 17 | 0.093845662 | Dhoreet | 103 | 2,179 | 2,946 | C3 |
| 18 | 0.094422946 | Rubjech | 335 | 2,514 | | |
| 19 | 0.120913421 | Nyawemyal | 325 | 2,839 | 3,553 | C4 |
| 20 | 0.135389814 | Dok | 293 | 3,132 | | |
| 21 | 0.139604164 | Dhorguan | 150 | 3,282 | | |

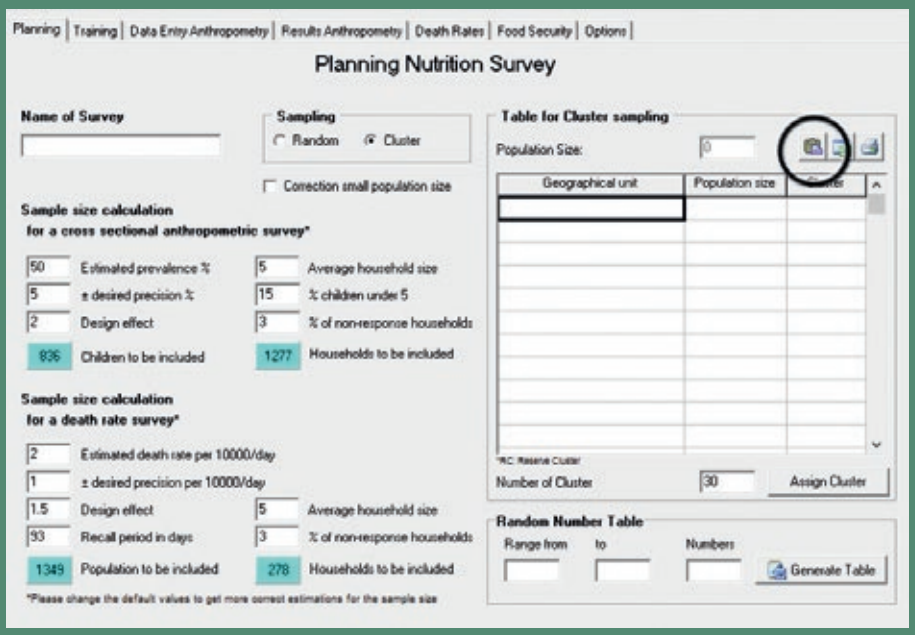
*If population figures are not available by site, PPS is not relevant and sites (clusters) are selected a simple random.

6. **Take a simple or systematic random sample within each cluster** - In this case, 30 households are selected at random in the field for all clusters. If one cluster (village) is not reachable, the first backup cluster is visited and 30 households are surveyed.

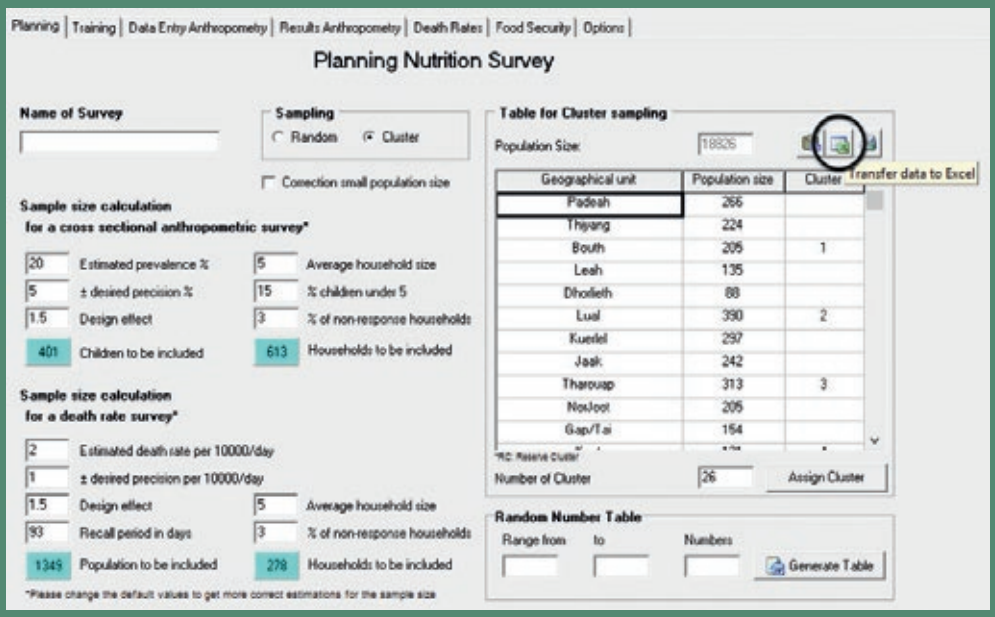
**EXAMPLE 2 (USING ENA SOFTWARE)**

In the scenario described above, selecting the clusters to visit (step 5 above) could also be done in ENA Software (http://www.nutrisurvey.de/ena/ena.html). ENA is designed for nutritional surveys, but parts of it can be used for other surveys as well. To use ENA, first download the software (free of charge online or from the ICT store for ICRC computers). Then, follow steps 1 through 4 mentioned above, and replace steps 5 and 6 with the following:

1. Navigate to the Planning tab.
2. Copy the list of primary sampling units (in this case, villages) and the population of the secondary sampling unit (in this case, households) and paste them into the Table for Cluster sampling. Use the paste function (as shown below) to paste the data.



3. Under Number of Clusters, enter the number of clusters identified in step 3 above: in this case, 26. Note that ENA automatically assigns backup clusters (called Reserve Clusters or RC) in addition to the primary clusters, so these do not need to be accounted for.
4. Select Assign Cluster and see the results in the "Cluster" column.
5. Export the cluster assignment to Excel by using the export function (as show below).




6. Take a simple or systematic random sample within each cluster - In this case, 30 households are selected at random in the field for all clusters. If one cluster (village) is not reachable, the first backup cluster is visited and 30 households surveyed.

*How many clusters should I select?* Our recommendation is to select many clusters and fewer ultimate sampling units (households, individuals, etc.) instead of fewer clusters and more sampling units, as that will increase variation in the sample. Recommended cluster numbers for economic security and nutrition surveys covering large areas are given below.[69] Every case should, however, be reviewed.

**Standard:** 30 clusters | **Compromise:** 25 | **Minimum:** 20

## STRATIFIED CLUSTER SAMPLING[69]

Some situations may require a combination of sampling methods. This can facilitate field exercises and when used correctly, can even increase accuracy and facilitate inference of variables over widespread areas and/or over key regions/groups of particular interest. The combined use of stratification and two-stage cluster sampling is common in humanitarian studies. This entails simply stratifying the sample frame into non-overlapping heterogeneous groups, and then applying two-stage cluster sampling for each stratum separately. The sample size for each stratum will depend on the analytical requirements (see the pertinent sections under '*Proportionate stratification' and 'Disproportionate stratification*').

69 WFP, 2004.

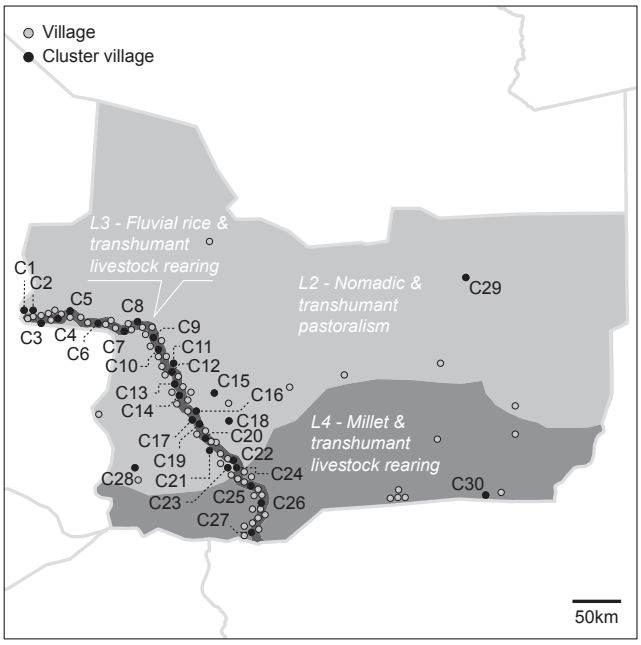



**Figure 33 - The example above shows a sample frame stratified by livelihood zones; villages are used as clusters.**

**When should stratified cluster sampling be used?** When the population of interest is heterogeneous and numerous and/or dispersed, and/or precise population figures are not available. Stratification can ensure that particular groups within a population are adequately represented in the sample without creating bias; the use of clusters can help to greatly expand coverage.

**How should stratified cluster sampling be done?** Stratified two-stage cluster sampling involves the following steps:

1. **Stratify the sample frame into non-overlapping groups** by using the appropriate method, from among those outlined under 'Proportionate stratification' and 'Disproportionate stratification'.
2. **Determine the primary sampling unit** (PSU, or cluster) and ultimate sampling unit that will be used (as described in the section on cluster sampling).
3. **Calculate the sample size per stratum**: Calculate the sample size for each stratum (see page 113 for 'Proportionate stratification' and page 115 for 'Disproportionate stratification').
4. **Determine the number of clusters and ultimate sampling units (USUs)** within each cluster: The total number of clusters will be the total sample size divided by the number of USUs (households, individuals, etc.) that can be visited in each cluster. For example, 30 households (USUs) in 30 villages (PSUs or clusters) amounts to a total sample size of 900 households. It is better to have more clusters with fewer samples than fewer clusters with more samples. Each cluster should have the same number of USUs, to ensure equal probability of selection.
5. **Determine the number of backup clusters**: Sometimes, certain clusters may not be reachable during the survey; the recommendation in such cases is to plan for backup clusters during the design phase. The number will depend on the context: three to five is usually sufficient for larger surveys.
6. **Identify the clusters to visit**: Randomly select the clusters for each stratum separately, in line with the number needed for that stratum, using the probability proportionate to size (PPS) method in MS Excel or ENA software.
7. **Take a simple or systematic random sample from each cluster**, in line with the sample sizes per cluster defined under steps two and three.
8. **Calculate survey weights**: If the survey employs disproportionate stratification OR if PPS was not used to select clusters, aggregated analysis of data will have to use weights for proper results. See page 129 for more information on sample weights.

**EXAMPLE**

Let us take the same scenario as above, of the assessment in Leer County, South Sudan. Now assume that the team would like to increase precision by ensuring proportionate representation by geographic region. They decide that in this case, proportionate stratification will be used with one stratum for each *payam* (second administrative unit). This will be followed by two-stage cluster sampling: in stage one, the villages to visit will be selected randomly; and in stage two, the households to visit in each village.

1. **Stratify the sample frame** – Data are organized in a way that makes it easy to calculate figures for the six payams in Leer County.
2. **Determine the primary and secondary sampling units** – Villages will be used as the primary sampling unit (or cluster) as they are the smallest units available, and households as the secondary sampling unit.
3. **Calculate the sample size needed for each stratum** – Because proportionate stratification will be employed, the sample size for the entire sample frame is determined with a sampling calculator using the basic formula for proportions. It is determined that for simple random sampling, 376 households need to be sampled to have a 5% margin of error and 95% level of confidence. As cluster sampling will also be used, the sample size is multiplied by 2 to account for the DEFF. The total sample is 752. The sample is divided up by payams, and according to the proportion of the population in each payam.
4. **Round off fractions** – Some estimates end up as fractions. These are rounded off, and the sample size adjusted accordingly.

| Strata *(payam)* | Population | Sample size |
|---|---|---|
| Adok | 3,906 | 157 |
| Gandor | 2,519 | 101 |
| Guat | 1,621 | 65 |
| Leer | 6,153 | 247 |
| Pilleny | 2,154 | 87 |
| Thonyor | 2,473 | 99 |
| Total | 18,826 | 756 |

5. **Determine the number of clusters per stratum** –The team determines that they can visit up to 30 households per village. The number of clusters per stratum is determined by dividing the population of each payam by the sample for that payam (e.g. for Adok - 157 total households / 30 households per village = 5.2 villages, rounded off to 6 villages). The rounding off of villages increases the sample size to 870.
6. **Determine the number of backup clusters** – It is the rainy season in South Sudan, and the team is sure that some villages will not be accessible. They decide that it would be ideal to have up to 1 backup cluster in each payam.
7. **Identify the clusters to visit, using PPS in Excel or ENA Software** – The same method is used as that described under two-stage cluster sampling, but individually for each stratum (i.e. repeated six times). ENA Software is recommended in this case, because in Excel much of the work would have to be done manually, increasing the likelihood of errors.

| Strata (payam) | Population | Sample size (n) | Clusters (n/30) | Sample size adjusted to number of clusters (n2) |
|---|---|---|---|---|
| Adok | 3,906 | 157 | 6 | 180 |
| Gandor | 2,519 | 101 | 4 | 120 |
| Guat | 1,621 | 65 | 3 | 90 |
| Leer | 6,153 | 247 | 9 | 270 |
| Pilleny | 2,154 | 87 | 3 | 90 |
| Thonyor | 2,473 | 99 | 4 | 120 |
| Total | 18,826 | 756 | 29 | 870 |

8. **Take a simple or systematic random sample** – In this case, 30 households are selected at random in the field for all clusters. If one cluster (village) is not reachable, the first backup cluster is visited and 30 households surveyed.

## NON-PROBABILITY SAMPLING

Unlike probability sampling, non-probability sampling method does not use random selection throughout the process (it may, however, be used at certain stages); therefore every individual or entity does not have an equal chance of being selected for the sample. It is not possible to quantify precision or bias when using this method. If results are to be generalized to the entire population of interest, they must be triangulated with evidence showing similarities between the sample and the overall population of interest. Every effort should be made throughout the process to minimize bias, and the limitations of the results and areas of error/bias must be reported with the results in a transparent manner.

Non-probability sampling is often used in exploratory studies, rapid assessments and/or evaluations, to collect detailed data that complement less in-depth or detailed data collected at a statistically relevant level (i.e. using probabaility sampling), and when probability sampling is not possible owing to resource or access constraints or lack of baseline sampling data. Probability sampling is often preferred, but results from a study that uses non-probability sampling can help analysts or researchers to understand the most pressing issues and needs, and serve as a guide for further studies.[70]

### SAMPLE SIZE CALCULATION

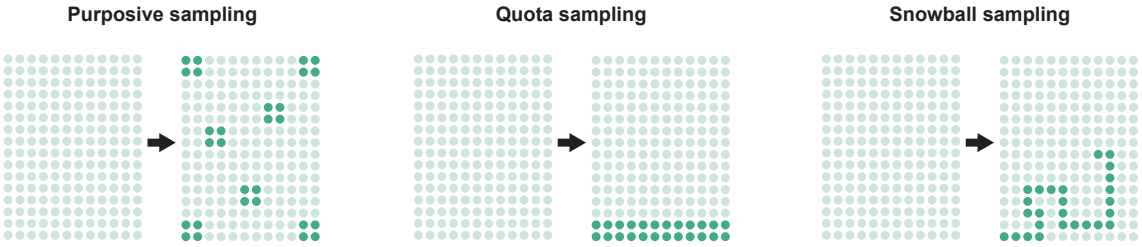
In sampling that is not statistically relevant (non-probability sampling), the sample size is determined by the survey designer according to specific objectives and feasibility of a given sample size. The sample size should be large enough to enable proper representation of and diversity among the subjects to be sampled, but not so large as to make the information collected redundant.

For purposive sampling (the most commonly used method of non-probability sampling), the convention for household food security and livelihood assessments is to sample between 50 and 150 households for each reporting domain that the assessment aims to draw conclusions from.[71] As the goal in sampling is to capture diversity, the sample design must explicitly reflect the heterogeneous characteristics (geographic location, displacement status, livelihood group, etc.) of the main indicators measured in the population of interest.

The **reporting domain** is the domain that you would like to summarize findings and report back on. For example, if the report is summarizing data on residents and IDPs separately, the two reporting domains are residents and IDPs.

### NON-PROBABILITY SAMPLING METHODS

Three methods are commonly used in economic security surveys: purposive sampling, quota sampling and snowball sampling.




Additionally, convenience sampling might be used in rare cases where no other alternative exists.

70 IASC, 2012.

71 ACF, 2010; WFP, 2009.

## PURPOSIVE SAMPLING

In purposive sampling, a sample representing a cross-section of regions and populations of interest is drawn; these regions and populations are selected on the basis of characteristics defined by the analyst. Accessibility, gaps in available information/knowledge, and specific regions and/or populations of interest: these are some of the key subjects that are often considered. Purposive sampling is the method of non-probability sampling that is most preferred.

**When should purposive sampling be used?** Purposive sampling may be used during the first stages of a crisis in rapid assessments and/or evaluations, to collect detailed data that complement less in-depth or detailed data collected at a statistically relevant level (i.e. using probabaility sampling), and when probability sampling is not possible owing to resource or access constraints or lack of baseline sampling data. Purposive sampling is also very useful when conducting in-depth multi-sector assessments with complex questionnaires.

**How to do purposive sampling?** In purposive sampling, a sample is selected that represents the heterogeneous characteristics of the population on a large enough scale. There is no one single method of purposive sampling; the following, taken from the Assessment Capacities Project (ACAPS),[72] can be used in a wide variety of humanitarian studies.

1. **Identify the population of interest and the characteristics of its location**: Follow the directions in Chapter 3: **Analysis design** on page 28, in the section titled "Population of interest".
2. **Define the sample frame.**
3. **Identify the gaps in information**: Identify the populations and locations for which essential information is lacking, and that need to be prioritized in the study.
4. **Develop a site matrix**: Develop a matrix highlighting the combinations of sites to be visited according to population and location characteristics (as outlined in Chapter 3: **Analysis design** on page 28, in the section titled "Population of interest"), and identify possible sites. See columns one through three in the table below.
5. **Determine the sample size**: Determine how many samples can be surveyed and sites visited in the time available. Follow the guidance in Chapter 3: **Analysis design** on page 35, in the section titled "Feasibility". If, as a result of your calculations, you discover that there will not be enough time for the required number of questionnaires to reach a representative sample, the questionnaire can be simplified or the heterogeneity of the target population redefined. The more heterogeneous the area, the more sites that will have to be visited. It is important to decide how many sites to visit and how much information to accumulate: more sites will mean more detail, but too many sites might mean a surfeit of repetitious information. The sample size can be adjusted once you are in the field, on the basis of preliminary information.
6. **Site selection**: Using the matrix developed below, review all possible sites, and select the ones to visit, taking into consideration the number and accessibility of sites that need to be visited per location/population combination, the findings of previous assessments and secondary data reviews, and local knowledge of the context.

72 ACAPS, 2012.

**Table 11** - Template for site-selection matrix in purposive sampling.

| LOCATION CHARACTERISTICS | POPULATION CHARACTERISTICS | SITES WITH LOCATION AND POPULATION CHARACTERISTICS | JUSTIFICATION FOR SELECTION/ NON-SELECTION |
|---|---|---|---|
| Location type 1 | Population type 1 | Site 1 | Description… |
| | | Site 2 | Description… |
| | | Site 3 | Description… |
| | | … | … |
| | Population type 2 | Site 1 | Description… |
| | | Site 2 | Description… |
| | | Site 3 | Description… |
| | | … | … |
| | Population type 3 | Site 1 | Description… |
| | | … | … |
| Location type 2 | … | … | … |

**EXAMPLE**

A team would like to carry out an in-depth economic assessment of households in a region affected by a series of conflicts in 2012; the assessment will pay particular attention to aspects of food consumption, food production, nutrition, protection concerns and the impact of humanitarian programmes. The team hopes to learn more about the situation in light of the programmes already under way, in order to help adapt the response to it. The sample frame is a vast region in which over two million people are estimated to be living; many of these people are constantly on the move, either because they have been displaced by conflict or because of their nomadic way of life. The security situation is volatile, and the accessibility of people in need is in daily flux. The team has only a limited amount of time and resources, because they are also responding to the crisis with various forms of assistance. They use purposive sampling to select their sample, which involves the following steps:

1. **Identify the characteristics of the population of interest** – Four population types are identified according to the heterogeneity between and homogeneity within them, and the direct relationship between their way of life and the impact of the conflict on their household economic situation:
   P1 – IDPs: internally displaced persons
   P2 – Returnees: people who have recently returned to their places of origin
   P3 – Sedentary residents: residents of the country who have a fixed abode
   P4 – Nomads: people from the region who do not have a fixed abode.
2. **Identify the characteristics of the locations of interest** – Locations are stratified on two levels: urban or rural areas and livelihood zones (of which there are 5 in the area of interest). This was done because there may be similarities in the household economies of people living in each of these types of location; but there may also be great differences between them, and comparisons had to be made with baseline livelihood data from before the crisis. Data on rural and urban areas are taken from the government census, and data on livelihood zones are obtained from FEWSNET (http://www.fews.net/). There are 10 combinations of location type (e.g. livelihood z1 urban, livelihood z1 rural, livelihood z2 urban, livelihood z2 rural, etc.).

3. **Identify the sample frame** – Establish the location of these populations of interest. Their accessibility should be taken into account.
4. **Identify the gaps in information** – The review of secondary data revealed major gaps in information on animal rearing, which confirmed the necessity of including nomadic populations in the sample as well as aspects of animal rearing in the questionnaires/evaluations.
5. **Develop a site matrix** – The following matrix template is used to list the towns by location characteristics and by the four population types that have been identified. Information on the location of the various population types is collected from primary and secondary sources, and validated by field officers working in the region on a daily basis.

| Location | Site | IDP | Returnee | Sedentary residents | Nomads |
|---|---|---|---|---|---|
| Livelihood Z1 semi-urban | Town 1<br>Town 2<br>Town 3 | x | x | x<br>x | x<br>x<br>x |
| Z1 rural | … | | | | |
| Z2 semi-urban | … | | | | |
| Z2 rural | … | | | | |
| Z3 semi-urban | … | | | | |
| … | … | | | | |

6. **Determine number of questionnaires/interviews and population to sample in each site** – The team will need to conduct a series of household interviews amongst each population type (IDPs, returnees, sedentary residents and nomads), and key informant interviews among local authorities and traders and merchants. The team would like various wealth groups to be represented (of which 3 have been identified), as well as the population types. They are not included in the site matrix because baseline data by site are not available. With this in mind, it is estimated that a full day will be needed at each site to complete 16 household questionnaires (4 teams of 2 conducting interviews simultaneously), 2 market interviews and 1 interview with the local authorities.
7. **Determine the number of sites** –18 people are available for field work. These 18 will be divided into two smaller teams to enable greater coverage in the area. Taking into account logistical considerations and time for resting, it is estimated that each team can visit 20 sites over a five-week period. Thus, 2 teams working 20 days each at 1 site per day means that 40 sites can be visited. This will enable representation of 4 sites per location type (40 sites / 10 combinations of site type), and a total sample size of 640 households (40 sites X 16 households per site), 80 market interviews (40 x 2 markets per site) and 40 interviews with local authorities (40 x 1 authority per site).
8. **Site selection** – The matrix developed is used to identify 40 sites with 5 backup sites (one per livelihood zone). Inaccessible sites are eliminated, and sites with a more heterogeneous representation of population type (e.g. presence of IDPs, returnees, sedentary residents and nomads) are prioritized.
9. Once they are in the field, the teams work with local authorities, guides and others with local knowledge to identify the households to be interviewed, according to population type and wealth category.

## QUOTA SAMPLING

Quota sampling involves choosing and sampling people, households, institutions, etc. until a specific number has been reached. For example, a team would like to administer a questionnaire to four IDP households. It goes out and interviews four displaced households, which are selected either randomly or non-randomly (identified by the local authorities, the first four found, etc.). The degree of bias associated with this method depends on how the units for the sample are identified. Random selection helps to minimize bias.

**When should quota sampling be used?** Quota sampling may be used to administer questionnaires at sites identified in purposive sampling: for instance, when a fixed number of questionnaires has to be administered to a specific population group or at a specific location, and time permits only a certain number to be completed. Quota sampling ensures that a minimum amount of information is collected on key groups of interest within a larger sample (a minimum number of households headed by women, elderly people living alone, etc.).

**How should quota sampling be done?** Select people, households, institutions, etc. that fit certain criteria for the sample until the required number of samples is reached.

### SNOWBALL SAMPLING

In snowball sampling, a subject is identified according the criteria of the study, and then that subject recommends the next subject to be visited.

**When should snowball sampling be used?** Snowball sampling can be useful for studies undertaken in regions where no information is available (e.g. a zone that has been closed off to external visitors for an extended period of time, a rapid initial assessment during or after a disaster) and/or where populations are difficult to find and no one with local knowledge is at hand to guide the team.

**How should snowball sampling be used?** In the area of interest, select, for the first interview, the first person, household, institution, etc. that meets the criteria of the study. At the end of the interview, this person, household, institution, etc. identifies the next of his or her kind to be interviewed, and so on.

### CONVENIENCE SAMPLING

In convenience sampling, subjects that are easiest to contact/visit are chosen. This method can be extremely biased.

**When should convenience sampling be used?** Convenience sampling is recommended only for those rare cases when it is the only feasible option (only one road in a region is passable, only one contact is available, etc.).

**How should convenience sampling be done?** There are numerous ways of doing so: from visiting households, people or locations that are accessible, convenient and known to be responsive, to instant polls on websites.

## ANALYSIS OF SAMPLED DATA

Data analysis must take into account the sample design, not only in the methodological description but also in the calculation of statistics.

### SAMPLE WEIGHTS

If every sampling unit (person, households, etc.) has the same chance of being selected, then weights are not required in analysis. The data are considered to be "self-weighted" by the sample method. This is the case when samples are drawn by means of simple or systematic random sampling, proportionate stratification and probability proportionate to size, or PPS (cluster selection).

Sample weights are required for analysing data that are not self-weighted, such as those collected by means of disproportionate stratification and cluster selection without the use of PPS. These sample designs do not give each and every unit an equal chance of selection. For example, let us assume that equal allocation is employed to select 300 resident households and 300 IDP households out of a total of 7,500 resident households and 2,500 IDP households, and that the sample ratio is 50:50 (IDPs to residents) while the population ratio is 25:75 (IDPs to residents). In this case, creating descriptive statistics on the population as a whole based on the 50:50 sample ratio would result in bias: the results would be tilted toward the value of the variable associated with the IDPs.

## DESIGN WEIGHTS

Variables are weighted principally in two ways to compute descriptive statistics of disproportionately stratified data: by using 'design weights' or 'normalized design weights'.

**Design weights** (also known as 'pweights' in statistics) are essentially "the number of units represented by one sampled unit" (WFP, 2009). The formula is as follows:

$$W_h = \frac{N_h}{n_h}$$

where:

$W_h$ = design weight in sampling stratum h
$N_h$ = population of stratum h
$n_h$ = sample size of stratum h

Design weights inflate the number of sampling cases used in analysis, and imply a larger sample size than is the case. Statistical tests for differences and changes over time using design weights will then be misleading. To compensate for this, normalized weights are often used in statistical analysis.[73] Normalized weights are the design weight multiplied by the stratum's representation in the overall sample. The formula is as follows:

$$w_h = \frac{N_h}{n_h} \times \frac{n}{N}$$

where:

$W_h$ = design weight in sampling stratum h
$N_h$ = population of stratum h
$n_h$ = sample size of stratum h
$N$ = total population in the sample (all strata)
$n$ = total sample size (all strata)

## CHECK STRATUM REPRESENTATION IN SAMPLE DESIGN

The weights can be examined to understand the effectiveness in terms of representation of each stratum of the sample. The following are the simple steps to take:

1. **Design weight calculation** – First, multiply the sample size for stratum x by the design weight (column E). Then add up these figures for all the strata (total of column E). The value should be equal to the total population in the sample frame.
2. **Normalized weight calculation** – The average of the normalized weights set should equal 1. See column D in the example table below.
3. **Sample design** - In an effective sample design, the normalized weights will not deviate far from 1 (ideally between 0.5 and 1.5). Very large (greater than 2) or very small weights (less than 0.5) can decrease the accuracy of the results.[74]

73 Magnani, 1999.
74 WFP, 2009.

| STRATUM | A N | B n | C Design weight ($W_h$) | D Normalized weight (wh) | E n x $W_h$ |
|---|---|---|---|---|---|
| A | 1,500 | 100 | 15 | 1.00 | 1,500 |
| B | 1,000 | 100 | 10 | 0.67 | 1,000 |
| C | 2,000 | 100 | 20 | 1.33 | 2,000 |
| Total | 4,500 | 300 | - | 1.00* | 4,500 |

*Average of weights

**EXAMPLE**

In the scenario set in the Central African Republic (see the section titled 'Disproportionate stratification'), where disproportionate stratification was used, weights are required for statistics throughout the sample frame (e.g. food consumption levels of all beneficiaries).

1. Proportion of total population (column D) is calculated by simply dividing the population in each stratum by the total population.
2. Design weight is calculated by dividing the population in the individual stratum (column B) by the sample size of the individual stratum (column C).
3. Normalized weight is calculated by multiplying the design weight (column E) by the total sample size (col C, row 3) and dividing that by the total population (col B, row 3).

| A Strata | B Population | C Sample size | D Proportion of total population | E Design weight (Nh/nh) | F Normalized weight (Nh/nh) x(n/N) |
|---|---|---|---|---|---|
| IDPs | 5,176 | 223 | 0.4 | 23.21 | 0.87 |
| Residents | 6,672 | 223 | 0.6 | 29.92 | 1.13 |
| Total | 11,848 | 446 | 1.0 | - | Average = 1.0 |

## USING WEIGHTS IN DESCRIPTIVE STATISTICS

Weighting is particularly important in calculating averages. For example, let us say a random sample of 100 male and 100 female students was tested out of a total of 4,500 students, 3,500 of whom are male and 1,000 female. The sampling method used was disproportionate stratification, because the sample was not taken in proportion to the representation of men and women in the total student population.

The average for each stratum in this case can be calculated from the unweighted scores, as everyone had an equal chance of selection within his or her stratum. For example, the average score for the male sample is 6.8 and that for the female sample, 6.5. However, when calculating the overall average taking the unweighted average would result in inaccuracies, as women are overrepresented in comparison to men. There are two options in this instance: weight the variable and then calculate statistics or apply weights when averaging data from strata.

**Option 1 – Weighting the variable and then calculating statistics**

1. Calculate the weighted score by multiplying the value of the variable by the weight (see column D in the table below)
2. Take the average of the weighted score (column D) to arrive at a sample average of 6.73. If the average of the unweighted score is taken (column B), the result would be 6.65, which is biased towards women, owing to their overrepresentation in the overall sample.

| RECORD | A<br>Sex | B<br>Score | C<br>Normalized weight (wh) | D<br>Weighted score (Column B x C) |
|---|---|---|---|---|
| 1 | Male | 2 | 1.56 | 3.11 |
| 2 | Male | 2 | 1.56 | 3.11 |
| 3 | Male | 8 | 1.56 | 12.44 |
| 4 | Male | 9 | 1.56 | 14.00 |
| 5 | Male | 8 | 1.56 | 12.44 |
| 6 | Male | 7 | 1.56 | 10.89 |
| 7 | Male | 7 | 1.56 | 10.89 |
| 8 | Male | 7 | 1.56 | 10.89 |
| 9 | Male | 8 | 1.56 | 12.44 |
| 10 | Male | 10 | 1.56 | 15.56 |
| … | … | … | … | … |
| 101 | Female | 6 | 0.44 | 2.67 |
| 102 | Female | 7 | 0.44 | 3.11 |
| 103 | Female | 6 | 0.44 | 2.67 |
| 104 | Female | 6 | 0.44 | 2.67 |
| 105 | Female | 7 | 0.44 | 3.11 |
| 106 | Female | 5 | 0.44 | 2.22 |
| 107 | Female | 7 | 0.44 | 3.11 |
| 108 | Female | 8 | 0.44 | 3.56 |
| 109 | Female | 7 | 0.44 | 3.11 |
| … | … | … | … | … |
| 200 | Female | 6 | 0.44 | 2.67 |

**Option 2 – Applying weights when averaging data from strata**

An alternative method, particularly useful when data are available only at the stratum level and not for each individual record, is to apply weights when calculating averages. Following on from the example above, say that the data are tabulated as shown below:

| STRATA | A<br>N | B<br>n | C<br>Average score |
|---|---|---|---|
| Men | 3,500 | 100 | 6.8 |
| Women | 1,000 | 100 | 6.5 |

In order to calculate the overall average, the weight of each stratum must be taken into account. To do this, the following formula is used:

$$\text{Average} = (p_{h1} \times x_{h1}) + (p_{h2} \times x_{h2}) + \dots$$

where:

$p_{h1}$ = proportion of population in stratum 1 (Nh1/N)
$x_{h1}$ = average of variable in stratum 1
$p_{h2}$ = proportion of population in stratum 2 (Nh2/N)
$x_{h2}$ = average of variable in stratum 2
… repeated for each stratum

So following the example above, the average score for the overall population would be 6.73 calculated as follows:

$$6.73 = ((3{,}500/4{,}500) \times 6.8) + ((1{,}000/4{,}500) \times 6.5)$$

As in the example of averages based on variable weights, if the average of the unweighted scores was taken, the result would be 6.65. This is biased towards women because of their overrepresentation in the overall sample.

## FREQUENTLY ASKED QUESTIONS

**Do I need to sample a minimum proportion of my population of interest? If so, what is the minimum number of households that I should interview?**

The answer to the first part of the question is: yes; and to the second part: well, it depends. If you want your results to be statistically representative, you should follow the guide in calculating sample size based on the level of accuracy that you want for your analysis. This calculation is not made as a proportion of the overall population; it will depend on the precision required and the expected prevalence or standard deviation of the variable measured. Sometimes, the sample size required is even smaller than, say, 5 or 10%.

If your results do not have to be statistically representative (i.e. if you use non-probability sampling), the sample size should be large enough to enable proper representation of and diversity among the subjects to be sampled, but not so large as to make the information collected redundant. In this case, select the sample size by reporting unit.

**What do I do if certain sites selected for sampling become inaccessible in the middle of a field exercise?**

**EXAMPLE**

A team goes to the field to carry out a six-week assessment. They use purposive sampling to select 40 sites (4 samples of 10 unique combinations). During the second week of the assessment, fighting breaks out in a region that is part of their sample and 6 of the sites are no longer accessible.

First, during purposive sample planning, backup sites that meet the same criteria (e.g. displaced populations in region X) as the primary sites should always be selected in case of access issues during the exercise. If a site is not accessible, a backup site with the same characteristics can be chosen. When both primary and backup sites are not accessible, there are two options: 1) reframe the sample mid-exercise or 2) treat the sites as 'non-response'.

Where the area that has become inaccessible contains all possible samples of the population that meet certain criteria (e.g. all displaced populations in region X are inaccessible), reframing the sample would be an appropriate solution, as it is no longer possible to include certain population and location characteristics in the primary data collected in the field. The field teams should convene (if not physically, then over the internet or by phone) and review their original methodology for site selection and sampling; they should then adapt it to the new sample frame (which is now the original sample frame minus the inaccessible areas). All population characteristics and stratification methods (if used) should be taken into account. Any changes in the methodology, and any limitations as well, should be recorded in the assessment report.

If the inaccessible area is not the only location where a population meeting certain criteria may be reached (e.g. one site of displaced populations in region X is accessible, but only that one), then treating the sites as non-response would be appropriate. If that area was treated as 'non-response', the analysis should take into account the possibility of non-response bias related to that particular population and their unique characteristics being underrepresented in the sample, and to other effects directly related to that population's exclusion from the study (i.e. consider what components in the study would be significantly different if that population had been accessible). With this in mind, careful consideration should be taken in generalizing information back to the entire population of interest.

**If, after completing a certain percentage of a survey (e.g. 60%) employing probability sampling, findings are seen to be consistently the same, do I need to continue the survey until the entire 100% of the sample has been covered?**

**EXAMPLE**

A team is evaluating a livestock programme, in order to find out what percentage of beneficiaries had become more productive after a particular project/programme. A representative random sample was taken; after 60% of this sample was evaluated, the team learnt that 95% of that 60% had increased their production capacity by more than 45%. Does the team need to cover the entire sample, given the preponderance of the same result?

In order for the final results to have the desired degree of accuracy, the entire sample must be covered. If that is not done, not only will the accuracy of the results be compromised, but also the sampling method used will be called into question because, ultimately, every unit in the sample did not have an equal chance of being selected. If the results are the same, the objectives and the methodology of the exercise, and the information being collected, could be reviewed to ensure that the correct information is collected and the data-collection method sensitive to any trend that might emerge. Where results continue to be the same, purposive sampling with appropriate representation may suffice.

**How should I adjust the sample size when using more than one survey and/or data-collection method?**

**EXAMPLE**

A team would like to evaluate the success of certain microeconomic initiatives. Its focus will be to gauge success by project type and to identify the reasons for success or failure. The subjects under study are extremely heterogeneous in nature, resources do not permit in-depth interviews with every beneficiary and household visits are difficult because of the unpredictable security situation. The team therefore decides to attempt to contact all beneficiaries over the phone and a sample of beneficiaries at their homes with a more extensive questionnaire. What is an appropriate sample size for the in-person interviews?

First, the objective of each individual survey has to be fixed. Combining data-collection and sampling methods is an excellent way to expand the analysis, but without a plan you can end up with an accumulation of disparate information. Once the objective of each survey is established, then the appropriate sampling methods can be identified.

In the case mentioned above, one option is to develop a survey that collects data that shed light on the success or non-success of a project, via the criteria and certain other key contributing indicators that have been chosen to define or measure success and failure. Either the entire population (for small homogeneous populations) or a statistically representative sample can be used here, as the phone is a low-cost survey method. However, information collected over the phone is much less reliable than that acquired in person. Therefore, it is important to triangulate and supplement this information with the house visits. The house visits may also be an opportunity to probe deeper into certain issues. In this case, after completing the phone interviews, a preliminary analysis can be done to identify 1) the types of project that are successful/non-successful, and 2) the factors that may be contributing to their success/non-success. Subsequently, a purposive sample stratified by these key elements can be used for the house visits.

**How do I choose an appropriate sampling method and sample size if I want to use a 'control group' to compare characteristics?**

**EXAMPLE**

A team would like to evaluate the impact of livestock-health projects/programmes in a region inhabited by livestock herders; the livestock of around 40% of these herders have been vaccinated by the ICRC. The team want to know if areas covered by the livestock-health projects/programmes have a higher rating, measured by animal-health indicators, than those not covered. To this end, they would like their sample to include areas that the ICRC has not reached.

A control group is a group that is identical, or very similar, to the population of interest, but one in which any factors thought to have an influence on the variables being studied are removed (for instance, in the case above, this would be the population not receiving any assistance).

Using a control group in a humanitarian setting can be very difficult because of the large number of complicating factors. Take the case above: the areas not covered by ICRC vaccination programmes may be covered by other humanitarian or government livestock programmes; coping mechanisms among herders may vary; the livestock may be exposed to different kinds of disease; and so on. If a control group is used, every effort should be made to ensure that it is as similar as possible to the group of interest, with the exception of the influencing variable; and any difficulties in finding a pure control group, or the limitations of the one that is to be used, should be taken into account when planning the study.

In order to reach statistically relevant conclusions that can be generalized back to the population, the control group should be treated as its own sample frame and sample size calculated. If the analytical method chosen will be used to compare and report any differences that are statistically relevant, equal allocation can be employed and the formula for comparison studies used. If using one of these methods cannot not yield a relevant sample size, the questionnaire could be carefully reviewed (and possibly shortened for the control group). If a relevant sample size is still not possible, and it is believed that a control group is required, the study should proceed with caution, limiting conclusions to 'tendencies in the sample used and triangulating primary data with secondary data.

## NOTE

The graphic representations of probability and non-probability sampling methods are based on ACF's *Guide Méthodologique: Enquêtes de terrain – Echantillonnage*.

# REFERENCES

ACF, *Food Security and Livelihood Assessments: A Practical Guide for Field Workers*, April 2010. Available at: http://www.actionagainsthunger.org/publication/2010/04/food-security-and-livelihoods-assessments-practical-guide-field-workers.

ACF, *Guide Méthodologique: Enquêtes de terrain – Echantillonnage*, 11 November 2011. Available at: www.parkdatabase.org.http://www.parkdatabase.org/files/documents/0000_Echantillonnage_ACF.pdf.

ACAPS, *Technical Brief*: *Purposive Sampling and Site Selection in Phase 2*, March 2012. Available at: https://www.acaps.org/resources.

IASC, *Multi-Cluster/Sector Initial Rapid Assessment (MIRA)*, March 2012. Available at: https://docs.unocha.org/sites/dms/Documents/mira_final_version2012.pdf.

ICRC EcoSec, *Technical Brief: Sampling*, August 2013.

ICRC EcoSec, Sample Calculator v2.3, March 2015. Available at: http://intranet.gva.icrc.priv/ecosec/topics/data-and-analysis/index.jsp.

Magnani, Robert, *FANTA: Sampling Guide*, December 1999. Available at: http://www.fantaproject.org/monitoring-and-evaluation/sampling.

MSF, *Nutrition Guidelines,* 1995. Available at: http://www.ennonline.net/fex/12/revised.

Scheuren, Fritz, *What is a Survey,* 1997. Available at: https://www.whatisasurvey.info/overview.htm.

SMART, *Sampling Methods and Sample Size Calculation for the SMART Methodology,* June 2012. Available at: http://smartmethodology.org/survey-planning-tools/smart-methodology/.

Trochim, William M.K, *The Research Methods Knowledge Base*, 3rd ed., Atomic Dog, 2006. Available at: http://www.socialresearchmethods.net/kb/measlevl.php.

Tulane University, *Practical Analysis of Nutritional Data*: *Survey Methods.* Online resource accessed in January 2015.

UN DESA, *Designing Household Survey Samples: Practical Guidelines*, 2005. Available at: unstats.un.org/unsd/demographic/sources/surveys/Handbook23June05.pdf.

WFP, *Comprehensive Food Security and Vulnerability AnalysisGuidelines*, 1st ed., 2009. Available at: http://documents.wfp.org/stellent/groups/public/documents/manual_guide_proced/wfp203208.pdf.

WFP, *Sampling Guidelines for Vulnerability Analysis*, December 2004. Available at: https://resources.vam.wfp.org/sites/default/files/VAm%20sampling%20guide%202004.pdf.

WFP, "Sample size calculation and development of sampling plan", Training presentation, 2010. Available at: https://resources.vam.wfp.org/Assessment-Tools/Sampling.

WHO, *The EPI Coverage Survey*, WHO, Geneva, 1991.

WHO, *The World Health Survey: Sampling Guidelines for Participating Countries*, WHO, Geneva, 2006.

# CHAPTER 6
# DATA TREATMENT

**Data treatment** refers to the handling, management, cleaning and when required, manipulation of raw data. Data come in many different formats, shapes and sizes; they may hold personal or sensitive information; they may have to be accessible at any given moment; they may be part of the historical record: all these factors need to be taken into account in the way data are 'treated'. This chapter touches on elements of each: it moves from the initial processing of data to consider such matters as the manipulation of data, data integrity and data security. The first part of the chapter focuses on structured data in tabular form. The last two sections, on data integrity and data security, are applicable to both structured and unstructured data.

## DATA PROCESSING

**Data processing** here refers to the methods for initially collating and manipulating data to meet analytical requirements.

### DATA REPLICATION

Data replication refers to the consolidation of new data, or the updating of existing data, collected by one or more users employing the same data-entry template at one central location. In a connected database environment, this is usually performed by the database software through a network – as is the case for some of the ICRC's main databases (PROT6, EPMT, WPA, FSS, etc.). In a disconnected environment, this usually has to be done manually, which is the case for many of the ICRC's ad hoc data-collection exercises.

In disconnected environments, files can be sent to a central location (through email, via USB keys, etc.) and then consolidated in one file. This process is much easier if data are collected using a standard collection form (like the one described in Chapter 4 Primary-data collection). One way to consolidate multiple files is by simply copying and pasting. This method is adequate to the task when there are not too many files to consolidate. A macro such as RDBMerge (http://www.rondebruin.nl/win/addins/rdbmerge.htm) can be used when there are a large number of data files to consolidate.

### CHECK ROWS AND COLUMNS

When you look at a data set for the first time, you must first study the structure of the rows and columns. If the data storage sheet was created by the person processing the data, this will entail no more than checking if data were entered correctly. If the data were reported in a format unknown to the person processing them, it will take him or her a short while to get to know the data. It should be apparent that the following rules have been followed:

- one record should take up only one row; and
- each variable should have its own column, and multiple-choice questions with many possible **responses should be given a unique column for each possible response.**

**EXAMPLE**

The example below is a typical case of one record taking up more than one row (see row 21). Here comments are given their own row, i.e. their own 'record'. As for any constant or variable, comments should be given their own proper column. It may be more appealing visually to have them directly under the data they are associated with, but that will cause problems at the analysis stage.

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1. INTRODUCTION | | | | | | 2. DEMOGRAPHICS | | | |
| 2 | UniqueID | Enumerator | Date | District | Place | Questionnaire # | Head of HH Sex | Head of HH Marital Status | HH Status | Adult Members |
| 3 | Automatic | Domain | Date (DD.MM.YY) | | Domain | Number | Domain | Domain | Domain | Number |
| 4 | UniqueID | Enumerator | Date | District | Place | Quest | HHHSex | HHHMarital | HHStatus | MemAdult |
| 20 | Ningerum16ND | ND | 14.01.14 | North Fly | Ningerum | 16 | Female | Widow | Resident | 2 |
| 21 | Ningerum1RG | RG | 14.01.14 | North Fly | Ningerum | 1 | Male | Married | Returnee | 2 |
| 22 | Ningerum2RG | RG | 14.01.14 | North Fly | Ningerum | 2 | Male | Married | Resident | 2 |
| 23 | Ningerum3RG | RG | 14.01.14 | North Fly | Ningerum | 3 | Male | Married | Resident | 2 |

## MULTIPLE-CHOICE DATA

Multiple-response data (where more than one response is possible) should have one column for each response. This makes it easier to calculate descriptive statistics such as frequency. The example below is taken from a monitoring and review exercise in connection with a microeconomic initiative in Iraq: one of the variables of interest in this instance was the difficulties, if any, that beneficiaries faced during the project. The question in the questionnaire allowed for more than one possible response.

| 09. Difficulties (problems) the beneficiary faced during implementation (encircle up to 5 of offered below): | | | | | | |
|---|---|---|---|---|---|---|
| Weather | Inputs | Production | Location | Debt | Second job | Market / sale |
| Skills /Knowledge | Legal issues | Illness | Quality of received/bought goods | | other* | |

The variable was stored in the database as a dichotomous variable ('yes' if it was selected or 'no' if it was not), where each possible response was allocated its own column. The data could also have been stored as '1' for 'yes' and '0' for no, thereby making it possible to use mathematical functions in frequency analysis.

| BL | BM | BN | BO | BP | BQ | BR | BS | BT | BU | BV | BW |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 09. Difficulties (problems) the beneficiary faced during implementation. | | | | | | | | | | | |
| 09a. Weather | 09b. Inputs | 09c. Production | 09d. Location | 09e. Debt | 09f. Second job | 09g. Market/sale | 09h. Skills/knowledge | 09i. Legal issues | 09j. Illness | 09k. Quality of received/ bought goods | 09l. Other *(specify)* |
| HHS_09a | HHS_09b | HHS_09c | HHS_09d | HHS_09e | HHS_09f | HHS_09g | HHS_09h | HHS_09i | HHS_09j | HHS_09k | HHS_09l |
| No | No | No | No | No | No | No | No | No | No | No | No |
| No | No | No | No | No | No | No | No | No | No | No | No |
| No | No | No | No | No | No | No | No | No | No | No | No |
| Yes | No | No | No | No | No | No | No | No | No | No | No |
| No | No | No | No | No | No | No | No | No | No | No | No |
| No | No | No | No | No | No | No | No | No | No | No | No |
| No | No | No | No | No | No | No | No | No | No | No | No |
| No | No | No | No | No | No | No | No | No | No | No | No |
| No | No | No | No | No | No | No | No | No | No | No | No |
| No | No | No | No | No | No | No | No | No | No | No | No |
| No | No | Yes | No | No | No | No | No | No | No | No | No |

Some electronic data-collection/entry tools store multiple-choice data in one record. In this case, the data need to be split to facilitate analysis so that each possible response is in fact its own distinct dichotomous variable. For example, the data below are responses to a question on debt where households with debt were asked what they used the debt money for. There were eight possible responses: food, health, buy household items, education, pay utilities, buy productive assets, pay rent or other. More than one response was possible. The data were collected using a mobile data- collection tool, and the output database put all responses into one variable, shown below.

| What did you the debt money for? |
|---|
| M_HH_DebtUse |
| HH_Item |
| Rent |
| |
| |
| Rent |
| HH_Item |
| Food,Health |
| |
| |
| |
| |
| Rent |
| |
| Food |
| |

From the data, you can see that one of the records has selected two choices: the used the money on both food and health. Below are two simple ways to do this in Excel.

**Using the IF and SEARCH functions**

- Create a new column needs to be created for each of the possible responses (food, health, buy household items, etc.)

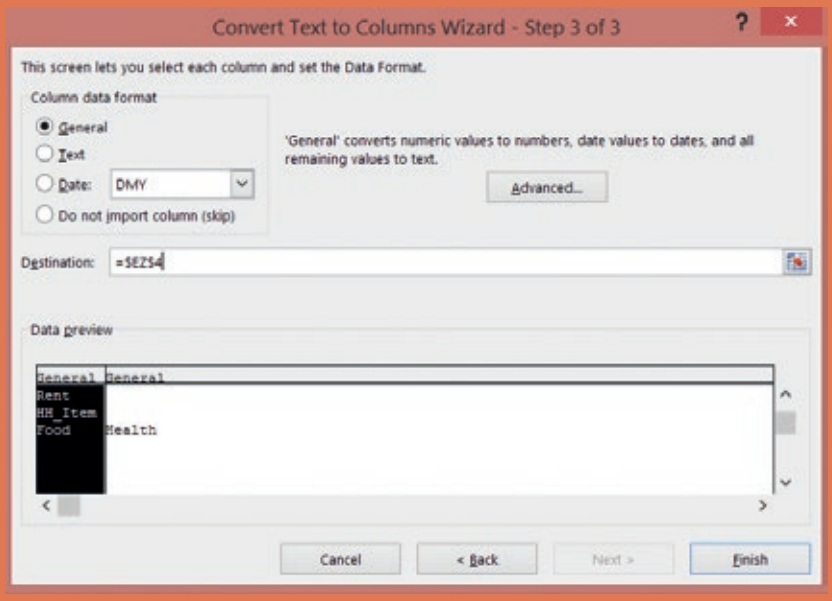
| EC | ED | EE |
|---|---|---|
| Derived variables | | |
| A Debt use food | A Debt use health | A Debt u |
| A_HH_DebtUse_Food | A_HH_DebtUse_Health | A_HH_D |

- For each new variable (column), use the IF and SEARCH functions following the logic below to tell Excel to SEARCH for the response (i.e. "Food") and give a value of 1 if it finds the response (e.g. food), else give a value of 0.

=IF(SEARCH("Food",EF4),1,0)

**Text to Columns Function**

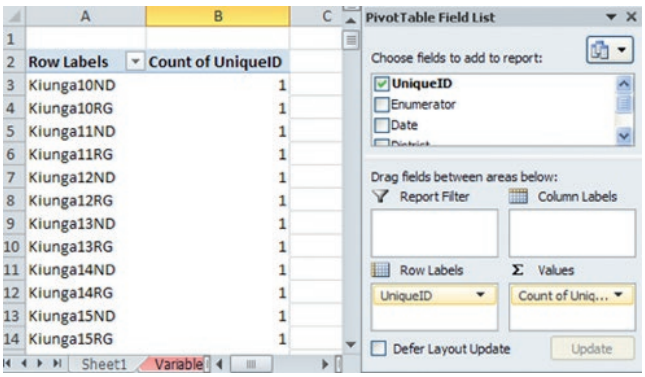
The section option is to use the 'Text to Columns' function in the Data Tools located in the Data tab of Excel. You need to tell excel the delimeter – the character used to separate each response – and where to put the results.




## CHECK RECORD ID

In every data set, there should be a unique ID for each record. Where that is not the case, these IDs should be added before the data are manipulated. See Chapter 4: **Primary-data collection** for more information on developing unique IDs.

If the data set does have a unique ID, check to confirm that all record IDs are in fact truly unique (i.e. no duplicates and no records without unique IDs). In Excel, uniqueness can be checked by means of filters, pivot tables or conditional formatting. Pivot tables can deal with larger data sets more rapidly. Simply create a pivot table with the unique ID variable as row labels and values. The count for each unique ID should be equal to 1. A count of more than 1 means that the unique ID was used more than once.



## RENAMING VARIABLES

Variables often have needlessly long names, usually associated with the question in the questionnaire. Have you ever seen a pivot table in which the name of the variable is something like this: *How many hectares of land did you cultivate last year?* Short names are much better for the purposes of analysis: whatever data processing and analysis software is chosen, it is this name that the software will use to refer to any data associated with that variable; and short names are obviously more convenient than long ones. In Excel, this name is the main ID for the variable when the Name function and pivot tables are used.

Creating short names for variables is a two-step process. First, insert a row below the row with the long variable name. Then, design a short name based on the following logic:

- each short name should be unique to that variable (i.e. no two variables have the same short name);
- short names should contain no spaces or special characters: +,",*,ç,%,&,",*,ç,%,#, etc.; and
- a short name should be the very last row before the first data record.

**EXAMPLE**

The example below shows category names in the first row, long variable names in the second, and short names in the third. The short names are, of course, short and unique, but they also follow a logic that makes it easier to work with them during analysis.

| AG | AH | AI | AJ | AK | AL | AM | AN |
|---|---|---|---|---|---|---|---|
| | | | | | 4.3. FOOD SOURCE | | |
| 8. Legumes, nuts & seeds | 9. Milk & milk products | 10. Oils & fats | 11. Sweets | 12. Spices, condiments, beverages | Starchy staples source | Fruits & vegetables source | Meat & fish source |
| FCLegumes | FCMilk | FCOils | FCSweets | FCSpices | FSStaples | FSFruitsVegetables | FSMeatFish |
| 1 | 0 | 0 | 0 | 0 | Production | Gift | |
| 1 | 0 | 1 | 1 | 1 | ICRC | Gift | |
| 1 | 0 | 1 | 0 | 0 | ICRC | Cash | |
| 1 | 0 | 1 | 0 | 0 | ICRC | Cash | |
| 1 | 0 | 1 | 1 | 1 | ICRC | Cash | |

## RESHAPING DATA

Reshaping is the term used to describe the process of transforming a data set from one format to another: long form to wide form, records from columns to rows, etc. The required shape – of the data – may depend on the type of analysis to be performed, the type of graphic that has to be created or the data's interoperability with certain statistical packages or databases.

### TRANSPOSITION

Transposition is the process of switching axes in a matrix: for example, switching the variables from rows to columns and vice versa. Data can be transposed in Excel by using the 'transpose' function under Paste Special.

**TRANSPOSE**

| | A | B | C |
|---|---|---|---|
| 1 | 0.1 | 0.1 | 0.2 |
| 2 | 3.4 | 5.6 | 7.2 |
| 3 | 2.0 | 2.2 | 3.1 |

| | 1 | 2 | 3 |
|---|---|---|---|
| A | 0.1 | 3.4 | 2.0 |
| B | 0.1 | 5.6 | 2.2 |
| C | 0.2 | 7.2 | 3.1 |

### LONG FORM VS WIDE FORM

Data in long form are data that keep a separate record for each individual instance of that record (i.e. in the example below, ID 1 has only one variable for year but there are multiple records for ID 1). Storing data in wide form means keeping only one record for any given record, which may have many instances associated with it (i.e. ID 1 below has only one record but four different variables for income).

Long-form data

| ID | Year | Sex | Income |
|---|---|---|---|
| 1 | 80 | F | 1,000 |
| 1 | 81 | F | 1,100 |
| 1 | 82 | F | 1,150 |
| 1 | 83 | F | 2,220 |
| 2 | 80 | M | 2,000 |
| 2 | 81 | M | 2,000 |
| … | … | … | … |

wide-form data

| ID | Sex | Income80 | Income81 | Income82 | Income83 |
|---|---|---|---|---|---|
| 1 | F | 1,000 | 1,100 | 1,150 | 2,220 |
| 2 | M | 2,000 | 2,000 | … | … |
| 3 | F | … | … | … | … |

In the example above, the long form of the data is useful for calculating statistics on years (e.g. other variables can be added for the year 1980 for ID 1), grouping data by income levels, etc. The wide form is useful for performing a time series analysis or creating a time series graphic (evolution of income levels by year).

## QUALITY CONTROL

Data should, before they are analysed, first be checked for errors. Some errors may not become apparent until analysis gets under way (e.g. an outlier that is not really one, but a typo instead); but it is best to catch as many as possible to avoid having to back-track during analysis. Listed below are some common errors and techniques for correcting them. Note that some of these errors can be mitigated by means of data-entry controls in the data-entry platform.

| ERROR | DESCRIPTION | CORRECTION TECHNIQUE |
|---|---|---|
| Outliers | An **outlier** is an observation that is at an abnormal distance from all other measures. A pre-screening for outliers should be done at the data-treatment phase to see if outliers are in fact real values or errors. | ▪ Conditional formatting |
| Gaps | There may be gaps in data that are associated with non-response (e.g. informants refused to respond, were not available to respond or did not know enough to respond) or with irrelevant questions (i.e. question was of no pertinence to the respondent). The best practice is to ensure that people using the data know what 'gaps in data' mean so that they can be analysed correctly. | ▪ Conditional formatting<br>▪ Find and replace<br>▪ 'Null value' is used instead of blanks, or if differentiation is required, 'non-response' and 'not applicable' are used. A common practice is to use -9999 or n/a for null values. |
| Zero values | Zero values will have a direct impact on any calculations preformed on a data set. For example, the average of 0 and 1 is 0.5 while the average of a blank value and 1 is 1. Care must be taken in this regard; otherwise, the resulting data analysis may be wrong. | ▪ Find and replace<br>▪ Zeros are never used in place of 'null value', but always used where the value is actually meaningful (a true value of 0) |
| Extraneous errors | Extraneous errors are associated with irrelevant data or information added to a particular record. If these data add nothing to the data set or analysis, they can be removed. | ▪ Manual review |

| ERROR | DESCRIPTION | CORRECTION TECHNIQUE |
|---|---|---|
| Duplicate records | Duplicate records are two or more records that are the same (e.g. the same household is listed twice or the same respondent's data recorded twice). | ■ Filters<br>■ Conditional formatting (in Excel, conditional formatting of duplicate values)<br>■ Pivot tables + count number of unique records |
| Geographic mishaps | Geographic data errors are common, particularly in paper-based surveys where data entry is open-ended. These are common occurrences:<br>■ The data collector confuses the difference between the city or town (exact location) and the administrative unit (region), particularly when the two have the same name<br>■ The city/town is attributed to the wrong administrative unit<br>■ The same place is given different spellings | ■ Mitigation in data entry through the use of cascading selects, either in mobile form or data-entry tool (see example on page 57).<br>■ Review data entry carefully with data collectors; mistakes are often made by the same data collector. |
| Skip logic incorrectly used | Particularly in paper-based forms, respondents may answer questions that are of no relevance. | ■ Create a derived variable[75] that indicates if the question was relevant, and whether it had to be answered (e.g. if the question is relevant only for households that farm wheat, create a derived variable that indicates if a household farms wheat or not). Filter to identify irrelevant questions, then review with data collectors why the question was answered. |
| Numerical entries do not add up | For example, when responses to multiple questions should add up to 100% (the sum of the percentages of food from own production, food purchased at market, etc.), but do not. | ■ Review manually for wrong entries |
| Miscalculations | Miscalculations are common, especially when a lot of data are being processed. Derived variables are often a source of difficulty in this regard. | ■ Check all calculations manually<br>■ Correct the calculation in one record and replicate this in all others (in the case of Excel, drag and drop) |
| Spelling mistakes | Spelling mistakes can cause problems during the data analysis phase, because it may result in one category being treated as several different ones. For example, if one person writes 'male' under 'head of household' and another writes 'man', the households that are categorized under 'man' may not be calculated if the keyword used for the calculation is 'male'. | ■ Manual review<br>■ Find and replace common spelling mistakes |

75 For more on derived variables, see Chapter 7 Quantitative analysis.

## DUPLICATE RECORDS

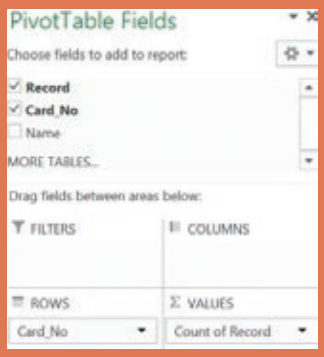
Some common examples of duplication in data sets for humanitarian work are listed below:

- A survey form is entered twice (or more) in the database by error.
- A mobile data-based survey is started and stopped, and then started again from the beginning. Say that this survey began with a particular household; when a member of the household arrived after the survey was under way, the household asked that the survey start again from the beginning. If the first record is not deleted in the field, the record will appear twice: first incomplete and then complete .
- When beneficiaries are being registered, some people or households may try to register themselves twice (or more).

This list is not exhaustive, and care should always be taken to check data sets for duplicate records before data analysis. Two techniques for checking for duplicates are described below.

### Checking for duplicates in one column

1. Create a pivot table of your data.
2. Drop the unique field (house number, registration number, telephone number, etc.) in 'rows' and the record ID in 'values'; the calculation performed is 'count'.




3. In the pivot table, see which fields have more than one record. These fields are duplicates (e.g. card number 23456 below has 2 records). Filters can be used in the original data set to examine deeper.

| Card number duplicates | |
|---|---|
| Row La | Count of Record |
| 12345 | 1 |
| 23456 | 2 |
| 34567 | 1 |
| 45678 | 1 |
| 56789 | 1 |
| 67891 | 1 |
| Grand Tot | 7 |

### Checking for duplicates in two columns

In some cases, a duplicate record can be spotted only by reviewing two or more pieces of data (e.g. name and phone number when more than one record could have the same name). This can be done quite easily by means of the COUNTIFS function in Excel.[76]

D2 =COUNTIFS($B$2:$B$8,B2,$C$2:$C$8,C2)

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | Record | Card_No | Name | Duplicate | | | | |
| 2 | 1 | 23456 | Abdoulaye | 2 | | | | |
| 3 | 2 | 34567 | Samuel | 1 | | | | |
| 4 | 3 | 45678 | Paul | 1 | | | | |
| 5 | 4 | 67891 | Peter | 1 | | | | |
| 6 | 5 | 56789 | Sonya | 1 | | | | |
| 7 | 6 | 23456 | Abdoulaye | 2 | | | | |
| 8 | 7 | 12345 | Paul | 1 | | | | |
| 9 | | | | | | | | |

See the example above, where the COUNTIFS statement counts the number of records that have the same card number and name. There are 2 records with the card number 23456 and the name Abdoulaye; all the other records have 1 card number and 1 name. Conditional formatting is used to highlight every cell in the Duplicate column with a value of more than 1, which enables easy identification of duplicates.

Other Excel techniques include the use of Filters and Conditional Formatting. There is also a tool to Remove Duplicates that is part of the Excel Tables tools. This should be used with great caution, as it deletes duplicates. Automatically deleting duplicates is problematic when a piece of data looks likes a duplicate but is not (duplication caused by error in data entry, unique ID not truly unique because of an error in registration cards, two respondents sharing the same phone number, etc.).

## DATA INTEGRITY

**Data integrity** refers to maintaining and assuring the accuracy and consistency of data over their entire life cycle.[76] Data are of value only if they are available for use and correctly interpreted when they are used. One initial step to optimize data integrity is to use proper methods of naming, sharing and storing files, and sound practices for recording metadata.

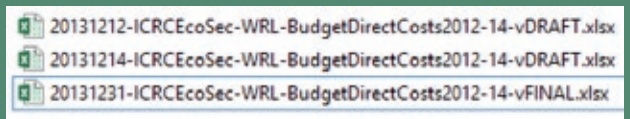
### NAMING FILES

Files should be given appropriate, informative and standard names to ensure they can be understood – and differentiated, one from another – by everyone using them. Ideally, when working with a team (e.g. unit or delegation) a standard format should be employed – a working language for naming files. The names of the files in a data set must include certain essential information, as shown in the table below:

| SUBJECT | FORMAT | EXAMPLE |
|---|---|---|
| Date | YYYMMDD | 20140422 |
| Source | Short text | ICRC EcoSec |
| Location | ISO3 country code unless code is not commonly used or ICRC site code if the document refers to one site in particular | COL, MMR |
| Subject | Short text | PDM, Assmt Report, Strategic Plan |
| Language version (if multiple versions) | Two-letter code | EN, ES, FR, RU, etc. |
| Document version | Draft or final | vDraft, vFinal |

**EXAMPLE**

Series of data sheets with a standard file name indicating date of file, source of data, location covered by data, subject, and version. This dating convention – year followed by month followed by day – is useful for sorting data in electronic archives.




76 Wikipedia, Wikipedia entry on "Data integrity", accessed in April 2015.

## SHARING AND STORING FILES

Files should be stored on and shared through standard ICRC platforms to ensure the following: the files are accessible to anyone who needs them at any given moment; there is a proper archive so that historical information can always be accessed; and, the files are properly protected.[77] Personal data and internal information should not be stored in or transmitted through external systems before an IT expert has made sure that the data are sufficiently secure. Files that are too large to transfer via ICRC email can be shared through the ICRC's AdHoc FTP service. For more information, contact the ICT staff member in the delegation.

An electronic file storage or sharing system is either a **document management system (DMS)** or a **content management system (CMS)**. The ICRC currently uses DocShare; it will be using SharePoint, a CMS, in the future. Check with your local chancellery to ensure that you have access to these systems.

## METADATA

Metadata are data about data: they include but are not limited to information on the source of the data, the date the data were collected and copyright information. Data are often useless without metadata.

In CMSs, metadata are collected when documents are uploaded to the system as part of free text and key words (sometimes referred to as tags). Data should however include metadata outside their CMS in case they are removed and shared outside the system.

**METADATA SHOULD INCLUDE AT LEAST THE FOLLOWING:**

- type of data or information
- date of data or information
- source of data or information
- location covered by the data or information
- all modifications to the data set
- any limitations the data or information may have, and – wherever relevant – comments on how they/it can and/or cannot be used
- any constraints with regard to copyright or sharing
- in connection with protection of personal data, the legislative frameworks the data are subject to
- for DMSs/CMSs, complete forms of keywords/tags, to ensure that data can be found.

77 These systems are located within the ICRC's secure IT environment, which follows the ICRC's rules on personal data protection.

# REFERENCES

FAO, *Guidelines for Measuring Household and Individual Dietary Diversity*, 2011. Available at: http://www.fao.org/fileadmin/user_upload/wa_workshop/docs/FAO-guidelines-dietary-diversity2011.pdf.

ICRC, *Information Handling Typology*, September 2012.

ICRC, *ICRC Rules on Personal Data Protection*, ICRC, Geneva, January 2016.

WFP, *Food Consumption Analysis: Calculation and Use of the Food Consumption Score in Food Security Analysis*, February 2008. Available at: http://www.wfp.org/content/technical-guidance-sheet-food-consumption-analysis-calculation-and-use-food-consumption-score-food-s

# CHAPTER 7
# QUANTITATIVE ANALYSIS

**Quantitative analysis**, for the purposes of this guide, refers to any analysis method that quantifies something, or measures or expresses something in numerical terms. Quantitative analysis uses both quantitative and qualitative data, and a variety of techniques. Quantitative methods are commonly used together with qualitative methods of data collection and analysis, and are often triangulated against each other. Quantitative analysis involves statistics, which may be descriptive or inferential: both descriptive and inferential statistics may involve the comparison of data between groups, across space or through time to uncover relationships, patterns and trends.

**Descriptive statistics** is used simply to describe a population in terms of measured values, whether it is a sample of a population or an entire population of interest. These values provide a simplified summary of characteristics: in the form of averages, maximums and minimums, proportions, etc. Unlike inferential statistics, descriptive statistics does not generalize or extrapolate the statistics back to a population larger than the one from which the data were collected. It may be the end analysis or the starting point for inferential statistics.

**EXAMPLE**

The average household dietary diversity score for displaced and returnee households interviewed (n=151) in three villages in the Hauts Plateaux of Kalehe, in the Democratic Republic of the Congo, is 3.45.[78]

**Inferential statistics** investigates models and hypotheses to make predictions or draw inferences about a population based on observations taken from a sample, or to test the probability of observed differences being true or false (or something in between), with a quantifiable level of confidence, precision and significance. These are some of the methods of inferential statistics that are commonly used in humanitarian work: extrapolation of estimates of means and proportions, chi-squared, t-tests, analysis of variance and regression analysis.

**EXAMPLE I**

The average household dietary diversity score of households in Essouk, in Mali, is 3.82 (95% confidence level, 3.5 to 4.14), and 30.8% have a score of less than 4 (95% confidence level, 20.5 to 41%).[79]

**EXAMPLE II**

The WHO predicts that if there are no changes in the control measures for the Ebola virus in West Africa, by 2 November 2014, the cumulative numbers of confirmed and probable cases are likely to look like this: 5,740 in Guinea, 9,890 in Liberia and 5,000 in Sierra Leone.[80]

**Comparative statistics** compares two or more subjects, processes or phenomena. It can be descriptive and/or draw inferences.

**EXAMPLE**

In Jordan and Syria, 1% and 3% of all households were food insecure; 15% of Jordanian and 18% of Syrian households were at risk.[81]

This chapter focuses on some of the quantitative methods that are commonly used in humanitarian work. Most of the techniques presented employ Excel tools, as Excel is the software generally used at the ICRC. Techniques beyond the scope of Excel are also presented; in this case, IBM's Statistical Package for the Social Sciences (SPSS) is used. The guide will use the Excel Analysis ToolPak, an add-in for Excel with techniques for advanced statistical analysis. It is standard with the software, but, to be displayed, it has to be added to the toolbar. For help in loading the ToolPak, visit https://support.office.com.

78 ICRC DRC, June 2014.

79 ICRC Mali, June 2014.

80 WHO Ebola Response Team, October 2014.

81 ACTED, August 2013.

## VARIABLES AND STATISTICS

The quantitative method or methods (statistics) to be used will be determined by the measurement scale of the variable.

**Table 12 - Variables and statistics**

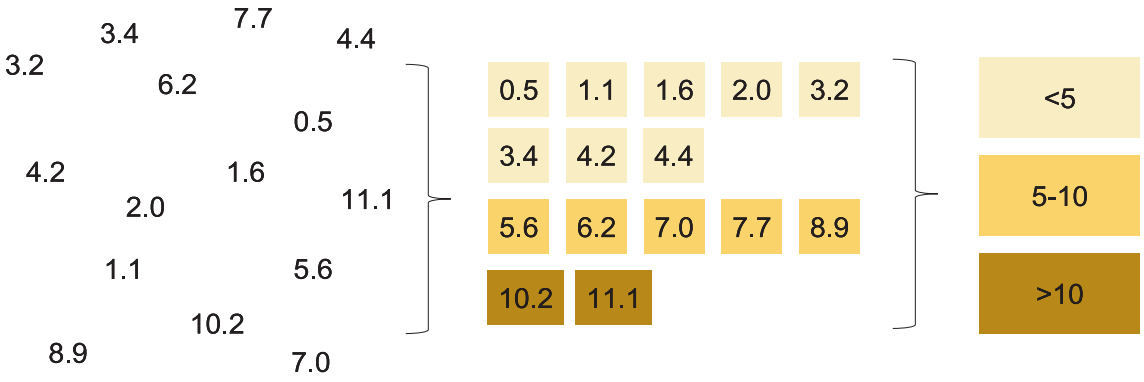
| VARIABLE | DEFINITION | EXAMPLE | STATISTICS |
|---|---|---|---|
| Nominal | Attributes are uniquely 'named' with no implied order. Nominal data with only two categories are dichotomous. | ▪ Phone number<br>▪ Red, blue, green<br>▪ Resident, Returnee, IDP, Refugee<br>▪ Man/Woman<br>▪ Yes/No<br>▪ Included/Excluded | ▪ Distribution (proportion, ratio and frequency) |
| Ordinal | Attributes can be ranked in a meaningful order with categories | ▪ High, medium, low<br>▪ 0-5, 6-10, 11-15, >15<br>▪ Likert scales | ▪ Distribution (proportion, ratio and frequency)<br>▪ Median and percentiles |
| Interval | Can only have a finite number of real values (whole numbers) and the distance between each number is meaningful but arbitrary (e.g. the distance between 1 and 2 may not be the same as between 2 and 3). Intervals do not have a true 0. | ▪ Temperature<br>▪ Time of the day<br>▪ Oedema | ▪ Distribution (proportion, ratio, frequency and rank)<br>▪ Median and percentiles<br>▪ Addition/subtraction-nalysis of variance |
| Ratio | Can have an infinite number of real values, and the value of 0 is meaningful. The distance between two numbers is the same (e.g. one person can be twice as tall as the other, 100 and 200 dollars is the same difference as 200 and 300 dollars, etc.). | ▪ Age<br>▪ Household size<br>▪ Height/weight<br>▪ Income/expenditure<br>▪ Distance<br>▪ Dependency ratio<br>▪ % of expenditure on food | ▪ Distribution (histogram and rank)<br>▪ Median and percentiles<br>▪ Addition/subtraction<br>▪ Mean, standard deviation and standard error of the mean<br>▪ Analysis of variance<br>▪ Coefficient of variation |

The table above displays the variables in a logical order to demonstrate how nominal variables are the least 'sensitive', and ratios the most sensitive, in terms of data analysis (i.e. the more sensitive a variable, the more types of test that can be run on it).[82] Proportions of nominal and ordinal data might be treated as ratios.

### GENERALIZATION

**Generalization** is the process of making data less detailed: all elements are grouped into new general categories; it could be said that generalization is essentially a form of categorization.

82 Trochim, 2006.




For example, let us say that in a particular survey, the Household Dietary Diversity Score is represented as an integer between 1 and 12. The real data are then expressed as one of 12 values. Data can, however, be generalized to groups that may be more useful to analyse, such as 'low score', 'medium score; and 'high score' – leaving only three possible 'values' for the data.

**EXAMPLE**

Data on the size of households are usually collected at discrete intervals. An analysis of household membership may just want to know how many households had 1-3 members, how many had 4-6 and how many had more than 6.[83] Most analytical software packages offer functions to generalize data. In Excel, the IF function is a simple tool. The following method can be used:

- Create a new column for the generalized category (e.g. HHSizeCategory).
- Use the IF function consecutively to define the result for each category. Start with the largest values (e.g. in this case, greater than 6) and move in decreasing order. The formula will look something like: =IF(K4>6,">6",IF(K4>3,"4-6",IF(K4<=3,"1-3"))).
- Drag the formula to calculate for all households (records) and check the results.

| J | K |
|---|---|
| Household size categorized | Total number of household members |
| HHSizeCategory | HHSize |
| 1-3 | 3 |
| 4-6 | 6 |
| 4-6 | 4 |
| 4-6 | 5 |
| 1-3 | 3 |
| >6 | 7 |
| 4-6 | 4 |

## CALCULATED DERIVED VARIABLES

A calculated derived variable is created from other variables by means of a calculated expression. For example, the Household Dietary Diversity Score (HDDS) is an indicator that is calculated from multiple variables asking whether a not a family ate food from a certain food group in the last 24 hours (one variable per food group).

There are numerous ways to calculate derived variables, but all depend on the way in which the variable is defined. A few examples of derived variables are listed below.

83 The cut-offs here are just examples; the actual choice of cut-off points will depend on the context and on the analytical objectives.

| DERIVED VARIABLE | CALCULATION |
|---|---|
| Household Dietary Diversity Score[84] | Sum of number of food groups eaten over the last 24 hours |
| Food Consumption Score[85] | Sum of scores calculated by multiplying the weight of a food group by the number of days, out of the last seven, on which the food group was eaten |
| Body Mass Index | Weight in kg *divided by* (Height in metres)$^2$ |
| Expenditure per capita | Total household expenditure *divided by* Number of household members |
| Population density | Population *divided by* Area |

**GOOD PRACTICES WHEN DERIVING NEW VARIABLES**

- **Create a new variable** (usually a column at the end of a table in a simple database) for your derived variable, which is separate from the base variables used to calculate it. Do not replace base variables with a derived variable. You may need the base variables for other analyses or you may have to update them later because of errors that you have found.
- **Preserve the link between the base variables and the derived variable.** This ensures that any updates or modifications to the base variables will be automatically reflected in the derived variable.

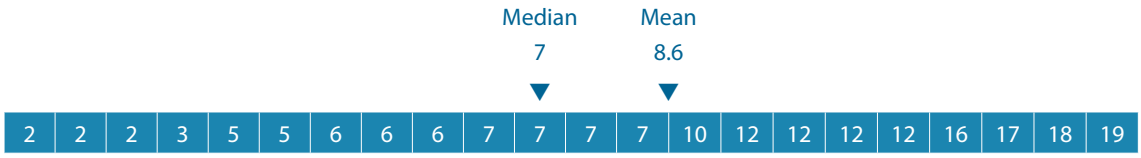
## DESCRIPTIVE STATISTICS

**Descriptive statistics** provide a simplified summary of characteristics – average values, maximums and minimums, proportions, etc. – to help describe data. The following section examines descriptive statistics of one variable – univariate analysis – including measures of central tendency, distribution and variance.

### CENTRAL TENDENCY

A measure of central tendency is a quantitative measure of the typical value among all values in a series of records. The most common measures of central tendency of numerical data are 'average', 'median' and 'mode'.

The mean (also referred to as the average) is the average of all values of the variable. The median is the central point of all values, or the value that is between 50% of lower and 50% of upper values. The simple graphic below shows the average and the median for a series of 21 values.



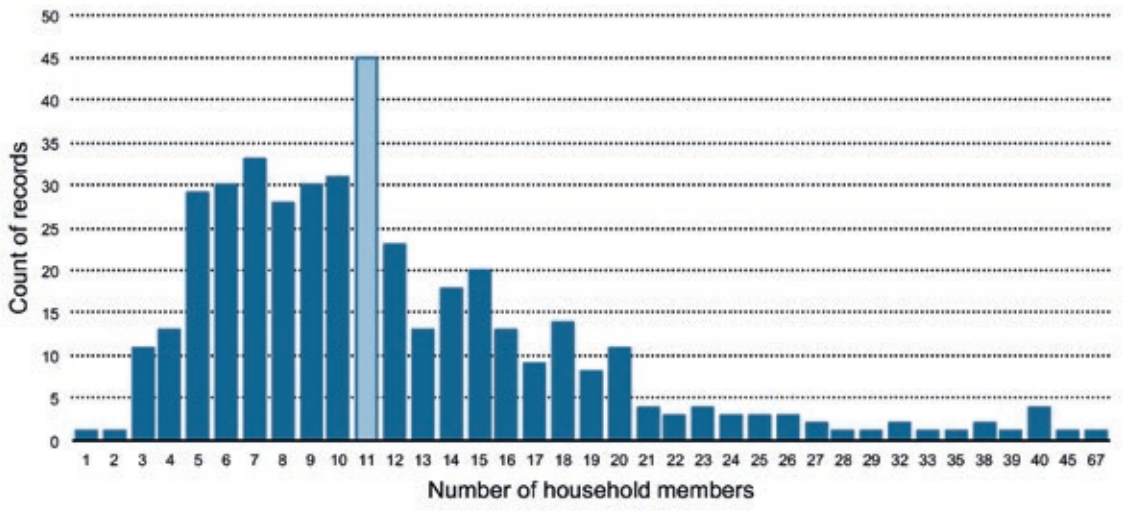

The mode is the value most frequently recorded for a variable. It is applicable to both integers (as shown above) and coded and/or categorized qualitative data. The following bar chart shows the count (number of records) with a given household size. We use a bar chart here because 'number of household members' is a discrete interval (i.e. we cannot have ½ a household member). The most frequently reported household size is 11 members (45 households reported having 11 members), which makes 11 the mode.

84 FAO, 2014.

85 WFP, 2008.

The structure of the data and the values of the variable(s) being analysed will determine what measure of central tendency should be used and how it should be calculated. The mode for the data in the graph above is 11. The average of the data is 12 and the median is 11. As the data are 'skewed' (i.e. affected by outliers as visualized in the bar chart), the most appropriate measure of central tendency in this case would be the mode or the median (also 11).

The table below shows, in a simplified way, when each measure of central tendency might be relevant, and the Excel techniques for calculating them.




| MEASURE OF CENTRAL TENDENCY | HOW TO CALCULATE IT IN EXCEL | WHEN TO USE IT |
|---|---|---|
| **Mean**<br>The average of all values; so, also called "average" | ▪ MEAN or AVERAGE function<br>▪ Pivot table | Variables that are normally distributed and that do not have a lot of outliers or abnormal values. |
| **Median**<br>The central point of all values | ▪ MEDIAN function | The median may be a more suitable measure, and may reflect the actual tendency more accurately, if the data are not normally distributed and/or the mean is heavily affected by outliers or abnormal values. |
| **Mode**<br>The most frequent value of the variable | ▪ MODE function<br>▪ COUNTIF function<br>▪ Pivot table | The mode can be useful when dealing with a limited number of integers or with qualitative data. It can, however, be misleading if there is one particular value that is abnormally represented, and a lot of variation in all other values. |

## DISTRIBUTION

**Distribution** is the process of arranging all the values for a variable to find the frequency with which they occur. The **frequency** is the count of records that have a given value or a value within a specified range (referred to as a group or class). When data are in the form of continuous variables, it is useful for understanding where they are concentrated within the full range of their values; and when data are in the form of categorical and discrete variables, it is useful for understanding the representation in each class.

## PROPORTIONS AND RATIOS

**Proportions** can be used to simply describe the representation of a key characteristic or value in relation to everything else. They can be reported as either a fraction or a percentage. **Ratios** are another way of expressing the representation of one key characteristic or value in relation to another. The example below shows how proportions (expressed as decimals and percentages) of male-headed households, and the ratio of male to female-headed households, are calculated.

| NUMBER OF HOUSEHOLDS | PROPORTION OF MALE-HEADED HOUSEHOLDS EXPRESSED AS A DECIMAL | PROPORTION OF MALE-HEADED HOUSEHOLDS EXPRESSED AS A PERCENTAGE | RATIO OF MALE TO FEMALE- HEADED HOUSEHOLDS |
|---|---|---|---|
| 57 female-headed<br>361 male-headed<br><br>Total 418 | 361 / 418 = 0.86 | (361 / 418) x 100 = 86.4% | 361 / 57 = 6:1 |

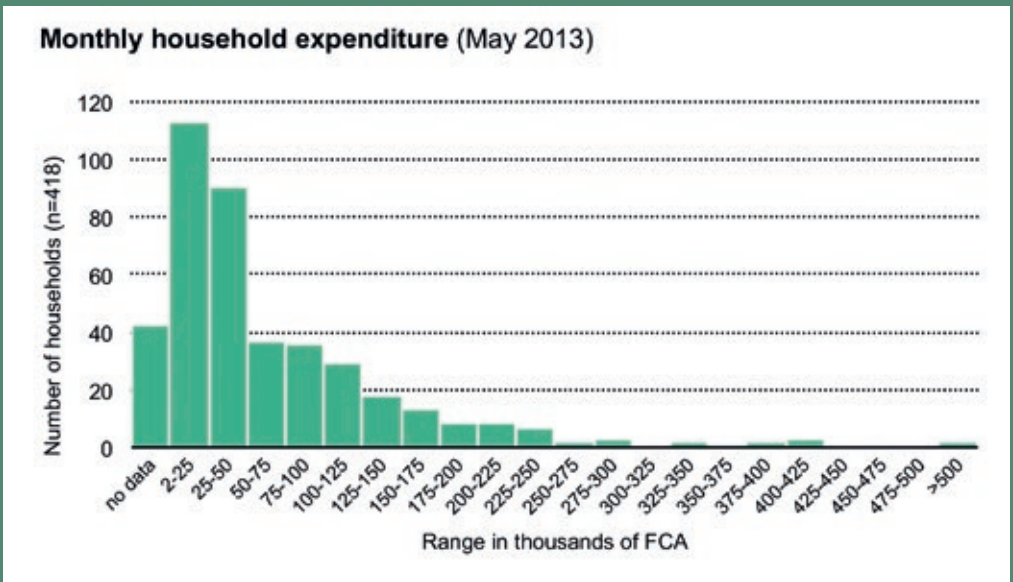
You can say that 86% of all households are headed by men, or that the ratio of households headed by men to that of households headed by women is 6:1. You could also say that *one in seven households was headed by a woman*. Whether to use a proportion or a ratio will depend on the analytical and communication objectives: on such questions as which figure is likely to have a greater impact on the audience: 86% or one in seven?

## HISTOGRAMS

**Histograms** are used to depict the frequency of continuous data. A histogram is a graphical representation of the distribution of data: the y-axis shows the frequency and the x-axis the values; the histogram may be accompanied by a table that shows the number of records for each range of values.

**EXAMPLE**

The following is a histogram of monthly household expenditure; it was derived from a household economy assessment in northern Mali. It shows both the range of values and where they are concentrated.



This histogram was created in Excel. The following technique, from the Analysis ToolPak in Excel, was used.

- Arrange the data in columns: one containing all records and values and a second containing an exhaustive list of possible values that you will use to calculate frequency for (called the 'Bin'; you can also think of them as buckets in which you will put all the records falling within a given range).

| Expenditures_1Month | Bin |
|---|---|
| 40'000 | 25'000 |
| | |
| 117'500 | 50'000 |
| 127'500 | 75'000 |
| 246'950 | 100'000 |
| 7'500 | 125'000 |
| 22'750 | 150'000 |
| 232'500 | 175'000 |
| 41'500 | 200'000 |
| 535'000 | 225'000 |
| 127'500 | 250'000 |
| 215'500 | 275'000 |
| 150'100 | 300'000 |
| 86'600 | 325'000 |
| 86'300 | 350'000 |
| 115'000 | 375'000 |
| 82'960 | 400'000 |
| 90'750 | 425'000 |
| 81'500 | 450'000 |
| 34'900 | 475'000 |
| 196'500 | 500'000 |
| 166'500 | |

- In the main menu, navigate to Data > Data Analysis.
- Select Histogram.
- Under Input Range, insert the range of cells representing your data.
- Under Bin Range, insert the range of cells representing your bins.
- Select your output and then select OK.
- The result is a table of the number of records in each 'bin' range (in the table below, the first column on the left). The default table created by Excel should be arranged so that it is understandable to anyone using the data. The range associated with the bin values should be clarified and what the frequency represents, spelt out.

| | | Monthly household expenditures (May 2013) | |
|---|---|---|---|
| *Bin* | *Frequency* | Range in thousands of FCA | Number of HHs (n=418) |
| 25'000 | 113 | no data | 42 |
| 50'000 | 90 | 2-25 | 113 |
| 75'000 | 37 | 25-50 | 90 |
| 100'000 | 36 | 50-75 | 37 |
| 125'000 | 29 | 75-100 | 36 |
| 150'000 | 18 | 100-125 | 29 |
| 175'000 | 13 | 125-150 | 18 |
| 200'000 | 9 | 150-175 | 13 |
| 225'000 | 9 | 175-200 | 9 |
| 250'000 | 7 | 200-225 | 9 |
| 275'000 | 2 | 225-250 | 7 |
| 300'000 | 3 | 250-275 | 2 |
| 325'000 | 0 | 275-300 | 3 |
| 350'000 | 2 | 300-325 | 0 |
| 375'000 | 1 | 325-350 | 2 |
| 400'000 | 2 | 350-375 | 1 |
| 425'000 | 3 | 375-400 | 2 |
| 450'000 | 0 | 400-425 | 3 |
| 475'000 | 0 | 425-450 | 0 |
| 500'000 | 0 | 450-475 | 0 |
| More | 2 | 475-500 | 0 |
| | | >500 | 2 |

**Figure 34** - Excel output (at left) and translated table (at right)

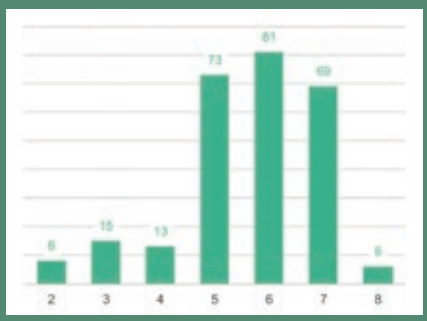
## FREQUENCY TABLES

**Frequency tables** can be used to explore frequencies for categorical or discrete variables (either ordinal or nominal), both quantitative and qualitative.

**EXAMPLE**

The frequency table and bar chart below show the Household Dietary Diversity Score (HDDS) – from 2 to 8 – for a sample of 265 households. The charts tell us that the majority of households (84%) have scores between 5 and 7. Tables and bar graphs are simple ways of depicting the frequency of qualitative data.

| HDDS | Frequency | Percentage of total |
|---|---|---|
| 2 | 8 | 3.0% |
| 3 | 15 | 5.7% |
| 4 | 13 | 4.9% |
| 5 | 73 | 27.6% |
| 6 | 81 | 30.6% |
| 7 | 69 | 26.0% |
| 8 | 6 | 2.3% |




Frequency tables are useful for analysing structured qualitative data, where responses are in the form of categories or could be categorized ('yes' or 'no', level of satisfaction, etc.), or discrete quantitative data (as in the example above). The construction of frequency tables will depend on the data (discrete quantitative variable, single-response or multiple-response categorical variables) and the way they are structured (long form or wide form). The two most commonly used methods in Excel for creating frequency tables are pivot tables, the COUNTIF function and the SUM function, depending on the structure of the data table as well as the preference of the analyst.

### DISCRETE VARIABLES AND SINGLE-RESPONSE CATEGORICAL VARIABLES

For discrete quantitative variables and single-response categorical variables (where only one response is possible), data are stored as one single variable (in one column); pivot tables are the easiest way to count the number of records associated with each group. The following technique is commonly used:

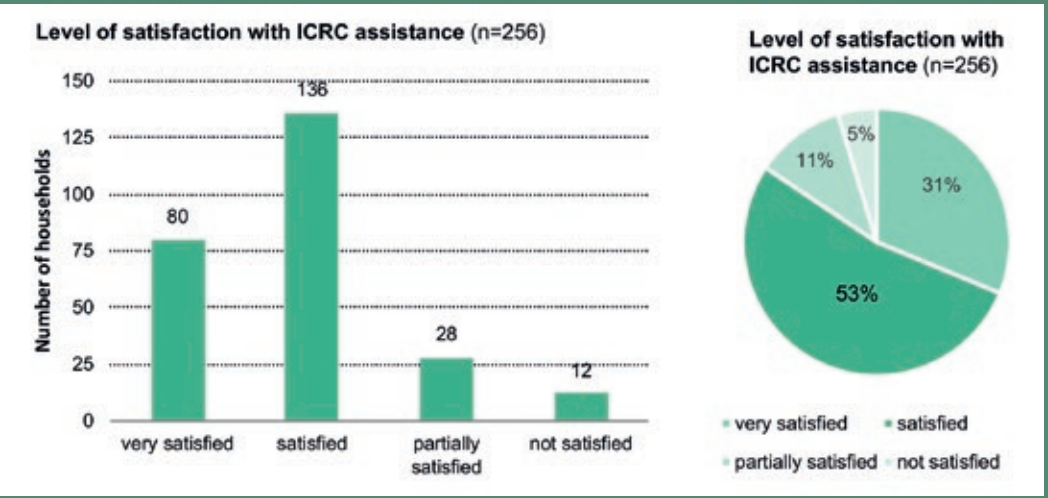
- Create a Pivot containing the qualitative variable for which you want to calculate the mode AND the unique ID for your records.
- In the pivot table, add the qualitative variable for which you need to calculate the frequency under Row Labels, and the unique ID variable under Values.
- Ensure that the value field setting is Count.
- Your pivot table will show a count of records for each category or discrete value.

**EXAMPLE**

The frequency table below represents data collected after a relief distribution in northern Mali; one of the variables was level of satisfaction with the food and household items. The frequency table below counts the number of households that reported on each 'level of satisfaction' with ICRC assistance.

| Level of satisfaction | n |
|---|---|
| Very satisfied | 80 |
| Satisfied | 136 |
| Partially satisfied | 28 |
| Not satisfied | 12 |
| **Grand total** | **256** |

The data may be displayed in a bar chart or a pie chart for a stronger visual representation of the results.




## MULTIPLE-RESPONSE CATEGORICAL VARIABLE

**Multiple-response data** (where more than one option in a list of options can be selected) are analysed slightly differently, as they are not part of a whole, but a series of possibilities. We add a reminder here that data are most easily analysed when they are stored as a series of dichotomous variables, and each response is its own variable (one column), with something to indicate whether it was selected or not (1 or 0, yes or no, X or blank, etc.). This was discussed in Chapter 4: **Primary- data collection.** If the data were collected and entered in long form, we recommend that they be reshaped into wide form, which can be more useful in frequency analysis.

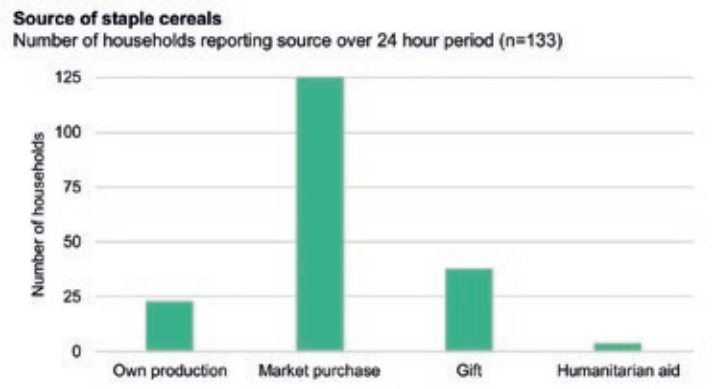
**EXAMPLE**

The following data are responses to a question on the sources of cereals; respondents could provide up to two sources and data collectors were to file their answers into one or more of six categories. Data were recorded in long form (columns FY and FZ). They were later reshaped during the data-treatment phase into wide form (columns GA to GD) using the IF and OR commands in Excel (see formula bar). Only four of the sources were preserved (A = Own production, C = Market purchase, D = Gift and F = Humanitarian aid) during the transformation into wide form, because no respondents selected B = Work-for-food, E = Gathered or other.

5.2 CONSOMMATION ALIMENTAIRE - Source de la nourriture
5.2101

| Céreales Source 1 Alim_Cere | Céréales Source 2 Alim_Cere | Céréales de source A Alim_Cere | Céréales de source C Alim_Cere | Céréales de source D Alim_Cere | Céréales de source F Alim_Cere |
|---|---|---|---|---|---|
| C | | 0 | 1 | 0 | 0 |
| C | | 0 | 1 | 0 | 1 |
| F | | 0 | 0 | 0 | 0 |
| C | D | 0 | 1 | 0 | 0 |
| C | D | 0 | 1 | 1 | 0 |
| C | D | 0 | 1 | 1 | 0 |
| C | | 0 | 1 | 0 | 0 |
| C | D | 0 | 1 | 1 | 0 |

The SUM function in Excel can be used to create a frequency table: the analyst simply takes the sum of all records (rows) for each possible response (column). The SUM function works in this case because the responses were recorded as numbers (1 for selected and 0 for unselected response); therefore, if you count the number of 1s you will get the number of responses for that category. The results can be shown in a frequency table or bar chart (see below).




Proportions and pie charts cannot be used to compare the different responses, as the data are not part of a whole. That being said, you can report that production was a source of staple cereals for 23 out of 133 households (or 17.23%).

**TAG CLOUDS**

A tag cloud is a visual representation of text data: here, each piece of text is weighted with its frequency in a data set. The weight is reflected in font size and sometimes font color. Tag clouds are also known as 'word clouds'. They are normally used for single words. They can be a useful tool for representing qualitative data visually in a somewhat quantitative manner.


**Figure 35 - Tag cloud of coping mechanisms reported during a household economic security assessment in northern Mali (ICRC Mali, July 2013).**

The tag cloud above was created with a free online tool Tag Crowd (tagcrowd.com). There are other tools as well, such as Wordly (http://www.wordle.net/), ToCloud (www.ToCloud.com), Tagul (tagul.com) and Tagxedo (http://www.tagxedo.com/).

### RANKING

Ranking is used to order data from smallest to largest, or vice versa. It is useful for putting records in context. For example, Libya has a Human Development Index (HDI) of 0.769. This may not mean much to someone who is not familiar with the index. However, Libya's HDI is ranked 64th out of 186 countries, which puts matters in context by telling us that while many countries have a higher HDI than Libya, Libya is still in the upper 35%.

If each value is unique (i.e. no two records have the same value), a simple 'sort' in Excel can be used to rank values from minimum to maximum or vice versa. If some records have the same value, they should have the same rank. This can be done in Excel, using the RANK function.

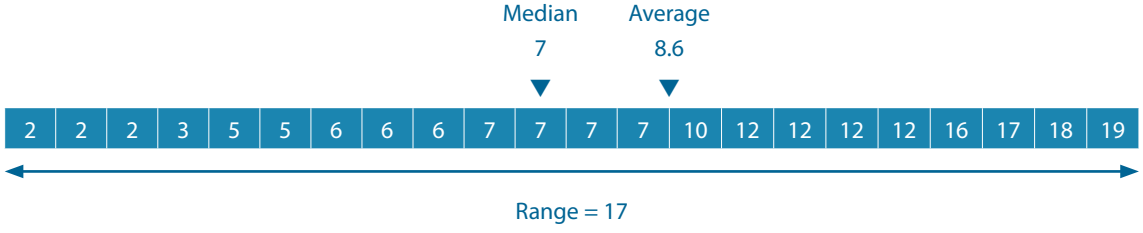
| TECHNIQUES TO CALCULATE DISTRIBUTION | HOW TO CALCULATE IT IN EXCEL | WHEN TO USE IT |
|---|---|---|
| **Histogram** Graphical representation of the distribution of data. | Histogram in Data Analysis ToolPak | Exploratory analysis of the distribution of continuous data |
| **Frequency table** Frequency of discrete or categorical data | Pivot table | Exploratory analysis of the distribution of discrete data Count of records of single or multiple-response categorical data |
| **Ranking** Data ordered from smallest to largest or vice versa. | COUNTIF function | Ordering quantitative data to understand the relationship between records. |

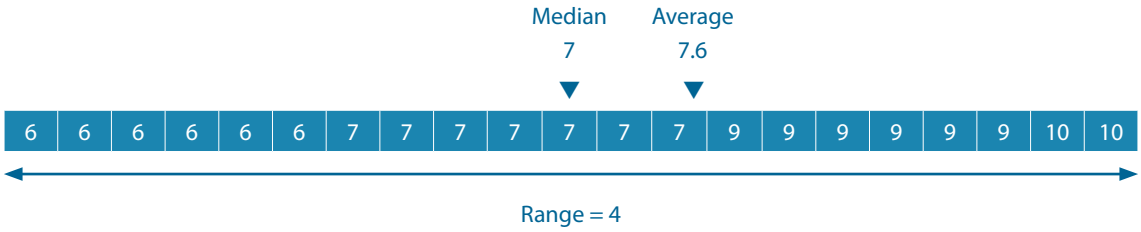
## VARIANCE

**Variance** quantifies how much the responses in a given variable are different from one another. It is extremely useful for evaluating the level of normality of the data, the extent to which the data fit the hypothesis or assumption stated at the beginning of an exercise or the value of data for making predictions. And it is helpful in answering questions such as these: Do the data have a lot of outliers? Are the data points consistent? Variance can be used with continuous data to supplement and complete measures of central tendency, as it can indicate the reliability of the mean, median or mode estimate. It is for this reason that variance is just as important as measures of central tendency.

### RANGE

The **range**, the simplest measure of the variance of data, is the difference between the highest and lowest values of the variable. It can provide an initial indication of dispersion. The following is an extension of the graphic on measures of central tendency; it demonstrates the relationship between median and average and range, and also shows how two data sets with similar medians and averages can have very different ranges. Which data set has more variability?

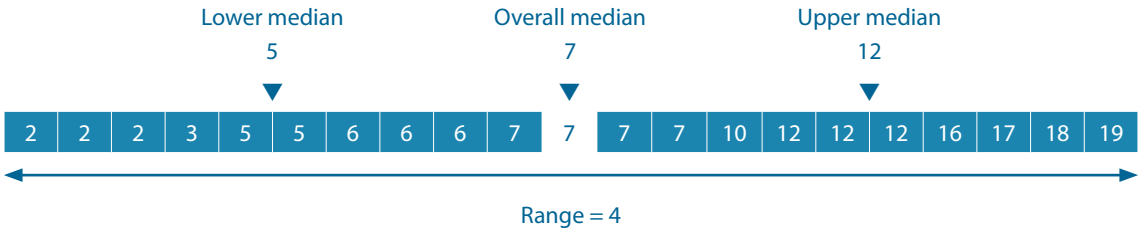





The range is the simplest measure of variance in terms of ease of measurement and understanding. However, it has its limitations. First, it looks only at the two extremes in the data set and ignores all other values; and, because of the way it is calculated, it is directly influenced by outliers. Second, it will be influenced by the sample size. One can imagine a range increasing with a larger sample (because of the greater likelihood of extreme values). Therefore, comparisons between ranges in two samples with different sample sizes can be misleading.

## INTERQUARTILE RANGE

The interquartile range is another simple measure of the variability of data: it breaks the data up into four equal quarters (quartiles) and calculates the range from the lower median (the median of the lower half of the data) to the upper median (the median of the upper half of the data). It is less influenced by outliers than the simple measure of range described above, but it has the same limitation: it does not consider all the data when calculating variability. The interquartile range for the first example above is 7: the upper median minus the lower median (see below).




There is no automated method in Excel for computing the interquartile range; but there is in certain statistical packages, such as SPSS. In Excel, this can be done manually: by ordering the data (e.g. lowest to highest), finding the overall median, then finding the median for the upper half and for the lower half of the data and finally, calculating the range between the lower and the upper median.

## VARIANCE AND STANDARD DEVIATION

The strongest measures of the variability of continuous data are variance and standard deviation, as they take into account all the values of the variable in question. **Variance** (sometimes denoted by the symbol σ2) measures dispersion from the mean. It is the average of the squared differences from the mean. **Standard deviation** (denoted by the symbol σ) is the square root of the variance. It is much easier to work with standard deviation than variance, as the number is in the same value range as the original data.

Standard deviation can be calculated using the STDEV function, where, like VAR, STDEV.P is used for exhaustive data (e.g. census) on an entire population and STDEV.S for a sample (e.g. sample of 100 households out of 1,000). In the example above, the variance of the data is 27.5 and the standard deviation 5.2.

A large standard deviation (relative to the data mean) can be understood to mean that records are at a distance from the mean (more variance in the data set). In this case, the mean may not be an accurate representation of the data. Likewise, a small standard deviation (relative to the data mean) can be understood to mean that records are not far from the mean (less variance in the data set). In this case, the mean may be an accurate representation of the data.

| MEASURES OF VARIANCE | HOW TO CALCULATE IT IN EXCEL | WHEN TO USE IT |
|---|---|---|
| **Range**<br>The difference between the highest and lowest value of the variable. | MAX - MIN function | The range is always useful for identifying outliers, getting a first impression of variation (distribution) and estimating how low and high values could be for continuous data. |
| **Interquartile range**<br>The difference between the upper and lower median of the variable. | Sort tool<br>MAX - MIN function after the data are divided into equal quarters | The interquartile range may be useful for showing, together with the range, how much a data set is influenced by outliers. The upper and lower medians can be reported together with the overall median. |
| **Standard deviation**<br>Square root of the variance. | STDEV function | Strongest measure of variance, and useful for checking whether continuous data are normally distributed and whether the average is a reliable estimate. |

### ALL-IN-ONE DESCRIPTIVE STATISTICS

The Descriptive Statistics tool in Excel's Analysis ToolPak is a simple tool for calculating, in one step, a broad range of descriptive statistics associated with a variable. To access the tool, navigate to Data > Analysis > Data Analysis. Then select Descriptive Statistics in the list of tools and fill out the appropriate parameters. In the example below, the tool was used to quickly calculate descriptive statistics for the monthly expenditure of the beneficiaries of microeconomic initiatives in Iraq. The unit of measurement is IQD. Interpretations of the results are in the right-hand column.

| Monthly expenditure: Descriptive statistics | | | |
|---|---|---|---|
| Mean | 474,467.6 | … | Mean or average |
| Standard Error | 26,077.6 | … | Standard deviation of the sample mean (σ/√n) |
| Median | 456,500 | … | Median |
| Mode | 240,000 | … | Mode |
| Standard Deviation | 340,010.2 | … | Standard deviation (σ) |
| | 115,606,923,50 | … | Variance (σ2) |
| Sample Variance | 3.31 | | |
| Kurtosis | 7.4 | … | Measure of the peakedness of the distribution |
| Skewness | 1.8 | … | Measure of the symmetry of the distribution |
| Range | 2,400,000 | … | Range = max – min |
| Minimum | 43,000 | … | Minimum |
| Maximum | 2,443,000 | … | Maximum |
| Sum | 80,659,500 | … | Sum of the values of all records |
| Count | 170 | … | The number of records for the variable[86] |

86 The number of records corresponds to the number of rows representing the unit of analysis (people, households, etc.).

## INFERENTIAL STATISTICS

**Inferential statistics**[87] use confidence intervals to make statements about the reliability and precision of estimates such as means, proportions and projections.[88] They calculate the interval within which a certain percentage of cases in the population would fall (the plus or minus figure reported in statistics). It helps answer questions such as these: How well does it actually represent the entire population of interest? What is the probable range of values?

**EXAMPLE**

For a household economy assessment in northern Mali, the Household Dietary Diversity Score (HDDS) was used as an indicator of food consumption. A probabilistic random sample of households in Essouk (in the Kidal Region) was taken. The average HDDS was 3.82. In order to ascertain the reliability of the estimate, the confidence interval on the mean was calculated at a confidence level of 95%. This was 0.32, which meant that if all the households in Essouk were measured, 95% of them would have an HDDS score between 3.5 (3.82-0.32) and 4.14 (3.82+.32).

Confidence intervals are usually measured at 90, 95 and 99% levels: in other words, if it is, say, 95%, it means that we can be 95% confident that the true value of the characteristic being measured (at population level) falls within the estimates based on the sample population. The confidence interval depends on the sampling distribution, the confidence level and the standard deviation (σ). The greater the deviation from the mean, the lower the confidence interval.[89]

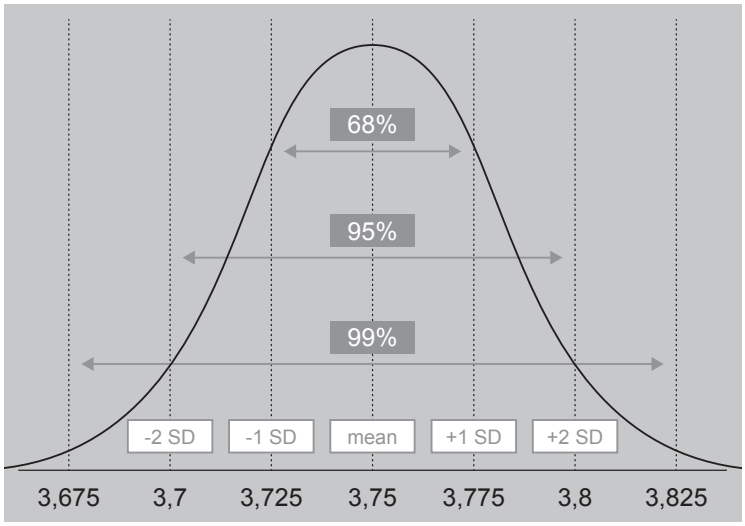



**Figure 36 - The graph above shows how, in a normal distribution, the confidence level increases with the standard deviation from the mean (i.e. the accuracy of the statistic increases as more units are included in the estimate). For example, 95% of cases are expected to fall within 2 standard deviations from the mean (in this case. between 3.7 and 3.8).**

### CONFIDENCE INTERVAL ON THE MEAN

There are two main ways in Excel to calculate the confidence interval on the mean. The first is by means of the CONFIDENCE function and the second is to check the option in the Descriptive Statistics tools in the Analysis ToolPak. The confidence interval thus arrived at is then added to or subtracted from the mean to calculate the possible range of values and the required confidence level.

87 Inferential statistics require data sets to be either exhaustive or probabilistic samples. Samples where each unit within the entire sample frame did not have the same chance of selection (i.e. non-probabilistic or non-random sampling) cannot be the subject of inferential statistics. See Chapter 5: Sampling for more information on drawing a probabilistic sample.

88 WFP, 2009.

89 Scheuren, 1997.

**EXAMPLE**

The table on the right, created using the Descriptive Statistics tool in the Excel Analysis ToolPak, shows the calculations for an indicator: in this case, the area of land cultivated by IDP households in the south-eastern section of the Central African Republic. The confidence interval – 0.0355 at the 95% confidence level (CL) – is highlighted in yellow.

This figure can be added to and subtracted from the mean to show the range around the mean (i.e.. 0.2316 + 0.0355 and 0.2316 – 0.0355). The analyst can use this to confirm the precision of the mean estimate. The statistic can be included in the report, or in an annex to the report, and stated like this: "The mean value is 0.23 hectares (+/- 0.03 (CL95%))", or "The mean value is between 0.19 and 0.27 hectares (CL95%)", so that the reader is aware of the reliability of the mean estimate.

| | A | B |
|---|---|---|
| 1 | *IDP_Land_Cultivated* | |
| 2 | | |
| 3 | Mean | 0.2316 |
| 4 | Standard Error | 0.0180 |
| 5 | Median | 0.16 |
| 6 | Mode | 0 |
| 7 | Standard Deviation | 0.2986 |
| 8 | Sample Variance | 0.0892 |
| 9 | Kurtosis | 9.3067 |
| 10 | Skewness | 2.5354 |
| 11 | Range | 2 |
| 12 | Minimum | 0 |
| 13 | Maximum | 2 |
| 14 | Sum | 63.4584 |
| 15 | Count | 274 |
| 16 | Confidence Level(95.0%) | 0.0355 |

## CONFIDENCE INTERVAL ON A PROPORTION

Confidence intervals on a proportion estimate are calculated in a similar way (e.g. added to and subtracted from the estimated proportion); but they have to be calculated manually in Excel, where the z-score associated with the desired confidence level (see conversions below) is multiplied by the standard error. The following formula – of which the second part is the formula for calculating the confidence interval of the proportion estimate – can be used:

$$CI = z * \text{sqrt}\left[\frac{p * (1-p)}{n}\right]$$

where:

z = z-score associated with the desired confidence level (99%CL = 2.576, 95%CL = 1.96, 90%CL = 1.645)

p = sample proportion

n = sample size

This interval is then added to or subtracted from the estimated proportion to calculate the possible range of values at the given confidence level.

**EXAMPLE**

For the example above, the analyst calculates the proportion of the population that did not cultivate any land. The estimates show that 69.3% of the IDPs cultivated zero hectares, or no land. Using the formula above, the confidence interval is calculated at

$$CI = 1.96 * \text{sqrt}\left[\frac{.6934 * (1-.6934)}{274}\right]$$

The result is 0.0546, which can be added to and subtracted from the proportion to show the range around the estimated proportion (i.e. 0.6934+ 0.0546 and 0.6934 – 0.0546). The analyst can use this to confirm the precision of the proportion estimate. The statistic can be included in the report, or in an annex to the report, and stated like this: "69.3% of IDPs did not cultivate (+/- 0.05 (CL95%))", or "Between 63.9 and 74.8% of IDP households did not cultivate (CL95%)", so that the audience are aware of the reliability of the mean estimate.

| CONFIDENCE INTERVALS | HOW TO CALCULATE IT IN EXCEL | WHEN TO USE IT |
|---|---|---|
| **Confidence interval on the mean** The confidence interval around the mean statistics | CONFIDENCE function | Variables that are normally distributed and do not have a lot of outliers or abnormal values |
| **Confidence interval on a proportion** The confidence interval around the proportion statistic | Descriptive Statistics in Analysis Toolpak | The median may be more appropriate, and may be a more accurate reflection of the actual tendency, if the data are not normally distributed and/or the mean is heavily affected by outliers or abnormal values. |

## RELATIONSHIPS

A **relationship** is the correspondence, connection, or link between two or more variables of interest. Relationships between variables are explored to determine whether two or more variables are related and/or whether their relationships vary according to a pattern. A pattern in this case is a 'recurring' relationship or one that appears in a predictable manner. Relationships can be tested on two (**bivariate**) or more (**multivariate**) variables, and they may look at measures of association, correlation or causality.

**EXAMPLE**

A study on childhood malnutrition in rural communities in north-eastern India found that "[e]xposure to floods is associated with chronic growth retardation in Indian children, especially in those exposed at very early stages in life". The study does not establish a causal relationship, but indicates an association between floods and malnutrition. The underlying cause of malnutrition is thought to be the adverse effects of flooding on crop productivity.[90]

**Association** refers to the general relationship between two variables. **Correlation** is a measure of association: it measures the strength and the direction of the relationship between two variables. **Causation** measures the degree to which one or more variables can cause (or predict) the value of another or of others. Neither association nor correlation establishes causality (causal relationship). The method of analysis depends on the type of variable and, in surveys, the sampling method employed. Some commonly used methods are described below.

### CROSS-TABULATION

**Cross-tabulation** is used to explore and test the association between **two categorical variables** and their characteristics. It is a table in a matrix format that breaks down responses by discrete factors or categories, and counts the frequency of each in comparison with other variable(s). The matrix can be called a crosstab, cross-tabulation or a contingency table; the data values are usually reported as counts and/or percentages.

In order to create a cross-tabulation, data need to be organized in a table or spreadsheet, with each record assigned its own value, and the variables for comparison placed side by side. The general practice in cross-tabulations is to put independent variables in columns and dependent variables in rows, and for values to add up to 100% at the bottom of the columns for independent variables. Cross-tabulations can be regarded from two different angles (swapping rows and columns) when it is not clear which variable is dependent and which independent. One or both may be correct so long as they follow the data-collection logic, and provided the sample (if a sample was used) was drawn properly and the statistics reported correctly.

90 Rodriguez-Llanes, et al., 2011.

**EXAMPLE**

The following cross-tabulation uses data from interviews with the families of Lebanese returnees from Syria, conducted as part of a household-level vulnerability analysis. Two qualitative categorical variables were collected to analyse living conditions: housing type and risk of eviction. Two cross-tabulations were created to explore the relationship between the variables.

Table 1 examines the risk of eviction by housing type. It essentially tries to answer questions such as this: How many households occupying a whole house or apartment are at risk of eviction? Of the households in a whole house or apartment, 37.1% answered 'yes'; as did 26.5% of those in informal settlements and 24.1% of those in a separate room of a house or apartment.

**Table 1 - Household assessment:** Housing type and risk of eviction

| Risk of eviction | All | Housing type | | | |
|---|---|---|---|---|---|
| | | Whole house or apartment | Separate room of house or apartment | Informal settlement | Other |
| **Yes** | 81 | 49 | 14 | 18 | 0 |
| % within response | 30.9% | 37.1% | 24.1% | 26.5% | 0% |
| **No** | 181 | 83 | 44 | 50 | 4 |
| % within response | 69.1% | 62.9% | 75.9% | 73.5% | 2.2% |
| **Total** | 262 | 132 | 58 | 68 | 4 |
| | 100% | 100% | 100% | 100% | 100% |

Table 2 below looks at the same data as Table 1 from a different angle. It examines housing type by risk of eviction, and answers this question: How many households at risk of eviction are in a whole house or apartment, a separate room of a house or apartment, an informal settlement or somewhere else? Of all the households at risk of eviction, 60.5% are in a whole house or apartment, 22.2% in an informal settlement and 17.3% in a separate room of a house or apartment.

**Table 2 – Household assessment:** Risk of eviction and housing type

| Housing type | All | Risk of eviction | |
|---|---|---|---|
| | | Yes | No |
| Whole house or apartment | 132 | 49 | 83 |
| % within response | 50.4% | 60.5% | 45.9% |
| Separate room of house or apartment | 58 | 14 | 44 |
| % within response | 22.1% | 17.3% | 24.3% |
| Informal settlement | 68 | 18 | 50 |
| % within response | 25.9% | 22.2% | 27.6% |
| Other | 4 | 0 | 4 |
| % within response | 1.5% | 0% | 2.2% |
| **Total** | 262 | 81 | 181 |
| | 100% | 100% | 100% |

From Table 2 we can conclude that the risk of eviction is highest for those living in a whole house or apartment. At that time in Lebanon, inability to pay rent was thought to be one of the main causes of eviction, so the fact that people who had to pay more in rent were at greater risk makes sense.

In the example above, the vulnerability analysis made use of non-sampled data: conclusions were reached only about the families interviewed (and not generalized back to a larger population), and were intended to provide a directional basis for future studies (what to look for in terms of household vulnerability) or for comparison with the results of other studies. When sampled data are used – that is, when samples are drawn using probabilistic sampling, and where the aim is to extrapolate results to a larger population of interest – the results can be tested for their statistical significance (e.g. to find out whether the relationship is due to sheer chance or whether it is statistically significant). The most commonly used test in this case is **Pearson's chi-square statistic**.

**EXAMPLE**

The following cross-tabulation uses data from household interviews involving beneficiaries of microeconomic initiatives (MEI) in Iraq. One of the objectives of the exercise was to learn more about the sustainability of the projects beyond the programme. Therefore, beneficiaries were asked if they expected the project to continue in the future. The table below shows the results by project type for a random sample of beneficiaries.

**Household interviews:** MEI project type and chance of continuation in the future

| Chance of continuation | All | Project type | | | | |
|---|---|---|---|---|---|---|
| | | Agriculture | Craft | Livestock | Service | Trade |
| **Yes** | 160 | 2 | 23 | 44 | 29 | 62 |
| % within response | 72.4% | 50.0% | 85.2% | 59.5% | 87.9% | 74.7% |
| **No** | 61 | 2 | 4 | 30 | 4 | 21 |
| % within response | 27.6% | 50.0% | 14.8% | 40.5% | 12.1% | 25.3% |
| **Total** | 221 | 4 | 27 | 74 | 33 | 83 |

The data show that the chance of continuation is highest for craft, service and trade projects, and less clear for agriculture and livestock projects. As the data were collected from a simple random sample of beneficiaries, the chi-square test is used to test the significance of the results. The following steps are required to perform the chi-square test in MS Excel:[91]

- Create a cross-tabulation that shows the frequency of observed responses by each category (as in the table above).
- Create a table showing the frequency of expected values if there was no relationship between the dependent and the independent variable. To do this, the total number of observations of category X of the dependent variable (in this case, the project type) is multiplied by the total proportion of "Yes" and total proportion of "No" for each category. For example, the expected values for 'Yes' for Agriculture are calculated by multiplying 4 x .724 and 'No' is calculated by multiplying 4 x .276.
- If the value is less than 0.05, then the null hypothesis[92] is rejected and the relationship considered to be significant.
- The Chi-square function (CHISQ.TEST) in Excel is used to calculate the chi-square, by means of the formula = CHISQ.TEST(observed values matrix, expected values matrix)

In this case, the chi-square is 0.0087, which means that the relationship between project type and chance of continuation is statistically significant.

**ASSUMPTIONS FOR CHI-SQUARE TEST**

- The two variables are categorical and each observation is assigned to only one category.
- The two variables are independent of each other (e.g. not matched pairs or panel data).
- The sampling method is simple random sampling.
- The sample should be sufficiently large (see next point). Small sample sizes may lead to a Type II error (false negative). They may also require the use of another test, such as Ficher's Exact Test.
- The frequency in each cell is at least 5 for all cells in a 2x2 table and 5 for 80% of cells for a larger table, with no cells having 0 frequency.
- Continuous variables may be generalized to intervals, but it should be based on theory or practice that is normally used in a given context.

## LINEAR CORRELATION

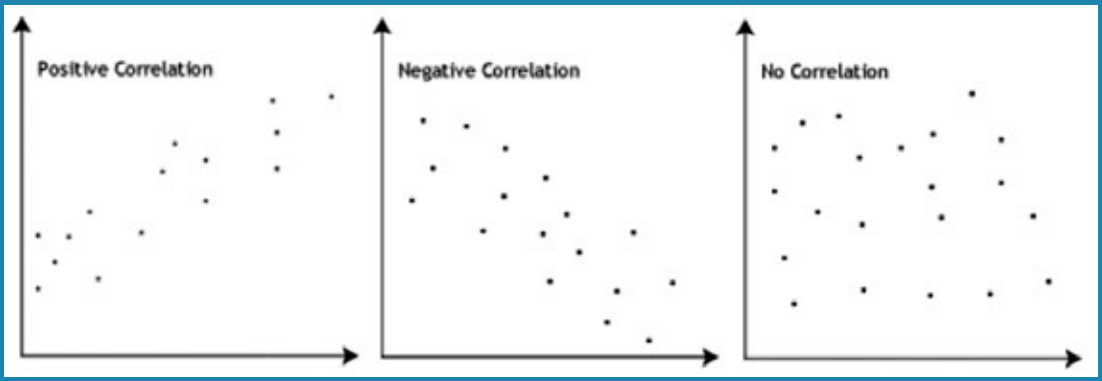
**Linear correlation** is a term used to describe the association between two normally distributed numerical variables thought to have a linear relationship. It can be measured in order to detect the presence of a relationship, or to evaluate the strength and the direction of such a relationship.

91 In SPSS ,this test can be performed under Analyze > Descriptive Statistics > Crosstabs…. > Statistics > Chi-square.

92 The null hypothesis refers to a general statement or default position that there is no relationship between two measured phenomena (Wikipedia, Wikipedia entry on "Null hypothesis", accessed in April 2015).

## SCATTER PLOT

A **scatter plot** is a simple graphic that is used to explore the relationship between two numerical variables. Scatter plots reveal patterns that can then be interpreted: Are the points arranged in an orderly pattern or scattered? Do they slope negatively, positively or not at all? Are there a lot of outliers?




Source: Laerd Statistics

Scatter plots are easily created with the charting tools in MS Excel. Each variable needs to be in its own column (side by side), and corresponding values should be adjacent to each other. The values in the left-hand column will be plotted along the x-axis, and those in the right-hand column, along the y-axis. The predictor variable is on the x-axis and the response variable on the y-axis: this is a convention that is generally followed.

### PREDICTOR VERSUS RESPONSE VARIABLE

A predictor variable is a variable that can influence the value of a response variable; and a response variable is a variable that can be influenced by the value of a predictor variable. As such, a predictor variable is an independent or control variable and a response variable, a dependent variable.

## BIVARIATE CORRELATION COEFFICIENT

The bivariate correlation coefficient (r) – or Pearson's correlation coefficient – is a statistic that describes the strength and direction of a linear association between two variables, where +1 is a perfectly positive correlation, 0 is no correlation and -1 is perfectly negative correlation.

### EXAMPLE

The following correlation matrix looks at the correlation between total monthly household expenditure (proxy for household income) and household dietary diversity in Léré, northern Mali. The matrix was created with the CORREL function in Excel (which calculates only the coefficient); the p-value was calculated using the Regression tool in the Analysis ToolPak.

**Table 13** - Correlation between total monthly household expenditure and household dietary diversity in Léré, Mali (July 2014)

| | **Total monthly expenditure** | **Household dietary diversity score** |
|---|---|---|
| Total monthly expenditure | 1 | |
| Household dietary diversity score | 0.1881* | 1 |

*Coefficient is statistically significant at the 5% level

The results show that there is a statistically significant positive relationship between the household dietary diversity score and total monthly expenditure: r =.1881 ($p<0.05$). If the outliers are removed (expenditure above 400,000), r increases to 0.2235 ($p<0.05$).

**ASSUMPTIONS OF CORRELATION COEFFICIENT**

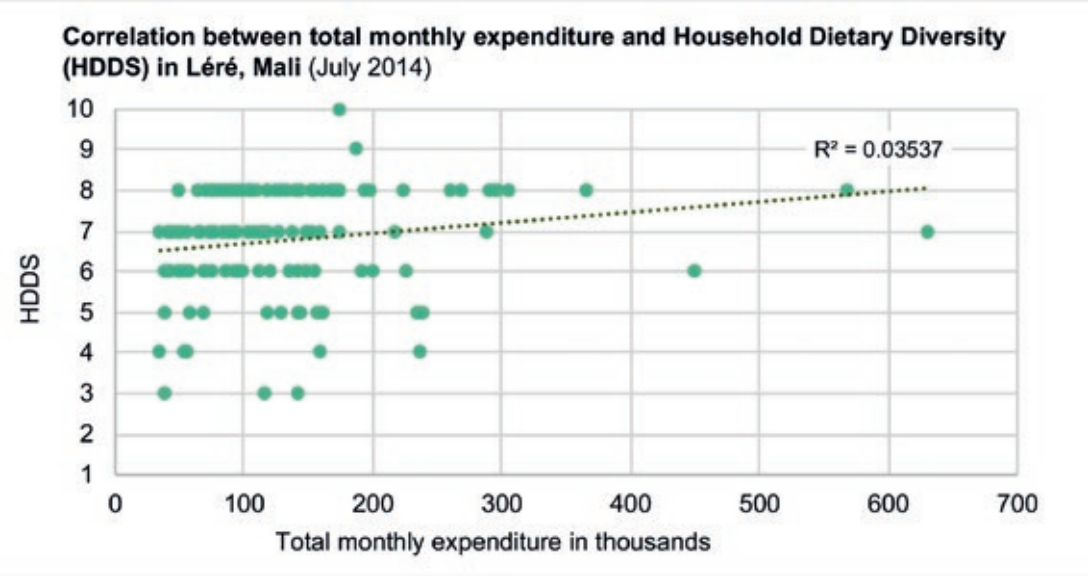
- Variable scale of measurement is an interval or ratio. If variables are ordinal, Spearman's rho may be a more appropriate statistic.
- Variable data are normally distributed.
- The association is assumed to be linear.
- Data are be free of outliers.

## COEFFICIENT OF DETERMINATION

The **coefficient of determination** ($r^2$) is the statistic used to examine how much of the variation in one variable can be explained by its relationship with the other variable: in other words, how much the value(s) of x can be used to predict the value(s) of y. $r^2$ supplements the interpretation of the results of r.

**EXAMPLE**

In connection with the example above, concerning the correlation between total monthly expenditure and household dietary diversity, the following scatter plot displays the data with household dietary diversity on the x-axis and total monthly expenditure on the y-axis. The scatter plot was created with the Scatter chart tool; the trend line and $r^2$ were developed by right-clicking the data points and selecting 'Add trendline'.




The trend line demonstrates the positive relationship described by the Pearson's correlation coefficient. In addition, $r^2$ shows that 3.5% of the variation in household dietary diversity can be explained by the linear relationship with total monthly expenditure. If the expenditure outliers are removed (expenditure above 400,000), $r^2$ increases to 0.049.

Studies in the social and behavioural sciences often yield small correlation coefficients. This is often attributed to the fact that variables may be influenced by a multitude of factors.[93]

93 Kenny, 1987.

**ALL-IN-ONE LINEAR CORRELATION**

The correlation coefficient (r) and the coefficient of determination ($r^2$), and their p-value, can all be calculated using the Regression tool in Excel's Analysis ToolPak. To access the tool, navigate to Data > Analysis > Data Analysis. Then, select Regression in the list of tools and fill out the pertinent parameters. The following example shows how the tool was used for the Léré data set on the correlation between HDDS and total household expenditure. The statistics of interest are in yellow: r is "Multiple R", $r^2$ is "R Square" and the p-value is the second p-value in the bottom table.

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.18806842 |
| R Square | 0.035369731 |
| Adjusted R Square | 0.028006141 |
| Standard Error | 93701.68393 |
| Observations | 133 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 42173238794 | 42173238794 | 4.80332711 | 0.030172365 |
| Residual | 131 | 1.15018E+12 | 8780005572 | | |
| Total | 132 | 1.19235E+12 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | 40300.51993 | 43069.81761 | 0.935702127 | 0.351148282 | -44901.85315 | 125502.893 | -44901.85315 | 125502.893 |

These can also be easily calculated using the SPSS correlation function or an online calculator such as this one from Social Science Statistics at: http://www.socscistatistics.com/.

## REGRESSION ANALYSIS

**Regression analysis** tries to estimate how one dependent (response) variable is influenced by one or many independent (explanatory) variable(s). It can be used to develop models on causal relationships, as part of prediction analysis (such as forecasting), in inferential statistics or to test hypotheses. One example of a regression is the coefficient of determination (discussed above), which quantifies the degree to which the variation in variable x is due to the linear relationship with variable y. This is one of the simplest 'regressions'. There are many others. Examples of regression used in humanitarian work are listed below:

- Coefficient of determination ($r^2$)
- Logistic regression
- Simple linear regression
- Multiple linear regression
- Multivariate multiple regression
- Ordinary least squares
- Canonical correlation analysis.

Regression analysis is a vast and complex field. Before models can be developed, a series of preliminary tests have to be carried out to understand the data. Regression analysis is beyond the scope of this guide, and is mentioned here only to illustrate the breadth of range of the analyses that can be performed. If assistance is required, a statistician should be sought out.

| RELATIONSHIP ANALYSIS | HOW TO CALCULATE IT IN EXCEL | WHEN TO USE IT |
|---|---|---|
| **Cross-tabulation**<br>Crosstab or contingency table in matrix format that breaks down responses by discrete factors or categories, and counts the frequency of each in comparison with the other | Pivot table | To explore and test the association between two categorical variables and their characteristics. |
| **Chi-square test**<br>Test of significance of relationship between two categorical variables | CHISQ.TEST function | To make inferences back to the population of interest. Data should be independent categorical data collected by means of simple random sampling, and sample sizes should be sufficiently large. |
| Scatter plot<br>Simple xy chart showing the relationship between two numerical variables | Scatter chart tool | To explore the relationship between two numerical variables |
| **Pearson's correlation coefficient** (r)<br>Quantitative estimate of the type of correlation between two numerical variables. | CORREL function<br>PEARSON function | To explore the strength and the direction of the relationship between two numerical variables |
| **Coefficient of determination** ($r^2$)<br>Quantitative estimate of the strength of the relationship between two numerical variables. | Trendline option in scatter chart tool | To explore the degree to which the variation in variable x is due to the linear relationship with variable y<br>To supplement understanding of the strength of the linear relationship. |

## COMPARISONS

**Comparison** in the context of this guide refers to the act of comparing an indicator or series of indicators between two or more entities (people, households, population groups, animals, institutions, etc.) and/or dimensions (geographic location, population group, etc.) in order to learn how they differ according to that entity or dimension. Some examples of the comparisons undertaken in humanitarian work are listed below:

- prevalence od disease in two different ethnic groups
- household dietary diversity score of residents and displaced people
- expenditure on food as percentage of overall expenditure, in urban and rural areas
- level of income before and after a shock

Comparisons can be between two or more entities and/or dimensions. For example, the prevalence of disease in two different ethnic groups can also be compared through space (if the groups are living in more than one place). Furthermore, trends can be explored by comparing over time (quarterly, annually, etc.).

**DISAGGREGATION**

Disaggregation in data collection and analysis involves breaking up a data set into two or more different components. For example, when reporting statistics on children who were treated at a health centre, it may be necessary to report on girls and boys separately. In this case, the data are disaggregated by gender. Disaggregation has to be considered not only in connection with data-collection tools (i.e. in this example, the data-collection forms would need to record the child's gender), but also in connection with the required end analysis: Will groups be analysed individually and/or comparisons made between them?

Comparison by definition involves two separate data sets (data from 1999 and 2000, data from one source in northern regions and another in southern regions, etc.) or sub-sets (stratified sample, paired group, etc.). Where primary data are being collected for the first time, special care must be taken before the process gets under way to ensure that it is done in a way that enables comparisons to be made. For sound comparisons to be made when secondary or historic data are used, data attributes must be checked before the data are used.

| ASK YOURSELF… | EXAMPLES – LET'S COMPARE | EXAMPLES – LET'S THINK ABOUT IT |
|---|---|---|
| Do the sources of data use the **same definitions of key terms?** | ▪ Adults are defined as anyone 18 and over in both data sets<br>▪ The composition of the minimum expenditure basket is agreed upon by the international community, and the prices are local market prices. | ▪ The data set from organization A considers anyone 18 and over as an adult, but organization B fixes adulthood at 16 and over<br>▪ The composition of the minimum expenditure basket changed in 2009, to reflect the international community's idea of a more appropriate definition. |
| **Were the methods used to collect** each indicator the same? | ▪ Household dietary diversity score based on 24-hour recall was the indicator chosen to measure food consumption, and teams from all locations used the same methods to collect data and measure indicators.<br>▪ Household interviews were used to collect data throughout the project cycle. | ▪ One data set includes remittances and gifts in the overall household income, but the other data set excludes it.<br>▪ The initial emergency assessment before the project/programme collected data on the community from key informants (the unit of observation is the community), but during monitoring, data were collected from a sample of households (the unit of observation is the household). |
| If samples were used, is **each entity or dimension sufficiently represented in the sample and are the populations of interest and the sampling units the same?** | ▪ Stratified random sampling was employed with two main strata (IDPs and residents) to ensure adequate representation of IDPs and residents in the sample.<br>▪ The same sample size and method were used in all three rounds of the monitoring exercise. | ▪ The unit of analysis was the entire region, not a sub-region, and the sample design was done with that in mind.<br>▪ The initial emergency assessment used a non-representative sample of 100 households that contained both those who would and those who would not eventually receive assistance; but the monitoring exercise included only beneficiaries in a probabilistic simple random sample. |

Entities or dimensions can be compared, superficially, by merely noting their differences. This can be done by just looking at the data and asking: Which mean is higher? Which mean is lower? Are they the same? For a deeper analysis, you may wish to describe the differences through percentages or test the differences for statistical significance. A few techniques are listed below. They are commonly used to compare the descriptive statistics associated with different entities or dimensions. A few tests for confirming statistical validity are also described below.

## ABSOLUTE DIFFERENCE

Comparing descriptive statistics associated with two different entities or dimensions is often the first exploratory step in comparison data analysis. It entails comparing means, proportions, frequencies, etc. A simple comparison of descriptive statistics does not establish statistical significance. The analysis should take into account the representation in each group (sample size, response rate, geographic coverage, etc.) during the comparison, to enable proper understanding of the scale; if possible, statistical tests can be run to confirm or reject the validity of the results.

**EXAMPLE**

A monitoring exercise in the Central African Republic, undertaken after seed was distributed to resident and IDP households in the south-eastern section of the country, collected data on the amount of land cultivated and the distance to the fields: here, land under cultivation was a proxy measure for food production, and distance to the fields an indicator for use in land-access analysis. The sample was stratified by IDP and resident households so that comparisons could be made between the two groups. The table below shows the number of households in each group and the minimum, average and maximum land cultivated in hectares.

Table 1 was created using an Excel Pivot Table; the descriptive statistics that were to be displayed were defined in the Value Field Settings Options; and n was calculated using the count function in the Value Field Setting options, which counts the number of records in each group. The absolute difference in average hectares cultivated by IDPs and Residents is 0.17 (or 0.40-0.23).

**Table 1 - Amount of land cultivated in hectares in 2010-2011**

| | n | Min | Average | Max |
|---|---|---|---|---|
| IDPs | 274 | 0.00 | 0.23 | 2.00 |
| Residents | 248 | 0.00 | 0.40 | 3.15 |

## PERCENT DIFFERENCE

The **percent difference** is a simple mathematical formula to describe the difference as a percentage.

$$\% \text{ difference} = \frac{|(x - y)|}{(x + y) / 2} \times 100$$

## RELATIVE PERCENT DIFFERENCE

The **relative difference** is a term used to compare two quantities or numbers while taking into account the sizes of the numbers or quantities being compared. It can be calculated with a simple mathematical formula. The fraction can be calculated using the following formula, and the percentage by multiplying the result by 100.

$$\text{relative } \% \text{ difference} = \frac{(x - y)}{y} \times 100$$

where:

y = value of variable for reference entity/dimension

x = value of variable for relative entity/dimension

## RATIO

A **simple ratio** can also be an effective way of expressing comparisons as it provides a number that an audience can grasp without difficulty. Ratios are calculated using simple division, and can be reported as numbers (formula below) or percentages (formula below multiplied by 100).

$$\text{Ratio} = \frac{y}{x}$$

where:

y = value of variable for reference entity/dimension
x = value of variable for relative entity/dimension

**EXAMPLE**

On average, during the 2010-2011 agricultural season, the figures for land cultivated per household were 0.23 hectares for IDPs and 0.40 hectares for residents.

If the statistics are calculated with the residents as the reference entity, the following could be reported:

- Residents cultivated on average 0.17 hectares more land/household than IDPs (difference).
- Residents cultivated on average 73.9% more than IDPs (relative percent difference).
- Residents cultivated on average 1.7 times more than IDPs (simple ratio expressed as a fraction).

If the statistics are calculated with the IDPs as the reference entity, the following statistics could be reported:

- IDPs cultivated on average 0.17 hectares less land/household than residents (difference).
- IDPs cultivated on average 42.5% less than residents (relative percent difference).
- IDPs cultivated on average 0.57 as much as residents (simple ratio expressed as a fraction).

- The percent difference in the land cultivated by the residents and IDPs is 53.9%.

**COMPARING WORDS**

Selecting the wrong words when reporting comparisons can easily cause confusion and misinterpretation of the data and the analysis. Some general guidelines are listed below.[94]

| Statistic | Formula | Terminology | Example |
|---|---|---|---|
| Absolute difference | x - y | more than<br>less than<br>from __ to __ | Four is **more than** three<br>Three is **less than** four<br>The value went **from** three **to** four |
| % difference | \|(x – y)\| / ((x+y)/2)*100 | % difference is | The **percent difference between** three and four is 28.6%. |
| Relative % difference | (x-y) / y | times more than<br>times less than<br>% more than<br>% less than | Four is .33 **times more than** three<br>Three is .25 **times less than** four<br>Four is 33% **more than** three<br>Three is 25% **less than** four |
| Ratio | x / y | times as much as<br>% as much as<br>% of | Four is 1.3 **times as much as** three<br>Four is 130% **as much as** three<br>Three is 75% **of** four |

94 Schield, 1999.

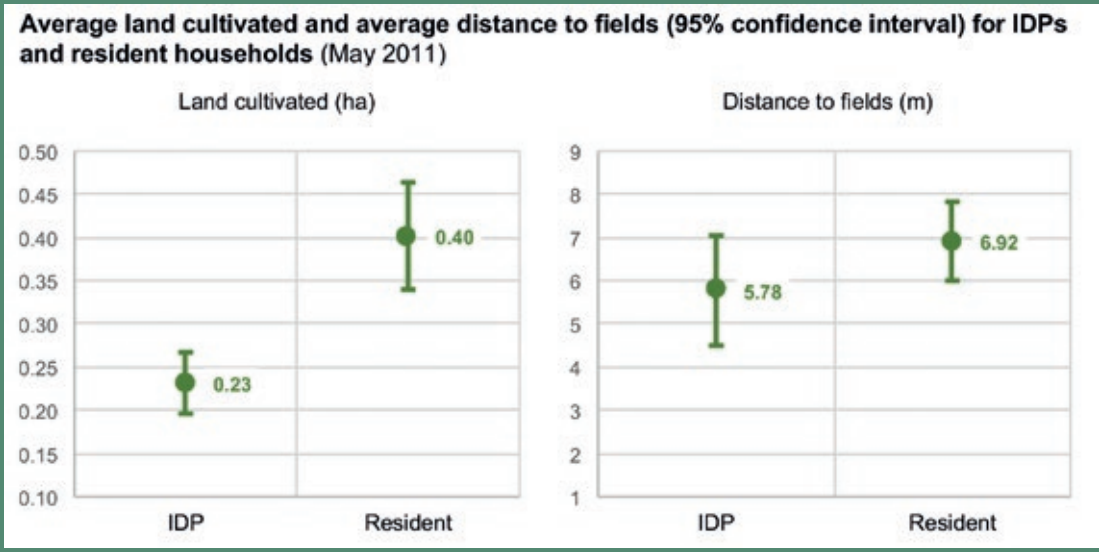
## STATISTICAL SIGNIFICANCE OF COMPARISONS

### VISUALIZING THE DIFFERENCE

If a statistically relevant sample is collected for each group, the confidence intervals can be plotted together with the mean to visualize if the differences are statistically significant. If the minimum and maximum range of values do not overlap, the difference between the means is statistically significant at the level of confidence identified when calculating the statistics. If they overlap, the difference is not statistically significant.

**EXAMPLE**

The graphic below shows the mean values, and 95% confidence intervals on the mean, for data on land cultivated and distance to fields captured during the post-distribution monitoring exercise in the Central African Republic. There is no overlap between the mean values on land cultivated, which means that the difference in the hectares cultivated is statistically significant; but the confidence intervals on the mean values of distance to fields do overlap, which means that the difference is not statistically significant.




The confidence interval ranges can be added to an Excel chart in the Chart Tools. Under Chart Layout select Add Chart Element > Error Bars > More Error bars. Under Error amount select Custom, and define the range as the confidence interval for the group.

### T-TEST: COMPARING THE MEANS OF TWO GROUPS

The t-test is another way to check if the means of two groups are statistically different from one another. T-tests can be performed on one sample or on two samples, using one of the following tests in the box below.

| | |
|---|---|
| **TWO INDEPENDENT SAMPLES** | **T-test on two samples assuming equal variances** – The variance between the two samples is known to be the same. This test is also known as the 'Student's t-test'. |
| | **T-test on two samples assuming unequal variances** – The variance between the two samples is either unknown or known to be different. |
| **TWO DEPENDENT SAMPLES** | **Paired sample t-test** – For paired samples, such as 'repeat exercises' involving the same population. This test is also known as the t-test for 'dependent means', 'matched pairs' or 'matched samples'. |

Furthermore, a t-test may be either 'one-sided' or 'two-sided'.[95] A **two-sided** test allots half of the alpha (statistical significance) to the difference in one direction and the other half to the relationship in the other direction. For example, the test will tell either that there is no difference between the means of the two samples or that there is a difference and that one sample is either significantly greater *or* significantly smallerß than the other. A **one-sided** test allots all of the alpha (statistical significance) to the difference in one direction. It is stronger for detecting an effect in one direction by not testing the effect in the other direction. For example, the test will tell you either that there is no difference between the means of the two samples or that there is a difference in a specified direction (assumed and identified before data collection).

A one-sided test should be used only when the direction of the relationship is well defined, and when there is no question about not identifying the precise direction of the relationship.

To perform a t-test for two samples in Excel, the following steps can be followed:

1. Select the type of test to use.
2. If you have chosen to use one of the tests for two independent samples, but do not know whether or not variances are equal, perform an F-test of the equality of the variances. This can be done using F-Test Two Sample for Variances in the Excel Analysis ToolPak. If the p-value of the F-test is less than 0.05, the variances cannot be assumed to be equal, and the t-test assuming unequal variances should be used.
3. The t-test can be performed in MS Excel using either the T.TEST function or the Analysis Toolpak, where one of the three tests mentioned above must be selected. The T.TEST function requires you to select either a one-sided or a two-sided test. The tools in the Analysis Toolpak test and present results for both.

**EXAMPLE**

Following on for the example above, concerning beneficiaries of a seed distribution in the Central African Republic – on the amount of land cultivated and the distance to fields – the means are tested for statistical difference. First, the F-test is used to determine if the variances are the same. This is done using F-Test Two Sample for Variances in the Excel Analysis ToolPak. The p-value for the F-test is less than 0.05, which means that the variances cannot be assumed to be equal. As a result, the t-Test: Two-Sample Assuming Unequal Variances tool in the Excel Analysis ToolPak is used to test if the means are statistically different.

The following table shows the t statistic, t critical values and the p-value. If the **t stat < –t critical** or the **t stat > t critical** and the **p-value is less than 0.05**, the null hypothesis can be rejected and the difference between the means is significant. The results show that the mean area of land cultivated by IDPs is significantly less than that cultivated by residents ($p<0.05$); however, the distance to the fields is not statistically different. These results correspond to the confidence interval plots.

| Variable | t Stat | t critical (two-sided) | p-value (two-sided) |
|---|---|---|---|
| Land cultivated (ha) | -4.6952 | 1.9660 | 0.0000 |
| Distance to fields (m) | -1.4295 | 1.9670 | 0.1538 |

**ASSUMPTIONS FOR T-TEST**

- Numerical variables (discrete intervals and continuous variables)
- Data are normally distributed
- Sample sizes are sufficiently large and data collected randomly
- Sample sizes between groups do not have to be equal

95 Also commonly referred to as 'one-tailed' or 'two-tailed'.

## ONE-WAY ANOVA: COMPARING MEANS OF THREE OR MORE GROUPS

The analysis of variance test (ANOVA) examines the differences in the mean and in the variance among three or more groups. It essentially generalizes the t-test to more than two groups. The one-way ANOVA test shows only whether two or more of the groups have statistically different means. The results are interpreted by comparing the f statistic with the f critical value. If the f statistic is greater than the f critical value, and the p-value of the ANOVA test less than 0.05, and if the null hypothesis is rejected (meaning there is a statically significant difference), a *post hoc* test can be used to determine which combinations are different (i.e. which groups are different for that particular variable).

### EXAMPLE

The following table show the results of analysing average monthly expenditure against cost-of-living index (proxy indicator for purchasing power) between resident, returnee and IDP populations in northern Mali. The statistics were run using the ANOVA: Single Factor tool in the Excel Analysis ToolPak. The first table shows the descriptive statistics for each group; the results of the ANOVA test are shown in the second table. It can be seen in the first table that the average expenditure against cost of living residents is higher for residents (i.e. assumed greater purchasing power by the analyst) than for IDPs, and lowest for returnees. The second table shows that the F value is higher than the F critical value, with a p-value less than 0.05; this means that the difference between the means is significant and not due to chance alone.

Anova: Single Factor

SUMMARY

| *Groups* | *Count* | *Sum* | *Average* | *Variance* |
|---|---|---|---|---|
| IDPs | 29 | -882,836 | -30442.61 | 8,716,739,529 |
| Residents | 812 | 3,041,870 | 3746.15 | 7,965,932,942 |
| Returnees | 260 | -9,447,004 | -36334.63 | 18,473,576,334 |

ANOVA

| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
|---|---|---|---|---|---|---|
| Between Groups | 3.33282E+11 | 2 | 1.66641E+11 | 15.9257 | 0.0000 | 3.0039 |
| Within Groups | 1.14891E+13 | 1098 | 10463658099 | | | |
| Total | 1.18224E+13 | 1100 | | | | |

The t-test shows that the real differences in monthly expenditure against the cost-of-living index are between residents and returnees, where the t stat is greater than the t critical and the p-value is significant ($p<0.05$). There are no real differences between IDPs and the others. The same results are obtained when an ad hoc test in SPSS is used.

| Groups | t Stat | t critical (two-sided) | p-value (two-sided) |
|---|---|---|---|
| IDPs and Residents | -1.9406 | 2.0423 | 0.0618 |
| IDPs and Returnees | 0.3056 | 2.0167 | 0.7614 |
| Returnees and Residents | 4.4572 | 1.9671 | 0.0000 |

### ASSUMPTIONS FOR ONE-WAY ANOVA[96]

- Dependent variable is an interval or continuous ratio
- Independent variable is two (typically three) or more groups or categorical variables
- Observations are independent (i.e. different participants in each group, with no participant being in more than one group)
- No significant outliers
- Dependent variable is normally distributed for each group/category
- Each group has homogeneous variances

96 Laerd Statistics, online resource, accessed in April 2015.

| COMPARISON ANALYSIS | HOW TO CALCULATE IT IN EXCEL | WHEN TO USE IT |
|---|---|---|
| **Descriptive statistics** Comparison of means, proportions, ratios, etc. between two or more groups. | IF function in combination with relevant mathematical function Pivot table | To conduct an exploratory analysis of the differences between groups supported by triangulation or statistical test |
| **Percentage difference** Quantifies the difference in percentages of any descriptive statistic between two groups. | Mathematical formulas | To quantify statistical differences between two groups, in order to report statistics in another way |
| **Confidence interval plots** Compares the significance of the differences in the means of two or more groups | CONFIDENCE.NORM function or Descriptive Statistics tool in Analysis Toolpak and Excel chart and error bars | To explore the real differences in the means of two or more groups of statistically relevant data |
| **T-test equal variances** Compares the significance of the differences in the means of two groups with equal variances | T.TEST function or t-test: Two-Sample Assuming equal Variances tool in Analysis Toolpak | To explore the real differences in the means of statistically relevant data |
| **T-test unequal variances** Compares the significance of the differences in the means of two groups with unequal variances. | T.TEST function or t-test: Two-Sample Assuming unequal Variances tool in Analysis Toolpak | To explore the real differences in the means of statistically relevant data |
| **T-test paired sample** Compares the significance of the differences in the means of two paired samples. | T.TEST function or t-test: Paired Two Sample for Means tool in Analysis Toolpak | To explore the real differences in the means of statistically relevant data |
| **One-way ANOVA** Compares the significance of the differences in the means of three or more groups. | Anova: Single Factor tool in Analysis ToolPak | To explore the real differences in the means of two or more groups of statistically relevant data |

## TRENDS

**Trend analysis** studies past and current data on a given indicator to try and spot trends, including anomalies (irregular events or tendencies) and patterns (repeated events or tendencies). The aim is to acquire a better understanding of the relationship between two or more entities or dimensions. Some of the subjects for which trend analysis can be used in humanitarian work are listed below:

- evolution of the price of local wheat
- number of admissions per week to a hospital or clinic for treatment of malnutrition
- number of households affected by flooding from 1990 to 2010
- attacks against civilians by week over a 12-month period
- level of income before and after a project/programme.

Trend analysis can involve something as simple as examining one variable for one entity or dimension (e.g. price of local wheat in one market), or as complex as studying many different variables over many entities or dimensions (e.g. price of local and imported wheat in all markets in the country). Like comparisons, trend analysis, by definition, involves two separate data sets (data from date 1, date 2, etc.). Where primary data are being collected for the first time, special care must be taken before the process gets under way to ensure that it is done in a way that enables comparisons to be made. For sound comparisons to be

made when secondary or historic data are used, data attributes must be checked before the data are used. In addition to the scale of analysis and the methods used, trend analysis must also pay close attention to the timing and frequency of data collection in order to capture patterns and trends.

**EXAMPLE I**

To monitor the market prices of commodities, you may need to collect data only on a monthly basis, because you are trying to capture trends that are influenced mainly by seasonal changes or slow-onset shocks, and it may take some time for these to have an impact on the market. Data collection should take place at the same time of day, and on the same day of the week, to ensure the uniformity of the data for each month.

**EXAMPLE II**

Monitoring an outbreak of cholera or some other disease may require data analysis on a daily or weekly – not a monthly – basis, in order to capture trends and important developments, as the situation may be very fast-moving.

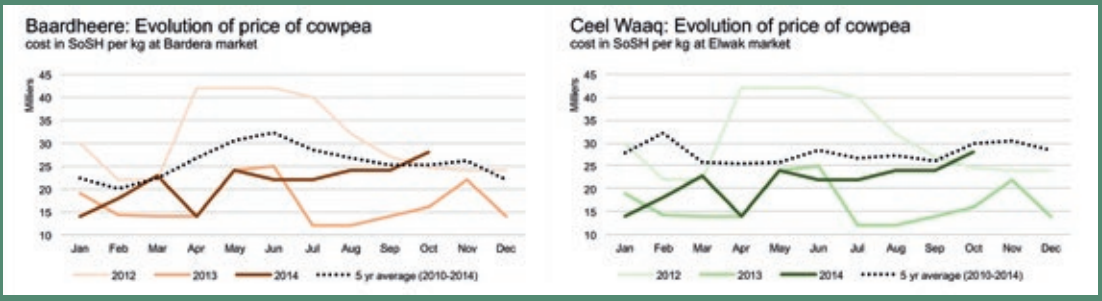
Like comparisons, trend analysis may begin by merely noting differences over time. This can be done by just looking at the data and asking: Was there an increase? Was there a decrease? Did it stay the same? For a deeper analysis, you may wish to describe the difference through percentages or test the difference for statistical significance. A few techniques that can be used in trend analysis are described below.

## TRENDLINE

Charts with trendlines are a powerful tool for visualizing trends over time. They can provide a simple linear illustration or a comparison of trends for different clusters (years, locations, etc.).

**EXAMPLE**

The two charts below show the evolution of the price of cowpea in two markets in Somalia; the data were collected and shared by the Food Security and Nutrition Analysis Unit in Somalia (FSNAU). The graphics were created using the line chart tool in Excel; the data were the raw data taken directly form the data source. The analysis was performed through simple visualization of the data. *Data source: FSNAU Somalia Integrated Database System.*



## RELATIVE PERCENT CHANGE

**Relative percent change** is a statistic, expressed as a percent, that quantifies the magnitude and direction of change of a variable from one point in time to another, where positive is associated with an increase and negative with a decrease. The formula is the same as relative difference (expressed as a percentage); however instead of comparing two different observations, two different periods of time are compared. The following is an adaptation of the formula:

$$\text{Percent change} = \frac{x_2 - x_1}{x_1} \times 100$$

where:

$x_1$ = value of variable at first point in time

$x_2$ = value of variable at second point in time

**EXAMPLE**

The price of cowpea in Bardera was 28,000 Somali shillings (SoSH) in October 2014, a 16.7% increase (from 24,000 SoSH) since September 2014 and 10.2% higher than the five-year average for October 2015.

## RATES

In the context of trend analysis, a **rate** is a type of ratio that compares two measures. These may be: two different variables or constants, such as distance and time (e.g. speed); two commodities or currencies (e.g. exchange rate); health indicators and time (mortality rate, incidence rate, etc.); social indicators and population (literacy rate); and so on. Simple rates can be calculated using simple division, and when they are reported, the units of measure of both variables (metres, days, people, households, etc.) should be used.

$$\text{Rate} = \frac{x}{y}$$

where:

y = value of variable for reference entity/dimension

x = value of variable for relative entity/dimension

**EXAMPLE I**

A rate that compares cases of Ebola (y) to deaths from Ebola (x).

As of 14 September 2014, the Ebola case fatality rate of patients with definitive outcomes in Guinea was 70.8% (95% CI, 69 to 73), which means that 7.1 of 10 cases of Ebola in Guinea with definitive outcomes ended with the patient's death.[97]

**EXAMPLE II**

This rate compares the value of 1 SoSH (x) to 1 US$ (y).

On 25 March 2015, the exchange rate was 704 SoSH to 1 US$.

97 WHO Ebola Response Team, October 2014.

**Change rates** look at the average amount of change per given time interval (daily, weekly, monthly, yearly, etc.).

$$\text{Change rate} = \frac{(x_2 - x_1) / t_{x2\text{-}x1}}{x_1}$$

where:

$x_1$ = value of variable at first point in time
$x_2$ = value of variable at second point in time
$tx_1$ = time one
$tx_2$ = time two

**EXAMPLE**

The population growth rate in Somalia in 2012 was 2.9% (from 9.91 million in 2011 to 10.19 million in 2012). The rate in 2011 was 2.8%. The highest growth rate, 10.98%, was in 1977.[98]
The population growth rate describes the change in total population from the previous year (t=1 year).

**USING THE RIGHT WORDS**

As with comparisons, using the wrong words when reporting changes can easily cause confusion, and misinterpretation of the data and analysis. Some general guidelines are listed below.[99]

| Statistic | Formula | Terminology | Example |
|---|---|---|---|
| Absolute change | $x_2$-$x_1$ | From __ to __ | The value went from three to four |
| Relative percent change | $(x_2$-$x_1) / x_1$ | Times more than<br>Times less than<br>Percent more than<br>Percent less than | The value is .25 times more than before<br>The value is .33 times less than before<br>The value is 25% more than before<br>The value is 33.3% less than before |
| Rate | x / y | ___ per/every ___<br>___ in ___<br>Was/is ___% | The rate is three incidents per day.<br>The weekly incident rate is 43%. |
| Change rate or growth rate | $((x_2$-$x_1) / t_{x1\text{-}x2}) / x_1$ | Rate was/is<br>At a rate of | The growth rate was 33.3%<br>The value grew at a rate of 33.3% |

## FORECASTING

**Forecasting,** in quantitative analysis, entails extrapolating from existing data to predict the future values of a variable. Forecasting may be confined to a single place and time or it may be a projection of future growth; it usually involves a series of assumptions (hypotheses) and predictions.

98 World Bank Data Bank, accessed in April 2015.
99 Schield, 1999.

**EXAMPLE I**

WHO projections of the spread of Ebola in three Ebola-affected countries

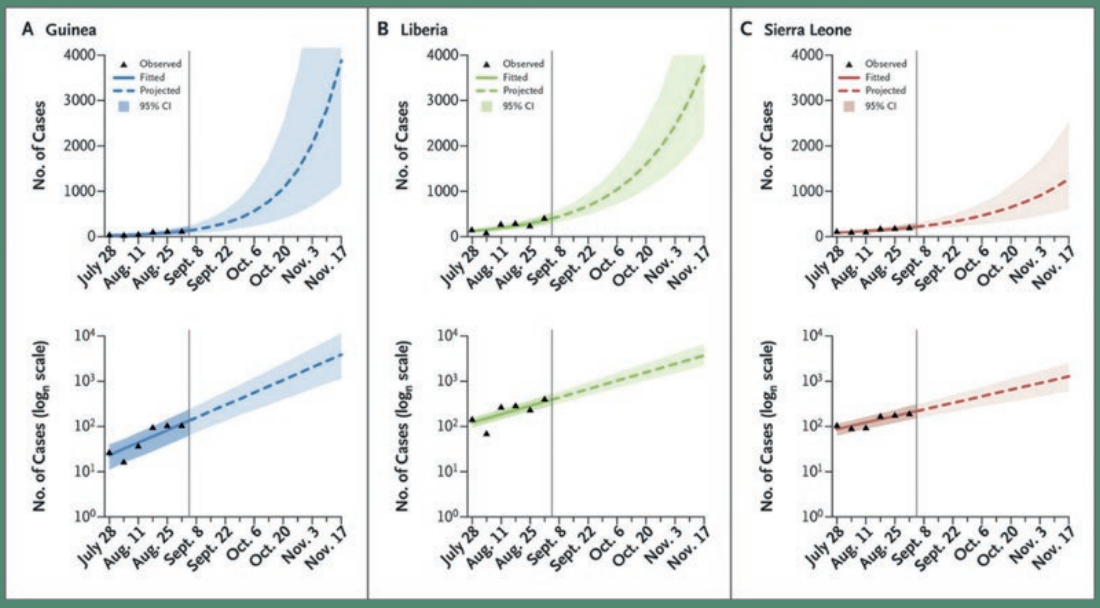



**Figure 37** - Observed and Projected Case Incidence. Observed and projected weekly case incidence in Guinea (Panel A), Liberia (Panel B), and Sierra Leone (Panel C) are shown on linear (upper panels) and logarithmic (lower panels) scales. Projections assume no changes in control efforts.
Source: WHO Ebola Response Team, October 2014.

**EXAMPLE II**

OXFAM GB forecast model of the number of people affected by natural disasters

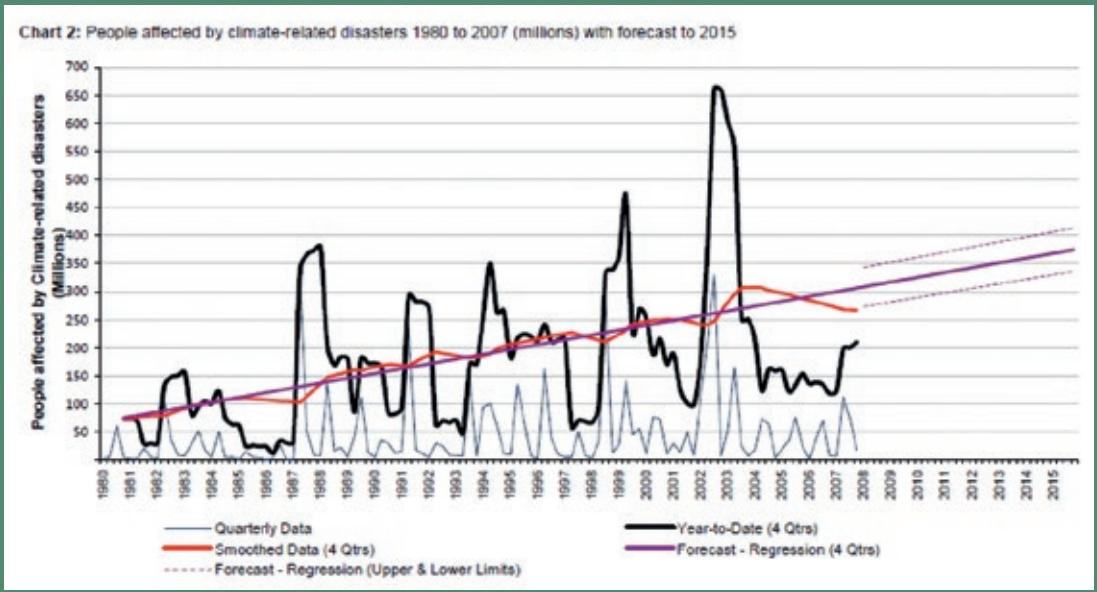



**Figure 38** - Based on historical data from the Emergency Events Database of the Centre for Research on the Epidemiology of Disasters. This forecast model of people affected by natural disasters used a combination of two things: 'double exponential smoothing' to add weight to more recent events over past, based on the assumption that there was an underlying trend in the data (e.g. growing number of people affected by natural disasters); and linear regression with 95% confidence intervals for future predictions.
Source: OXFAM GB, April 2009.

Quantitative forecasting involves some form of probability statistics and a regression model such as a linear regression or logistic regression. In ICRC operations, forecasting is done mainly through a combination of qualitative and quantitative indicators; the analysis is qualitative and comparative, as most forecasting is concerned with such matters as population movement, the spread of conflict, and the risk of violence in acute emergencies. This could be supplemented with quantitative statistical forecasting if data of sufficient quality are available. If assistance is required, a statistician should be sought out.

| TREND ANALYSIS | HOW TO CALCULATE IT IN EXCEL | WHEN TO USE IT |
|---|---|---|
| **Trendline**<br>Visualizes data through time | Line chart | To carry out a quantitative examination of historical, seasonal, annual, etc. trends in regularly collected data |
| **Relative percent change**<br>Quantifies the change in a value | Mathematical formulas | To quantify the change from one point in time to another |
| **Rates**<br>Compares the values of two or more variables | Mathematical formulas | To acquire an up-to-date snapshot of how the values of two or more quantitative variables perform against one another |
| **Change rate**<br>Quantifies the rate of change | Mathematical formulas | To quantify the scale of change from one point in time to another |

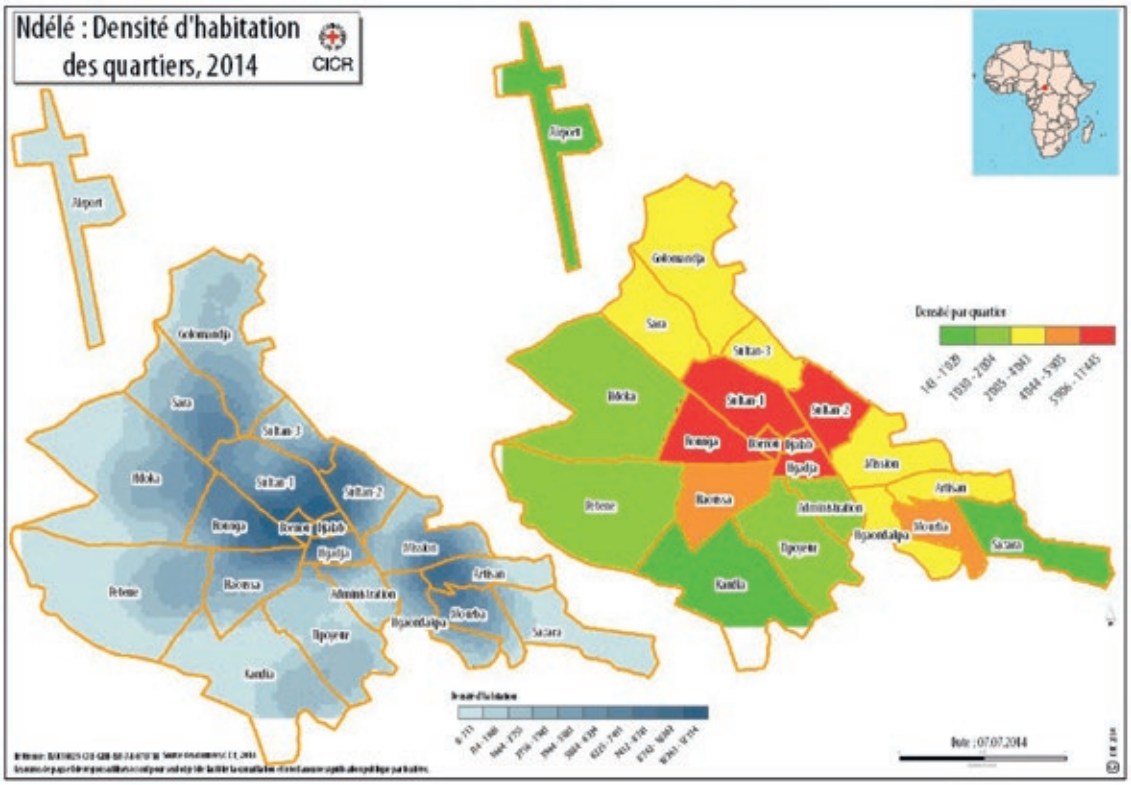
## ADDING DEPTH TO THE ANALYSIS

Here, **depth** refers to the profundity of the data, adding another dimension to the analysis. While not always required in data analysis, performing additional analysis can enrich understanding and the interpretation of the data. Adding depth to figures involves comparing or contextualizing the data. Some of the most commonly used statistics for adding depth to quantitative data analysis involve developing bivariate or multivariate ratios and proportions, comparative proportions and comparative analogies.

### BIVARIATE PROPORTIONS AND RATIOS

Bivariate and multivariate proportions and ratios use two or more variables to create a derived variable that combines the information from the these variables to provide new information. Two examples are provided in the sections on Trends and Comparisons. Another common example in humanitarian analysis is population density, the population is divided by the surface area to calculate the density of a population.

**EXAMPLE**
Population density in the town of Ndélé, in the north-eastern section of the Central African Republic. Map created by WatHab GIS using population figures against geographic area.




## CONTEXTUALIZATION

Comparing data from one event or time period to another can serve as a simple and powerful tool to contextualize data and information quantitatively.

**EXAMPLE**

"Over a million people have arrived in Lattakia and Tartous since the beginning of the conflict, swelling the local population by 50%." [100]

The statistic above – an example of relative percent change – compares the new arrivals with baseline population figures, thereby contextualizing the data. The first part of the statement reports the raw figures (over a million people) on the number of arrivals. This statement alone is an aid to understanding and planning. When it is compared to the baseline data, depth is added, through scale, to the statistic and the statement becomes a much more powerful guide to understanding not only the arrivals but those who are welcoming them.

## COMPARATIVE ANALOGIES

In a **comparative analogy,** a unit of measurement is compared with something that the audience can more easily understand or digest. Comparative analogies are particularly useful for expressing technical measurements. They can be used to contextualize area, volume, density, population size and more.

**EXAMPLE**

Comparative analogies developed from actual reported statistics:

- The seven-week conflict in Gaza in early 2014 caused substantial damage to an estimated 17,000 hectares of cropland, an area larger than the entire territory of Liechtenstein.[101]
- Dadaab refugee camp in Kenya is estimated to house more than 335,000 Somali refugees, 1.7 times the population of the city of Geneva.[102]

100 ICRC Syria, July 2014.

101 FAO, August 2014.

102 UNHCR online resource, accessed in April 2015; Rép. et canton de Genève, online resource, accessed in April 2015.

## OUTLIERS

An **outlier** is a piece of data or information that is at an abnormal distance from all other measures. Quantitatively, this can be a number that is much higher or much lower than others. Qualitatively, it can be a response that is very different from others or unexpected in some way. Outliers are essentially anomalies in a specific data set.

The first step in dealing with an outlier is to determine if it is an error or a true value. The second step is to determine how to deal with it. In any case, an outlier cannot be completed ignored unless it is an error, as its value is based on something, and removing it can cause false results or misrepresent the facts. However, treatment of the outlier can be adapted to the requirements of the data analysis. The table below provides general guidance for dealing with outliers during basic descriptive statistical analysis. More advanced statistical models may deal with outliers in greater detail by using transformations.

| CASE | EXAMPLE | WHAT TO DO |
|---|---|---|
| Outlier is an error | Record shows 22 adults members of a household in an area where average household size is 6. | ▪ Try to find the correct figure for that household, or remove it. |
| Outlier(s) affect(s) the results, but general trends and relationships are preserved | One household reports a monthly income of 80,000 while the maximum for all other households is 40,000, with an average of about 25,000. | ▪ Statistics can be used that accurately portray the relationships (median, max and min vs mean) and/or results may be presented **with and without** the outlier, together with an explanation.<br>▪ Advanced statisticians can use transformations to lessen the effect of the outlier. |
| Outlier(s) affect(s) the results in such a way that trends or key relationships are lost | One household reports a monthly income of 1,500,000 while the average of all other households is 25,000. | ▪ Results may be presented **without the outlier**, together with an explanation.<br>▪ A general rule of thumb when removing outliers is to delete those up to 3 or 4 standard deviations from the mean.<br>▪ Advanced statisticians can use transformations to lessen the effect of the outlier. |

## MISSING VALUES AND NON-RESPONSE

Sometimes missing values are obvious and sometimes they are not; their absence can be of due to a respondent's inability or unwillingness to respond, the way in which data were collected/recorded and/or the way in which data were entered into the data-entry sheet. Some common reasons for non-response are listed below:

- subject's refusal to respond or inability to do so, owing to lack of knowledge/information
- non-applicability
- errors/missing data.

Missing data may be said to be "missing at random" if there is no logical explanation for their absence; or "not missing at random" if there is a rational explanation (data collector A never collected data on variable X, livestock herding/farming households on the east side of the village did not respond when asked how many heads of cattle they had, etc.). The first step in dealing with missing values is to determine how they will be treated and how important the information contained in the missing values is.

| CASE | POINTS TO CONSIDER BEFORE ANALYSIS | THINGS TO DO IN ANALYSIS |
|---|---|---|
| Non-response | ▪ Can the variable still be analysed and, if yes, will the data be inherently biased? Who responded? Who didn't respond?<br>▪ How can we optimize response rate in the future? Do we need to change the way we collect this variable or measure this element? | ▪ Talk to the data collectors about response-rate trends.<br>▪ The analysis should report response rate or absolute number of responses.<br>▪ Recommendations for future studies should be made on response rate in sample estimate, indicators and data-collection methods. |
| Non-applicability | ▪ Can we distinguish between those that are non-applicable and those that are non-response or missing values, in order to check data quality?<br>▪ Did some who were not applicable respond? For example, did households without children respond to questions on children missing school? | ▪ Use filter to verify that the data were not applicable, and double check any other missing values before moving forward with the analysis.<br>▪ The analysis will be performed only on applicable data (i.e. n = applicable units). |
| Missing data | ▪ Why are the data missing? Were the questions irrelevant? Did the data collectors skip the questions? Were the respondents unwilling or unable to respond?<br>▪ Can the variable still be analyzed and, if yes, will the data be inherently biased? Whose data are missing? Which respondent(s)? Which data collector(s)?<br>▪ If it was indeed errors or data collectors skipping questions, how can we work in the future to have more complete dataset? | ▪ Use filters and talk to the data collectors to understand why data are missing.<br>▪ The analysis should report response rate and data errors<br>▪ Recommendations should be made for future studies, focusing on the reasons for the data's absence.<br>▪ More advanced statistical models, using imputation, may be employed to deal with missing data. |

Sometimes data-entry personnel use a zero value ("0") in place of missing or blank data, and some leave cells blank if the value is in fact zero. This needs to be considered in data collection, treatment and finally, analysis, to ensure that statistics are calculated properly. Every zero value will raise or lower the mean.

### CALCULATING NON-RESPONSE RATE

Non-response can be calculated in the same manner as frequency: the number of times non-response is reported is counted and this figure then divided by the total number of possible responses; that will yield a 'non-response rate'.

# REPORTING STATISTICS

## ROUNDING OFF NUMBERS

The number of digits will depend on how precise the data need to be, but here are some rules of thumb to follow:

- Use the maximum number of digits during intermediate calculations to ensure that small differences in the data are preserved.
- Final calculations should use two digits more than the original data when small differences need to be preserved. For example: If the average number of heads of livestock owned by poor households is collected as a whole number, values are low and variation minimal (7, 8, 10, 7, 6, 5, 8, 7, etc.), the average should be reported as a decimal with two places (e.g. 6.45) to account for small differences in the data set. In some cases, however, small differences are inconsequential. For example: If you take the average monthly income of three households reporting 750, 825, and 700, reporting the result with two decimal with place (758.33) adds nothing of value to the rounded-off average of 758.
- Round numbers off, upwards if the last digit is more than five (e.g. 12.12507 is rounded off to 12.13), and downwards if the last digit is equal to 5 or less (e.g. 12.12500 is rounded off to 12.12).
- When writing up narratives and quoting figures (except in tables), eliminate decimals as much as possible: the focus is on the magnitude of a trend or the differences it reflects, not on precise figures. Think about the people who will be reading the narrative and what they need to understand.

# REFERENCES

ACTED, *Food Security Situation and Livelihood Intervention Opportunities for Syrian Refugees and Host Communities in North Jordan*, August 2013. Available at: http://www.acted.org/en/food-security-situation-and-livelihood-intervention-opportunities-syrians-refugees-and-host-communit.

CRED, "Emergency Management Events Database EM-DAT", online resource. Available at: http://www.emdat.be/.

FAO, "Gaza: Damage to agriculture will have long-lasting effects", August 2014. Available at: http://www.fao.org/emergencies/fao-in-action/stories/stories-detail/en/c/241146/.

FAO, *Guidelines for Measuring Household and Individual Dietary diversity*, reprinted in 2013. Available at: http://www.fao.org/docrep/014/i1983e/i1983e00.htm.

ICRC, *"Syria: Growing needs of families displaced to coastal cities"*, 11 July 2014. Available at: https://www.icrc.org/eng/resources/documents/update/2014/07-11-syria-displaced-aleppo-tartus-latakia.htm.

ICRC, *Democratic Republic of the Congo: Initial assessment in Kahele, South Kivu*, June 2014.

ICRC, *Northern Mali: Assessment of the Economic Security Situation*, June 2014.

Kenny, David A., *Statistics for the Social and Behavioral Sciences*, Little, Brown, 1987.

Laerd Statistics, online resource. Available at: https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php.

OXFAM GB, *Forecasting the Numbers of People Affected Annually by Natural Disasters up to 2015*, April 2009. Available at: http://policy-practice.oxfam.org.uk/publications/forecasting-numbers-of-people-affected-annually-by-natural-disasters-up-to-2015-112416.

Rép. Et Canton de Genève, "Statistique Genève", online resource. Available at: http://www.ge.ch/statistique/.

Rodriguez-Llanes J.M., *et al*., "Child malnutrition and recurrent flooding in rural eastern India: A community-based survey", *BMJ Open*, Vol.1, Issue 2, 2011.

Scheuren, Fritz, *What is a Survey,* 1997. Available at: www.amstat.org/sections/srms/pamphlet.pdf.

Schield, Milo. *Common Errors in Forming Arithmetic Comparisons*, pp. 2-3, APDU *Of Significance*, 1999. Available from: www.statlit.org/pdf/1999SchieldAPDU2.pdf.

Trochim, William M.K, *The Research Methods Knowledge Base*, 3rd ed., Atomic Dog 2006. Available at: http://www.socialresearchmethods.net/kb/measlevl.php.

UNHCR, "Refugees in the Horn of Africa: Somali Displacement Crisis", online resource. Available at: http://data.unhcr.org/horn-of-africa/region.php?id=3&country=110.

World Bank World Data Bank, "World Development Indicators", Online resource. Available at: http://databank.worldbank.org/.

WFP, *Food Consumption Analysis: Calculation and Use of the Food Consumption Score in Food Security Analysis*, February 2008. Available at: http://www.wfp.org/content/technical-guidance-sheet-food-consumption-analysis-calculation-and-use-food-consumption-score-food-s.

WFP, *Comprehensive Food Security and Vulnerability AnalysisGuidelines*, 1st ed., 2009. Available at: http://documents.wfp.org/stellent/groups/public/documents/manual_guide_proced/wfp203208.pdf.

WHO Ebola Response Team, "Ebola virus disease in West Africa: The first 9 months of the epidemic and forward projections", *New England Journal of Medicine*, Vol. 371, No. 16, 16 October 2014, pp. 1481-1495. Available at: http://www.nejm.org/doi/full/10.1056/NEJMoa1411100?query=featured_home.

# CHAPTER 8
# QUALITATIVE ANALYSIS

**Qualitative analysis** is defined, for the purposes of this guide, as analysis of information in a non-quantifiable manner. Qualitative analysis uses inductive reasoning (generalizations from descriptions or observations) to uncover emerging themes, patterns and concepts that lead to insights and understanding in support of information requirements and that also support decision-making.

**EXAMPLE**
Displaced and returnee households in the three villages visited in Hauts Plateaux of Kalehe, in the Democratic Republic of the Congo, have a diet generally composed of cassava flour and potatoes; richer households (not interviewed) reportedly have a broader diet. This information is based on interviews with 151 displaced and returnee households, and on focus-group discussions with community leaders.[103]

Qualitative analysis is an iterative process that starts in the field. The data collector(s) and survey team should have a clear understanding of the objectives of the study and the questions that have to be answered; and should use appropriate techniques to generate information in a way that will enable it to be analysed. Data-collection methods can include note-taking, guides and checklists, structured or semi-structured tools, and recording devices. Plenty of space should be made available in the data- collection form for additional note-taking and for the data collector's observations, analysis and interpretation; all this is necessary for a complete qualitative analysis.

Like quantitative analysis, qualitative analysis can be descriptive and/or comparative. In qualitative analysis, however, transferability to like situations/contexts is discussed rather than inference back to a population of interest. Data collection and processing have a vitally important role in the end analysis and interpretation. Chapter 4 dealt with the subject of primary-qualitative-data collection. This chapter will consider some of the methods that can be used to process and analyse qualitative data; the treatment of the subject here is by no means exhaustive; every concept should be adapted to the information needs of the situation in question.

## EXTRACTION AND ORGANIZATION

Before qualitative data analysis can begin, data must be extracted and organized in a way that enable them to be exploited further, understood and analysed. This can be done during data collection within the methods used to gather and record data and information; or after data have been collected, during data treatment and analysis, using the methods for mining, categorizing and/or grouping or breaking up data and information.

Organization can entail first sorting through the data, highlighting or setting aside data that match information needs. Then data may be '**coded**' and data with like codes **categorized** or **clustered** systematically. Data treated in this way will seem less complex, will be understandable or analysable, and will enable patterns and relationships to be detected, concepts elaborated and theories tested.

The process of discovering relevant pieces of data and organizing them is iterative; it entails frequent refinement during data collection and throughout the analysis phase.

### EXTRACTION

The first step in organizing qualitative data is to identify what is relevant and what is not – and if possible, what is most and what is least relevant – to the subject matter. To do all this, the team must have a sound knowledge of the subject matter, the information requirements and any underlying concepts or theories. The team also has to familiarize itself with the data, and grasp all its subtleties.

103 ICRC DRC, June 2014.

The team or lead analysts should take ownership of the data, reading and rereading it, and taking notes and jotting down additional analytic memos. This process of getting to know the data, as it were, is referred to as establishing "data intimacy".[104] Strauss and Corbin refer to "theoretical sensitivity" as an important aspect of extracting qualitative data and the concepts contained in them. This is essentially an "awareness of the subtleties of meaning of data (Strauss and Corbin, p. 41)". 'Sensitivity' can be developed through literature review and professional and personal experience, and gradually throughout the analytic process.

The amount of work involved in extracting data, and quality of the results, will depend on the amount of data, the extent to which data collection was structured, the consistency of the responses and the consistency among data collectors (if more than one); another crucial factor is the analyst's 'intimacy' with the data and his or her 'sensitivity' in this regard.

Extraction may be particularly challenging when working with data from many sources, in various formats, and of different scales.

**EXAMPLE**

"…arriving humanitarian field staff interviewed consistently reported that as they rushed to reassemble data sets and coordinate the relief effort, they felt as though they were trying to drink from a fire hose of information. Yet these same respondents described not being able to collate, analyse and transform into the knowledge they needed to make decisions and brief their bosses (Harvard Humanitarian Initiative, *Disaster Relief 2.0: The Future of Information Sharing in Humanitarian Emergencies,* 2011, p.17)."

The example above, which refers to information management after the earthquake in Haiti in 2010, illustrates one of the key challenges encountered in contemporary large-scale humanitarian emergencies: data and information sources include reports not only from field workers from one unit within one organization, but from numerous other organizations and disaster-affected communities through social media and global digital volunteers. Lessons learnt emphasized the need for every data provider, processor and analyst to understand his or her role and to work together in collecting and sharing data and eventually, in turning data into information and knowledge.

Methods for data extraction may include highlighting or underlining key points in a report or data record sheet, entering structured data into a spreadsheet, text pattern-matching, data mining, and Web scraping. The main goal of any form of extraction is to extract as much relevant data and information as possible for the next steps in the analysis. Organization through coding and categorization can be a vital part of this process; it can also be done afterwards.

## DISTINCTION

Qualitative data should be collected (see Chapter 4: **Primary-data collection** ) in a way that allows the data processor and analyst to distinguish data by source and type, such as subject recall, subject quotes, community records and data-collectors' observations or analysis/interpretation. The analysis phase should carefully consider each piece of data – to separate them out for analysis and to triangulate them against each other. Chapter 4 discusses doing this at the data-collection phase, by establishing the differences between data collectors 'memos' and participants' responses or by separating one from the other.

## CODING

Coding is a technique used to attach qualitative data to particular subjects or classes – a word or short phrase – in order to reduce data to a form in which they can be analysed more easily. Coding helps capture key points that can be used as the basis for further analysis and to reduce noise in data without losing meaning (hence, data reduction). Codes can analysed quantitatively or qualitatively, and/or help to identify relationships and patterns in the data. They may be further organized into categories.

104 Saldaña, 2011.

There are many approaches to coding, and coding can be used in various ways at different stages in the analytical process. In this section, we regard codes as most closely attached to the data – nearest to the original words – which can then be further classified or 'categorized'. For our purposes, a code is different from a category in this way: coding simply transcribes data into something that is intelligible whereas categories are developed according to a given theme, pattern or concept. For example, data coded to *fewer meals*, *less expensive food*, *additional sources of income* and *debt* may eventually be placed in a category for coping strategies.[105]

**WHAT TO CODE**

Coding is most relevant for large amounts of qualitative data, either a large sample or a long interview, where the objective of the analysis is to eventually look for patterns in the data and/or frequency (e.g. quantitative analysis of qualitative data). It can focus on one or more topics, one variable or many, etc. Coding can be time-consuming and, as has already been emphasized, the analyst must feel comfortable with the data and with coding.

## CODING METHODS

Coding, like data discovery and data reduction, is an iterative process, and may be 'open' or 'selective'.

Open coding is essentially the first level of coding: initial codes are created based on a first look or a first reading of the data. These codes may be the most detailed and closest to the 'real' data themselves, and may form the basis for further exploration. The codes themselves are derived from the data and developed during analysis. Open coding is relevant in when the analyst wishes to capture all details, and is commonly used when concepts, models or theories are not yet established.

Selective coding takes the core variable of interest, and uses it as a framework for coding the data. It is more restrictive and therefore 'selective'; it may be secondary to open coding, and may be an element in refining data. The codes themselves are derived from concepts, models or theories, and may be developed before or during analysis.[106]

## CODING TYPES

The type of code used and its application (e.g. what text is given what code) may vary widely, according to the subject of interest, the context, and the perspective of the analysis. The subjectivity of the analyst will also, naturally, play a role in the process.

**IS IT THE RIGHT CODE?**

Identifying the right codes is itself a qualitative process. A few key questions, adapted from Saldaña (page 50-51, 2009)[107], are listed below. They can be considered when selecting codes. These questions may also be pondered while codes are developed and tested.

- Are the codes in harmony with the conceptual or theoretical framework?
- Do the codes relate to or address the key questions in the analysis?
- Do you feel comfortable and confident applying the codes to your data?
- Do the data lend themselves to the codes? Do the codes cut across sources (focus-group discussion, key informant interview, etc.)?
- Are the codes sufficiently specific? Or are they too broad?
- Can the codes be used as building blocks in analysis (construction of categories or taxonomies, development of patterns or themes, etc.)?
- Do the codes lend you to new discoveries, insights or patterns about your data?

105 WFP, 2009.

106 Hsieh and Shannon, 2005.

107 Saldaña, adapted version of a checklist taken from Flick (2002).

## CODING IN THE FIELD

Coding can start in the field, regardless of the exercise in question: an assessment, a monitoring or an evaluation exercise. Codes can be incorporated in memos: certain words and phrases can be underlined or WRITTEN IN CAPITAL LETTERS so that they stand out. When there are several data collectors, field teams may agree upon codes based on data collected, in the initial phases of the exercise – after which they can arrange to use these codes for all their memos, observations and quotes.

**EXAMPLE**

Coding during a rapid assessment (example taken from WFP, 2009)

"After day one in the field, the group debriefs to explore the causes of food insecurity. It produces two categories from the discussions and interviews conducted that day: food utilization and food access. After many household interviews on day two, team members agree to arrange their notes, quotes and observations according to the codes: insufficient food portions for children, no village health care and distant market."

## CODING FIELD NOTES, UNSTRUCTURED TEXT AND TRANSCRIPTIONS

Coding can also be done after data have been collected and while they are being collated. On paper or word processing software, formatting or highlighters can be used to code text. This can be particularly useful when working in a group to review text and draw up pertinent codes.

**EXAMPLE**

Colour-coded transcription of interview with female-headed household (example taken from WFP, 2009)

"During this season, we never prepare more than one meal a day. I go into the fields to work and the children must fend for themselves. When I return in the evening, I'm too tired and the children are not much help. So we eat cold whatever is left from the morning. I've sent my eldest son to look for work in the town. He hasn't sent back any money yet, so the children do not go to school."

**Coded version**

"During this season, we never prepare more than **one meal a day 1**. I go into the fields to work and the children must fend for themselves. When I return in the evening, I'm too tired and the children are not much help. So we **eat cold whatever is left 2** from the morning. I've sent my eldest son to **look for work in the town 3**. He hasn't sent back any money yet, so the **children do not go to school 4**."

**1= few meals**; **2= poor diet**; **3= seek revenue**; **4= no education**

## CODING VERBOSE ENTRIES FROM SEVERAL RECORDS

Let us say that you had, in a structured questionnaire, an open-ended question on coping mechanisms that you have to analyse. The data collectors said that the answers they were getting were similar to each other, which suggests that you may be able to categorize the data into groups. This is what, in database terminology, we call 'recoding verbose entries'. If you are dealing with comparatively few records, this may be done manually, as that is both the easiest and most accurate method. But this can be quite time consuming for a large number of records (e.g. 265). The objective is to be as efficient and as accurate as possible.

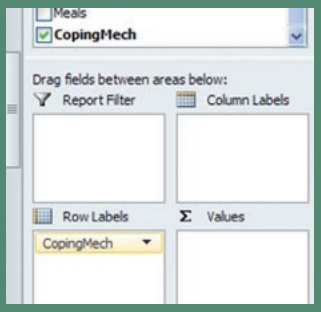
Excel can be used to attribute codes semi-automatically. This involves four main steps: 1) creating a unique list of responses; 2) creating a list of codes; 3) assigning each unique response a code; 4) matching the codes with all the responses in the list. There are a number of methods in Excel for doing each of these tasks. An example of coding in Excel is given below.

**EXAMPLE**

Coding numerous open-ended responses on coping mechanisms

Before recoding verbose entries, data should be entered into electronic format, and each record given a row of its own (see Chapter 6: Data treatment). Once data are in this format, they can be coded. The following is an example of semi-automated recoding.

- Create a unique list of responses.
- Create a pivot table with the full range of data (ensure that data are in a format where each record is equivalent to a row and each column equivalent to a variable). In the Pivot Table Field List, uncheck all fields except the one associated with the qualitative variable of interest (CopingMech, or Coping Mechanism, in the image below) and drag it under row labels.




- In the pivot table, highlight the unique list of records and copy and paste it into a new worksheet.
- Create a list of codes.
- Let's say you had a meeting with the data collectors, and having heard about their field experiences, you developed a list of codes for coping mechanisms. You add the list to your workbook, and give it a Name, CopingMechCode; include a code for 'no response/not sure' and one for 'other'.

| D |
|---|
| CopingMechCode |
| 1 - eat less or smaller meals |
| 2 - eat less meat or fish |
| 3 - eat less non-essential foods |
| 4 - borrow money/loan |
| 5 - call on family/friends |
| 6 - reduce expenses |
| 7 - sell assets |
| 8 - work more hours |
| 9 - more family members work |
| 10 - use own production |
| 11 - call for external support from community/NGOs |
| 12 - change markets |
| 13 - move |
| 14 - not a problem |
| 15 - no response/not sure |
| 16 - other |

- Code list: note that the numbers in the code list above are not hierarchical; they are meant only to serve as a sorting mechanism for the data.
- Assign each unique response a code.
- Review each unique coping mechanism and assign it a code. Simple copy and paste can be used. The analyst should consult the field team regularly – and people having 'intimacy' and 'sensitivity' with regard to the data, context, local culture and language – and report the methodology and thought process when presenting the results.

| | A | B |
|---|---|---|
| 1 | CopingMechOriginal | CopingMechMatch |
| 2 | - | 15 - no response/not sure |
| 3 | ask for help | 11 - call for external support from community/NGOs |
| 4 | ate less | 1 - eat less or smaller meals |
| 5 | become vegetarian | 2 - eat less meat or fish |
| 6 | borrow | 4 - borrow money/loan |
| 7 | borrow cash | 4 - borrow money/loan |
| 8 | borrow from friends/family | 4 - borrow money/loan |
| 9 | borrow money | 4 - borrow money/loan |
| 10 | borrow money from friends | 4 - borrow money/loan |

- Match the codes with all the unique responses in the list .
- To reincorporate the codes in the dataset, create a new variable in the main database. To do this, we can use the OFFSET and MATCH functions.
- First, create 'Names' for the original unique list (in this example, CopingMechOriginal). To do this, highlight the list, including the title, and navigate to the Formulas tab and select *Defined Names > Create from Selection*. Select *Create names from values in the: Top row*. Do the same for the assigned codes list (in this example, CopingMechMatch).
- Now highlight the top row and create a name for the list header by navigating to Formulas and Select *Create from Selection* and choose the Name CopingMechOriginal_Start.


- Navigate back to the Data tab and create a new row after Coping mechanisms and give it a long name – Coping mechanisms recoded – and a short one, CopingMechRecoded.
- We use the following formula to insert the coping mechanism code associated with the coping mechanisms reported:
- =OFFSET(CopingMechOriginal_Start;MATCH(AP5;CopingMechOriginal;0);1)
- Check the result. If Excel has returned errors (#N/A), you can fix it by simply adding an IF statement to our formula that tells Excel what to do with blank cells. Rewrite the formula, following the logic below:
- =IF(ISBLANK(AP5);"";OFFSET(CopingMechOriginal_Start;MATCH(AP5;CopingMechOriginal;0);1))
- This returns any cell which is blank as blank. This can be replaced with any other values, depending on the way missing or blank values are treated in the database.

Exceeding Excel, a computer-assisted qualitative data analysis software (CAQDAS) package, can help with the more advanced processing and analysis that can be used to systematically apply codes and perform analyses.

**POINTS TO REMEMBER WHEN CODING**

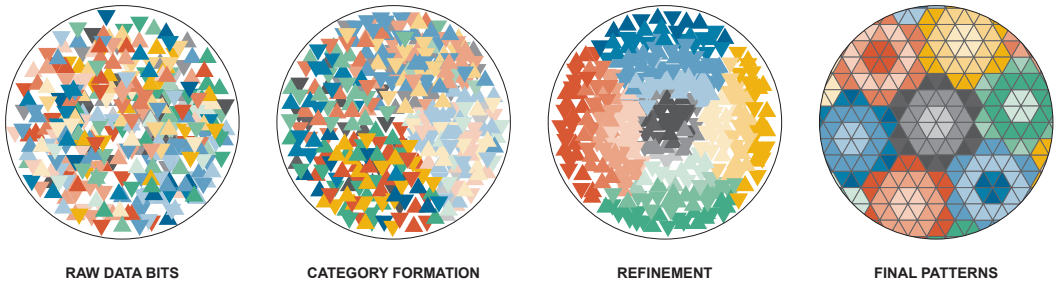
- Codes should emerge from the data and not be imposed on them.
- Coding should be systematic, and the methods recorded and shared. If more than one person is processing the data, a guidance sheet should be developed to make sure that everyone follows the same logic when assigning codes.
- Coding is iterative. Codes should be reviewed and adapted constantly, in consultation with all those who are 'intimate' with the data and 'sensitive' to them.
- Coding should be done on all relevant data (in whatever form or medium) on a given subject, until patterns and trends emerge. Because data can always be linked to their sources, end analyses can report correctly what data came from where.

## CATEGORIZATION

**Categorization** is the process of grouping data into categories according to specific criteria, which usually means that these data have something in common, such as a similar value (in the case of quantitative data) or a similar feature (in the case of qualitative data). Categories can be overlapping or non-overlapping, and a piece of data may be categorized in a number of different ways, depending on the analyst's perspective.

### THE KALEIDOSCOPE METAPHOR

Categorizing data can be compared to moving the bits of glass in a kaleidoscope: the bits of glass represent pieces of data, the two mirrors represent the categories and the two ends of the tube represent the overarching guide to creating the categories. Turning the handle on the kaleidoscope, allows you to visualize and compare many different combinations of pieces of glass (pieces of data) until a final pattern is revealed.




*Credit: Graphic adapted from Dye, Schatz, Rosenberg and Coleman (2000).*

In certain cases, a record may fit into one or more categories. Let us assume that livelihood strategies are categorized like this: fishing, agriculture, livestock farming and small business. A family could potentially fit into more than one category: agriculture and small business, for example. In other cases, records may fit into only one category. For example, if the categories for level of destruction are 'high', 'medium' and 'low', each record should fit into only one category.

**SOME POINTS TO CONSIDER WHEN DEVELOPING CATEGORIES**

| | |
|---|---|
| Categories should **examine, more or less, the same topic** | For example, let's say that we have created three categories of fruit: apples, oranges and unripe lemons. What are we trying to analyse? Do we not count ripe lemons? Do we count both ripe and unripe apples and oranges? More importantly, what is so special about unripe lemons that they need their own distinct category? |
| … and they should be **distinct** | Besides examining the same topic, categories should be distinct enough for there to be no confusion when using them. For example, let's say that you categorize food groups like this: 'vegetables', 'fruits', 'proteins', 'eggs', 'milk and milk products', 'oils' and 'sugar'. There is an overlap, as eggs and milk are sources of protein and could be considered as such. If 'proteins' is replaced by 'meat', 'fish' and 'pulses and nuts', data can still be analysed whether or not they contain protein. |
| Categories can be divided into **sub-categories**, and sub-categories into sub-sub-categories | The WFP's *Coping Strategies Index* lists diet-related coping strategies from a pilot study in Kenya. There are 12 strategies (codes or categories), which are grouped into four main categories: 'dietary change', 'increase short-term household food availability', 'decrease numbers of people' and 'rationing strategies'. The four broader categories will feed analyses that don't require such details as are contained in the sub-categories or codes.[108] |
| Categories can be **ordered** | For example, categories can be created for levels of damage to, say, a home: 'high', 'medium', 'low' and 'no damage'. In cases like this, it important to provide guidelines for determining what constitutes 'high', 'medium' or 'low' levels of damage, in order to ensure consistency of interpretation. Time and size are other examples of ordered categories. |

Quantitative analysis may include frequencies, proportions or comparisons. Analysis of categorized data will depend on data set (survey with 100 records, series of transcriptions of field notes, etc.) and on the way the categories were created and how records are matched with categories (one category per record, multiple categories per record, ordered categories, etc.)

## TYPOLOGIES

A **typology** is a type of categorization that puts subjects into groups, based on certain traits. Wealth ranking is a typology that is commonly used in livelihood analysis: levels of wealth are categorized and defined by common characteristics, such as assets, livelihood strategies and living conditions.

108 WFP, January 2008.

**EXAMPLE**

The following typologies of wealth groups of fishing households in Gaza were developed during an ICRC household economy assessment in 2008.[109]

"As the livelihood of fishery communities depends upon fishing, wealth is primarily determined by ownership of fishing equipment, with boat ownership as the only determining factor and wealth increasing with boat size and motorization. Four different wealth groups emerged during the key informant and community group discussions, from the 'very poor' households not owning any boat and working as labourers for wealthier groups, to the 'better-off' owners of trawlers or launch shanshullas."

| | Better off | Middle | Poor | Very poor |
|---|---|---|---|---|
| **% population** | 10% | 40% | 20% | 30% |
| **Household size** | 7-9 | 7-9 | 7-9 | 7-9 |
| **Characteristics** | ▪ Having trawler & ▪ launch shanshulla or launch hasaka boats ▪ Permanent employment ▪ Traders | ▪ Owning hasaka shanshulla ▪ Some also own 1 rowing boat or hook hasaka | ▪ Owning 1 rowing boat ▪ In MKH owning 1 hook hasaka ▪ Casual work | ▪ Not owning boat ▪ Working for the middle and better off ▪ Casual work |

## TRIANGULATION

In the social sciences, **triangulation** is the process of combining or comparing several sources and/or observations on a given topic, with the aim of increasing confidence in the result by decreasing the bias associated with 'one side of the story'. The end goal of triangulation is to reveal: converging results, complementary results and contradictions.[110]

Triangulation can be done on many levels, as data might come from various sources or via many different data-collection methods and on different scales. The following diagram shows how data may be collected using different methods and different sources (i.e. interviews with different people and observations), and on various scales (i.e. entire sets of data and single cases).

While data may be at different levels or on various scales, a link needs to be identified in order to triangulate them: What is the basis of comparison? For example, if we are trying to understand why people in a particular mining community don't have enough food to eat, and our data include household interviews, notes from field trips, historical reports and focus-group discussions with women from mining households, we would need to find out how the matter is discussed and/or recorded in each, and how these various data can be compared.

### TRIANGULATION MATRIX

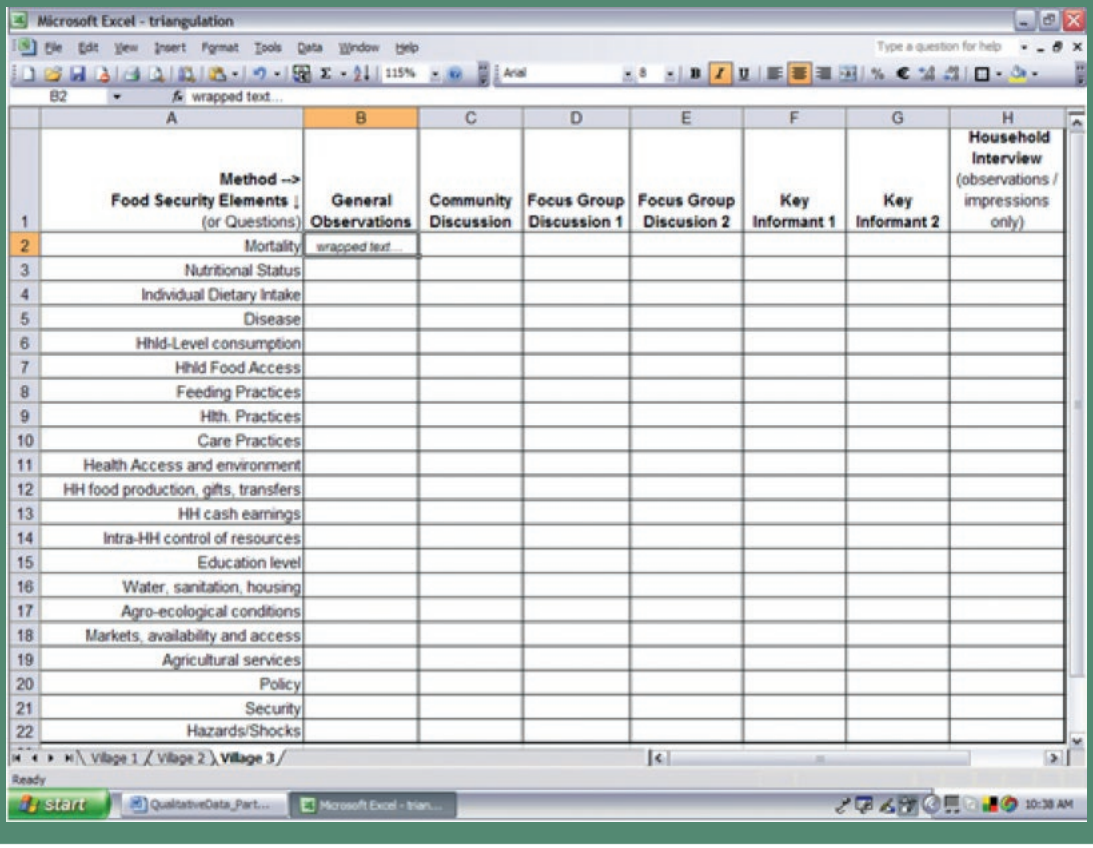
A triangulation matrix is a relatively simple tool for summarizing data and comparing them across sources and themes. It can help to facilitate discussions among the interpretation team on patterns in the data. The matrix may be quite large and may have too much information to serve as a 'pretty' visual with a 'jumping' pattern, but this may be a minor consideration, as the matrix's main purpose is to reveal information. It is then up to the analyst and the interpretation team to decide what to report and how to do so.

109 ICRC, 2008.

110 Flick, 2009.

**EXAMPLE**

The following table shows a triangulation matrix developed by the WFP (2009) in Excel. The matrix is organized so that indicators and variables are in rows and the various sources of data are in columns. A separate matrix is developed for each community.




| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | Method --> Food Security Elements ↓ (or Questions) | General Observations | Community Discussion | Focus Group Discussion 1 | Focus Group Discusion 2 | Key Informant 1 | Key Informant 2 | Household Interview (observations / impressions only) |
| 2 | Mortality | wrapped text... | | | | | | |
| 3 | Nutritional Status | | | | | | | |
| 4 | Individual Dietary Intake | | | | | | | |
| 5 | Disease | | | | | | | |
| 6 | Hhld-Level consumption | | | | | | | |
| 7 | Hhld Food Access | | | | | | | |
| 8 | Feeding Practices | | | | | | | |
| 9 | Hlth. Practices | | | | | | | |
| 10 | Care Practices | | | | | | | |
| 11 | Health Access and environment | | | | | | | |
| 12 | HH food production, gifts, transfers | | | | | | | |
| 13 | HH cash earnings | | | | | | | |
| 14 | Intra-HH control of resources | | | | | | | |
| 15 | Education level | | | | | | | |
| 16 | Water, sanitation, housing | | | | | | | |
| 17 | Agro-ecological conditions | | | | | | | |
| 18 | Markets, availability and access | | | | | | | |
| 19 | Agricultural services | | | | | | | |
| 20 | Policy | | | | | | | |
| 21 | Security | | | | | | | |
| 22 | Hazards/Shocks | | | | | | | |

Triangulation matrices can be used in both secondary-data analysis and primary-data analysis, or when the two are combined. In primary-data analysis, the matrix can be developed during the early stages of data collection and, in the field, completed by field teams at the end of each day during debriefing sessions. It can also be used by teams, when they are back in the office, to collate the data. In secondary-data analysis, the matrix can be developed by analysts during the early stages of data collation to help force the data into a format in which they can be compared quickly and easily.

**STEP-BY-STEP CREATION OF A TRIANGULATION MATRIX[111]**

- **Set up the matrix** so that the rows are associated with different themes and the columns with the different sources of data or the various methods employed when collecting primary data.
- **Ensure that the themes are specific** enough to meet the analytical objectives and to be accurately completed. Ambiguous themes lead to ambiguous results.
- **Record relevant bits of qualitative data** in their proper places within the matrix. Capitalize on all data, including observations, perceptions, direct quotations, etc., but use column headers to differentiate the various sources/methods.

## RELATIONSHIPS AND TRENDS

As discussed at the beginning of this chapter, the goal in qualitative analysis is to uncover themes, patterns, concepts and insights, which may disclose themselves at any stage, from data collection in the beginning to analysis at the end. The following section reviews various thought processes and techniques that can be used to identify the relationships and trends that can, in turn, help to uncover these themes, patterns, concepts and insights.

111 Adapted from WFP, 2009.

## TYPES OF RELATIONSHIPS

As mentioned in the previous chapter, a relationship is a correspondence, connection or link between two or more variables of interest. Cross-tabs and correlation coefficients are use to explore relationships in quantitative analysis. In qualitative analysis, too, relationships are explored by comparing data and groups, but more descriptive techniques are used. A list of semantic relationships and their basis is given below. It can serve as a guide for exploring relationships and for explaining the findings.[112]

| TYPE | FORM OF RELATIONSHIP |
|---|---|
| Strict inclusion | X is a kind of Y |
| Spatial | X is a place in Y; X is a part of Y |
| Cause-and-effect | X is a result of Y; X is a part of Y |
| Rationale | X is a reason for doing Y |
| Location for action | X is a place for doing Y |
| Function | X is used for Y |
| Means to an end | X is a way to do Y |
| Sequence | X is a step (stage) in Y |
| Attribution | X is an attribute (characteristic) of Y |

In qualitative analysis, these relationships can be explored through the use of matrices, diagrams, maps, timelines and storylines. These relationships may lead to the identification of additional patterns (recurring and predictable relationships).

## MATRICES AND FRAMEWORKS

A **matrix** is a tool that can be used to look at the intersection of two constants, variables or processes. Like cross-tabs in quantitative data analysis, matrices are used to explore relationships and make comparisons between locations and/or groups. Qualitative matrices, however, place descriptions and words, not numbers, in each cell.

112 Source: J.P. Spradley, 1979. Taken from Whitehead, 2005.

**EXAMPLE**

The following matrix for coping strategies was developed after a series of group discussions with communities in northern Mali in 2014. The objectives of this part of the discussion were: to identify the various coping strategies employed by community households when they have difficulties in obtaining food or earning an income; to discuss why households used the strategies they did; and to understand the effects (both positive and negative) of the strategy on the household from the point of view of the community. The end analysis allowed the team to develop a severity ranking of food and livelihood-related coping strategies that was the basis not only for analysing household-level data collected during this exercise in 2014, but potentially for future exercises as well.

| Level of severity | | Coping strategies | Effect on household economy | Reasons household may use strategy |
|---|---|---|---|---|
| 1 | Reversible effect on household economy | • **"Solidarité"** (support of family/friends)<br>• **Work more** (either more hours or engage in additional activities)<br>• **Replace certain household products** (either food or non-food) with products that are less expensive and do not have a negative effect on health/nutrition<br>• **Reduce daily expenditure** and/or **use savings** | Complementary food or income without any effect on household assets | Easy to use; socially acceptable |
| 2 | | • **Commodity loans**<br>• **Reduce food intake** (either quantity or quality)<br>• **Sell more production than expected** (animals, agricultural production, etc.) | Complementary food or income with a reversible effect on household assets | |
| 3 | | • **Monetary loans** (loan, credit, etc. from others or institutions)<br>• **Seek external aid** (beyond family and friend such as humanitarian aid) | Complementary food or income with an effect difficult to reverse on assets and/or slight effect on human well-being | |
| 4 | | • **Sell productive assets**<br>• **Remove children from school** | Complementary food or income with a permanent effect on household assets and/or significant effect on human well-being | |
| 5 | Irreversible effect on household economy | • **Move or migrate**<br>• **Sell irreplaceable family goods** (e.g. jewellery)<br>• **Engage in illegal activities or prostitution** | Loss of all household assets and/or an extreme effect on human well-being | No other solution |

## DIAGRAMS

A **diagram** is a symbolic representation of information.[113] Diagrams can be developed on paper, white boards or, even more creatively, using string and objects. They can be drawn or created by the analyst or together with a team or even as a participatory tool in the field. Final products can simply be transferred into electronic format in MS Word, PowerPoint or Excel using one of the smart art tools.

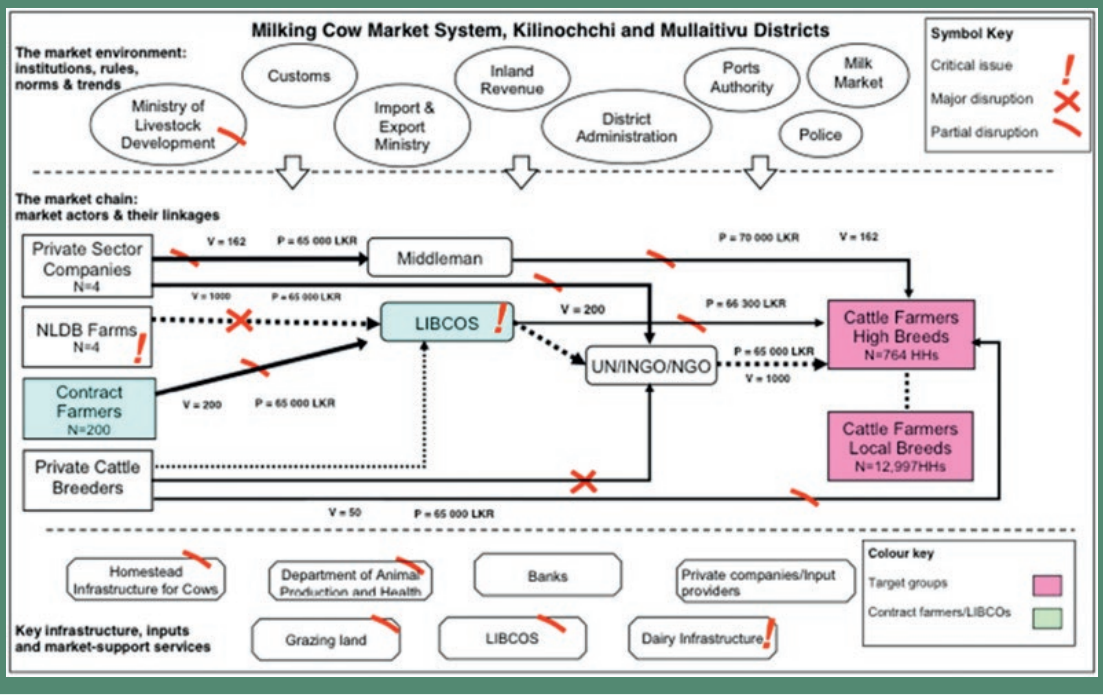
113 Wikipedia, Wikipedia entry on "Diagram", accessed in March 2015.

## NETWORK DIAGRAMS

A **network diagram** is a drawing of various subjects and variables and the set of dyadic ties between them.[114] A network diagram provides a useful structure for analysing a social process or entity. Market chain mapping and social network mapping –both network diagrams – are commonly used in humanitarian work.

### EXAMPLE

The following diagram produced by Oxfam (May 2012) represents the milking cow market system in the Kilinochchi and Mullaitivu districts in Sri Lanka. The diagram uses circles and squares to highlight subjects (actors, processes, institutions, services and infrastructure) and arrows to represent processes. It even highlights where these processes are working and where they are not.



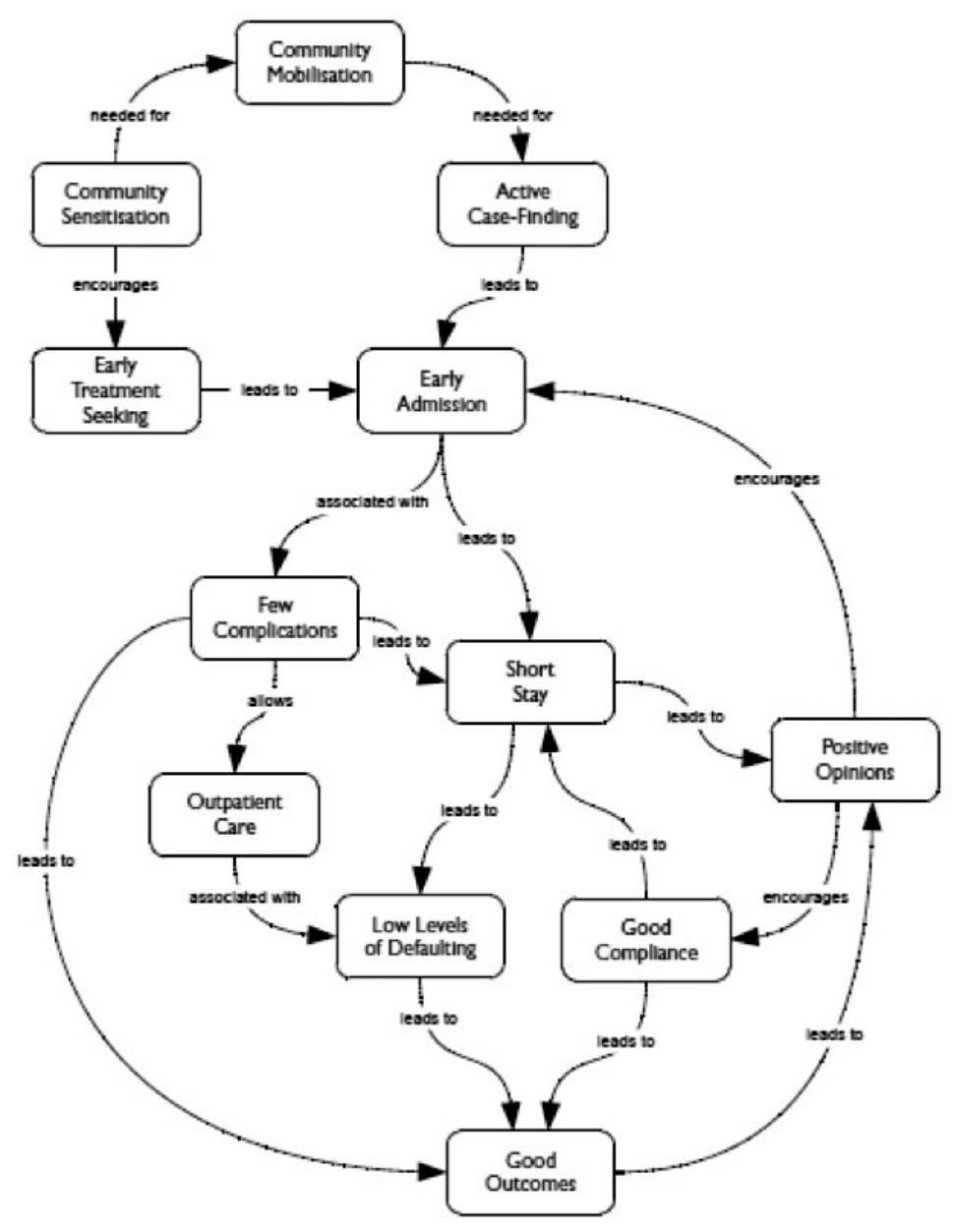

114 Wikipedia, Wikipedia entry on "Social Network", accessed in April 2015.

## PROCESS DIAGRAMS

A **process diagram** is a type of diagram that maps processes (events, decisions, activities, etc.) and their outcomes. It can be on one level (e.g. one process leading to one outcome) or on many (e.g. a process leads to an outcome that leads to another process and outcome, and so on).

### EXAMPLE

The following process diagram depicts the relationships between factors influencing the coverage and those influencing the effectiveness of community-based management of acute-malnutrition programmes. The diagram was used to develop a mixed-method approach to the evaluation of malnutrition programmes, which was called the Semi-Quantitative Evaluation of Access and Coverage (SQUEAC)/Simplified Lot Quality Assurance Sampling Evaluation of Access and Coverage (SLEAC) (FANTA, October 2012).

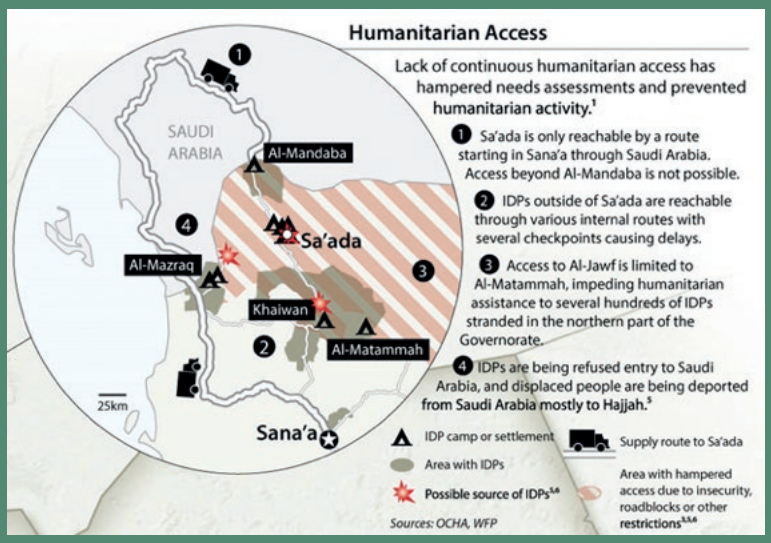


## GEOGRAPHIC MAPS

**Geographic maps** are useful for exploring relationships between a process or subject and space, and identifying patterns such as clusters, concentrations or gaps. They can be used in the field, during the data-collection phase, to map concepts together with informants or the field data-collection team. In the data-analysis phase, data can be coded or categorized by geographic location; this will help to map them and to explore patterns through space.

**EXAMPLE**

The following map prepared by UN OCHA ReliefWeb (December 2009) shows the various routes for reaching displaced people in northern Yemen. Mapping the data provides the analyst with a first glimpse of where exactly people may or may not be easily reached. It can also provide an indication of scale (as shown by the length of the routes), however faint, that could be supported by quantitative analysis or geospatial analysis.



## TIMELINES AND STORYLINES

**Timelines** can be used to explore trends and patterns through time. Like mapping, timelines can be used in the field in the data-collection phase, to walk through events together with informants or the field data-collection team. In the data-analysis phase, data can be coded or categorized by date, time of year, relation to an event (before, after, during, etc.); this will help to map them and to explore patterns through time. Seasonal calendars, timelines of political events, timelines of violent events and timelines of population movement: these are some of the qualitative timelines (which are different from quantitative time series) that are commonly used in humanitarian work.

**EXAMPLE**

In the example from UN OCHA Reliefweb on Yemen, the same analysis categorized various conflict events over time using the metaphor of 'waves'.

### Six waves of conflict

**Aug 2009 - present -** Sixth round of conflict between Al-Houthi and the Government of Yemen breaks out on 12 August. A ceasefire is announced by the Government on 4 September to allow for the distribution of humanitarian aid, however it only lasted a few hours.[6]

**May - Jul 2008 -** The government accuses Al-Houthi of violating the ceasefire agreement signed in Doha in February 2008, and the fifth round of conflict breaks out. On 17 July, the president announces a unilateral ceasefire.[7]

**Feb - Jun 2007 -** Fourth round of conflict spreads to districts outside of Sa'ada.

**late 2005 - early 2006 -** Third round of conflict starts as a confrontation between pro-government tribesmen and tribal fighters supporting Al-Houthis.

**Mar - May 2005 -** Second round of conflict erupts in north and west of Sa'ada, where the Al-Houthis found support and the mountainous terrain slowed the army.

**Jun - Sep 2004 -** Conflict starts southwest of Sa'ada city where the first head of Al-Houthi, Husein al-Huthi, sought refuge. Following al-Huthi's death on 10 September, the Government declared a unilateral end to the fighting.

A **storyline** is a narrative account – for our purposes, an account of humanitarian interest – that describes a sequence of events with a conclusion, or end result. An analytic storyline is less 'structured' than a process diagram or a timeline: it uses descriptive words to explain the transition and flow from one event or process to another. A qualitative analysis in the form of a storyline can be a powerful way to report on the development or the evolution of a situation or phenomenon.

# REFERENCES

Dye, J. F., *et al.,* (2000, January). "Constant comparison method: A kaleidoscope of data" [24 paragraphs], *The Qualitative Report*, Vol. 4, Nos 1/2, January 2000. Available at: http://www.nova.edu/ssss/QR/QR4-1/dye.html.

FANTA, *Semi-Qualitative Evaluation of Access and Coverage (SQUEAC)/Simplified Lot Quality Assurance Sampling Evaluation of Access and Coverage (SLEAC) Technical Reference*, FANTA, Washington D.C., October 2012. Available at: http://www.fantaproject.org/monitoring-and-evaluation/squeac-sleac.

Flick, Uwe, *An Introduction to Qualitative Research*, 4th ed., SAGE Publications, London, 2009.

Harvard Humanitarian Initiative, *Disaster Relief 2.0: The Future of Information Sharing in Humanitarian Emergencies*, Washington, D.C. and Berkshire, UK, United Nations Foundation and Vodafone Foundation Technology Partnership, 2011.

Hsieh, H.F., Shannon, S.E., "Three Approaches to Qualitative Content Analysis", *Qualitative Health Research,* Vol. 15, No. 9, November 2005, pp. 1277-1288. Available at: http://www.iisgcp.org/pdf/glssn/Supplemental_Reading_on_Coding_2.pdf .

ICRC, *Democratic Republic of the Congo: Initial Assessment in Kahele, South Kivu*, June 2014.

ICRC, *Household Economy Assessment in the Gaza Strip*, May-June 2008.

Oxfam/ECHO, *Emergency Market Mapping and Analysis of the Milking Cow Market System*: *Kilinochchi and Mullaitivu Districts, Northern Province, Sri Lanka, 1-18 May 2012.* Available at: http://www.emma-toolkit.org/report/milking-cow-market-system.

Saldaña, Johnny, *The Coding Manual for Qualitative Researchers*, SAGE Publications, London, 2009.

Saldaña, Johnny, *Fundamentals of Qualitative Research*, Oxford University Press, New York, 2011.

Strauss, Anselm, Corbin, Juliet, *Basics of Qualitative Research*, SAGE Publications, London, 1998.

UN OCHA ReliefWeb, *Yemen: Conflict and Displacement in the North*, December 2009. Available at: http://reliefweb.int/map/yemen/yemen-conflict-and-displacement-north-humanitarian-snapshot-dec-2009.

WFP, USAID, Feinstein Group, TANGO, CARE, *The Coping Strategies Index: Field Methods Manual,* 2nd ed., January 2008. Available at: https://www.wfp.org/content/coping-strategies-index-field-methods-manual-2nd-edition.

WFP, *Emergency Food Security Assessments (EFSAs) – Technical Guidance Sheet No. 9: Qualitative Data Collection and Analysis for Food Security Assessments*, WFP, Rome, 2009.

Whitehead, T.L. (2005), "Basic Classical Ethnographic Methods", *CEHC Working Papers*, TL Whitehead Associates. http://tonywhitehead.squarespace.com/tools-products/.

# CHAPTER 9
# COMBINING ANALYSES AND DRAWING CONCLUSIONS

In the previous chapters, we discussed various techniques of quantitative and qualitative analysis in which data and variables are reviewed, the appropriate techniques chosen and various kinds of exploratory analysis undertaken. The next logical step is to bring together all the data and information and make sense of it all. What does it mean? What can we say? What can we do? So what? Drawing conclusions involves combining all the different pieces of analysis and interpreting them to determine what exactly they mean, their relevance to the study and their validity. The process can involve the following tasks:

- **combining** all the analyses and secondary information to contribute to a greater macro-analysis
- performing **further micro** and **macro analyses** as necessary (iterative process)
- reviewing the **plausibility** and **validity** of the findings
- **interpreting** the findings
- **drawing conclusions.**

## COMBINING AND REANALYSING

The process of combining pieces of data and analysis to contribute to the overall analysis of a study is guided by such things as the objectives of the study and the key questions, the analytical tools identified and/or developed at the beginning of the study (indicators, criteria, frameworks, analysis plans, etc.), and possibly also by new insights, theories and models identified and/or developed during the study (unforeseen in planning).

### FITTING DATA TO A FRAMEWORK WITH PRE-IDENTIFIED CRITERIA

If the study makes use of a framework for analysis, combining analyses will involve making the data fit into the framework. Making the data fit specific criteria can be done at the record level (e.g. case monitoring) or on aggregated data (e.g. global risk index). These criteria can include qualitative and both quantitative variables, and the resulting analysis can be done using descriptive text, matrices and/or mathematical formulas.

**EXAMPLE**

In Chapter 3: Analysis design, we looked at the vulnerability framework for Lebanese returnees from Syria. Developing the framework was one step in the process. The next step was to feed the framework with data The following is a preview of the database. The datasheet was developed in Excel, using a simple format that was described in Chapter 4: Primary-data collection. Excel formulas were used to calculate the scores based on the results. The chart below show the individual and combined scores for a number of records. This simple analysis allows the team to see not only what a household's overall score is, but also what component is contributing most to their vulnerability. For example, one household may be 'income-rich', but may also have a high dependency ratio and poor living conditions.

| VULNERABILITY SCORE | | | | | | | | | VULNERAB |
|---|---|---|---|---|---|---|---|---|---|
| Overall | by Category | | | | | | | | 1 - HH Comp |
| Overall vulnerability | Household composition | Income | Living conditions | Food consumption | Health | Coping mechanisms | Protection | Assistance | Dependency |
| VULS | VULS_HHC | VULS_INC | VULS_LC | VULS_FC | VULS_HLT | VULS_CS | VULS_PRC | VULS_ASS | INDS_HHC_ |
| 20.13 | 5.0 | 3.0 | 3.1 | 1.0 | 3.0 | 3.0 | 1.0 | 1.0 | 5.0 |
| 19.33 | 1.0 | 2.3 | 3.0 | 1.0 | 3.0 | 3.0 | 1.0 | 5.0 | 1.0 |
| 21.08 | 1.0 | 2.3 | 2.8 | 1.0 | 5.0 | 3.0 | 1.0 | 5.0 | 1.0 |
| 22.25 | 1.0 | 3.0 | 3.3 | 1.0 | 5.0 | 3.0 | 1.0 | 5.0 | 1.0 |
| 23.75 | 2.0 | 4.0 | 2.8 | 1.0 | 5.0 | 3.0 | 1.0 | 5.0 | 2.0 |
| 17.54 | 1.0 | 3.7 | 3.4 | 1.5 | 3.0 | 3.0 | 1.0 | 1.0 | 1.0 |
| 21.04 | 1.0 | 3.7 | 3.4 | 1.0 | 5.0 | 3.0 | 1.0 | 3.0 | 1.0 |
| 21.38 | 1.0 | 3.0 | 2.4 | 1.0 | 5.0 | 3.0 | 1.0 | 5.0 | 1.0 |
| 16.50 | 1.0 | 2.0 | 2.5 | 1.0 | 5.0 | 3.0 | 1.0 | 1.0 | 1.0 |
| 19.88 | 1.0 | 3.0 | 2.9 | 1.0 | 5.0 | 3.0 | 1.0 | 3.0 | 1.0 |
| 22.83 | 1.0 | 3.3 | 3.5 | 1.0 | 5.0 | 3.0 | 1.0 | 5.0 | 1.0 |
| 18.13 | 5.0 | 3.0 | 2.1 | 2.0 | 3.0 | 1.0 | 1.0 | 1.0 | 5.0 |
| 17.71 | 1.0 | 4.3 | 3.4 | 1.0 | 3.0 | 1.0 | 1.0 | 3.0 | 1.0 |

The key to feeding data into a framework is to ensure the linkages (interoperability) between the analysis (the framework), the questionnaire and the database. A database or information management specialist can be of great help in developing technical tools.

## COMBINING ANALYSES

We have reviewed ways of looking at patterns, relationships and trends in quantitative and qualitative analysis. We studied statistics for quantitative analysis such as cross-tabs, correlation coefficients, timelines and geostatistics; and for qualitative analysis, we looked into matrices, diagrams, maps and timelines. We may, however, be required sometimes to combine qualitative and quantitative analysis, and to put together all the different pieces of analysis to uncover critical patterns, trends and interrelationships.

Some of the techniques discussed in qualitative analysis can be employed – triangulation matrices, bivariate or multivariate matrices, diagrams or timelines – that take into account both quantitative and qualitative data from a variety of sources and topics. The idea is to look beyond one piece of analysis and try to understand how all the pieces of data and analysis interrelate, using the analytical tools at our disposal (What should we be looking for? What can we expect to find?) and guided by any new insights we may have acquired (What interesting reflections did I put in my memos?).

## DISCERNING WHAT IS MEANINGFUL AND THE PLAUSIBILITY AND VALIDITY OF THE FINDINGS

Fitting data to criteria or putting together various analyses may confirm the validity of some models and reveal the inadequacy of others; it may also raise more questions. For example, the vulnerability and selection criteria for Lebanese returnees went through a number of iterations until the team found the right fit for the situation and the context. Reviewing may involve digger deeper into the data, rereading memos, performing further analysis and even collecting more data.

What is meaningful in the data may be self-evident. The following table from ACAPS[115] highlights three groups of questions that can be used to identify patterns and trends. These can serve as a guide for thinking through the analytical process.

**ANALYTICAL THOUGHT PROCESS FOR IDENTIFYING PATTERNS AND TRENDS**

| | |
|---|---|
| Define significant parts and make the implicit explicit | ■ Which details seem significant?<br>■ What does the detail mean?<br>■ What else might it mean? |
| Look for patterns | ■ How do the details fit together? What do they have in common?<br>■ What does this pattern of details mean?<br>■ What else might this same pattern of details mean? How else could it be explained? |
| Look for anomalies and keep asking questions | ■ What details don't seem to fit? How might they be connected with other details to form a different pattern?<br>■ What does this new pattern mean? How might it cause the meaning of individual details to be read differently? |

Source: ACAPS, 2013, p. 13.

115 Adapted from Stephen and Rosenwasser, 2012.

## PLAUSIBILIY AND VALIDITY

**Plausibility** and **validity** are terms used to describe the truthfulness and reliability of the data and analysis. Without evidence of plausibility and validity, data and analysis run the risk of being rejected and the credibility of the research team, of being compromised.

### PLAUSIBILITY

**Plausibility** is the appearance of having truth or of being credible. Plausibility is important in analysis because without it, we risk rejection of our analysis. An analyst can give plausibility to a questionable claim by supporting it with 'corroborating' and/or 'converging' evidence.

**Corroborating evidence** is made up of several pieces of evidence (data or analysis) that support a conclusion. **Converging evidence** consists of individual pieces of information that do not by themselves support a conclusion, but when combined, constitute a robust body of evidence in support of the conclusion. Converging evidence requires reasoning and data that are contextualized.[116]

**EXAMPLE**

The following pieces of data, collected for an economic security assessment in northern Mali in 2014, provide converging evidence that the increasing cost of wood (used for cooking) in northern markets have an effect on household income and the community's natural assets. Data were collected from three different primary sources: monthly market price monitoring, household surveys and focus-group discussions.

**Monthly market price monitoring**

- The price of wood in Léré is 733 West African CFA franc (CFA) for a small bundle (633% more than 3 months ago)
- The price of wood in Kidal is 3,000 CFA for a small bundle, (29% less than 3 months ago)
- The price of wood in other markets in the region ranges between 100 and 250 CFA for a small bundle

**Household survey**

| Location | Closest large market | Monthly household expenditures on energy (average in CFA) |
|---|---|---|
| Goundam | Léré | 5,395 |
| Essouk | Kidal | 4,583 |
| Tilemesi | Léré | 3,500 |
| Léré | Léré | 3,144 |
| Soumpi | Léré | 3,095 |
| Koumaïra | Léré | 2,952 |
| Kidal | Kidal | 2,928 |
| Tonka | Léré | 2,806 |
| Monzonga | Ansongo | 1,562 |
| Seyna | Ansongo | 832 |
| Basi Gourma | Ansongo | 604 |

**Focus-group discussions**

Focus-group discussions in Soumpi – Léré is the closest large market to it – revealed that there have been an increase in damage to the forest due to households cutting trees to obtain firewood, a coping strategy for market volatility.

Evidence may support more than one 'plausible' claim, and it is up to the analyst, with the support of peers, to highlight the most plausible one in the context of this particular analysis. The various claims should be communicated transparently in the report, with arguments in support of granting one more relevance than all the others.

### VALIDITY

**Validity** is the extent to which a concept, conclusion or measurement is well founded and corresponds to the real world. Validity has been discussed in several places in this guide, when addressing bias, statistical significance, confidence intervals, etc.

116 ACAPS, 2013.

Validity in qualitative analysis relies on the analyst's description of what it is to be well-founded and valid, and on corroborating and converging evidence; quantitative analysis can employ certain statistical tests to this end. The following table sets out a number of criteria for judging the validity of qualitative analysis and the equivalent in quantitative analysis.[117]

| TRADITIONAL CRITERIA FOR JUDGING QUANTITATIVE RESEARCH | | EQUIVALENT CRITERIA FOR JUDGING QUALITATIVE RESEARCH | |
|---|---|---|---|
| **Internal validity** | The extent to which a study is free of error or bias (type I or type II errors) | **Credibility** | The results are credible from the perspective of the participants |
| **External validity** | The amount of generalizability of the findings (confidence intervals) | **Transferability** | The degree to which the results can be generalized to other contexts or settings |
| **Reliability** | The reliability of the estimate (precision) | **Dependability** | The degree to which the results would be the same or similar if repeated |
| **Objectivity** | The decisions on and methods for collecting measurements and performing analyses are done in an objective manner | **Confirmability** | The degree to which the results can be confirmed or corroborated by others |

Triangulation and consultation, or peer review. are two methods that can be used to explore the level of agreement with the criteria mentioned above.

Triangulation, as discussed in previous chapters, is crucial for 'building' validity associated with all criteria. The validity of results may increase exponentially with the level of agreement between different pieces of data and analysis, methods of data collection and analysis, and/or sources.

Peer review, or consultation, entails working together with informants, data collectors, analysts, other sources of knowledge, etc. to review the analysis and determine the quality and validity of the results until a certain level of agreement is shared. Such reviews contribute to credibility and confirmability. ACAPS proposes a simple matrix for ranking degree of confidence that graphs 'agreement' (either through peer review/consultation or triangulation) against 'evidence'.

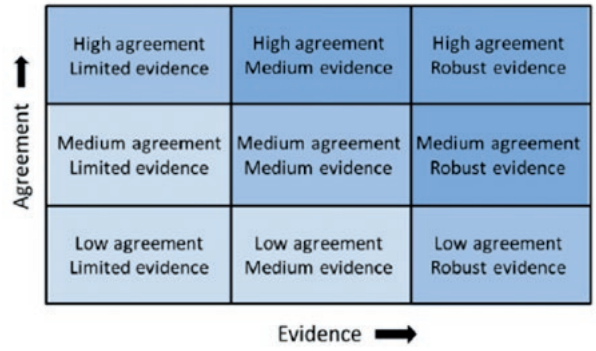



**Figure 39** - **Ranking confidence in qualitative analysis through analysis of the level of agreement and the level of evidence. Source: ACAPS, 2013.**

117 Trochim, 2006.

# INTERPRETATION

**Interpretation** is the process of determining what an analysis means through contextualization, use of experience, and selection of the most important findings – in order to draw conclusions.

## CONTEXTUALIZATION

How many times have you heard someone say, "It depends on the context"? Food insecurity in rural Central African Republic may have a very different meaning than in the urban United States; the destruction of 350 homes in a village of 5,000 inhabitants is very different from that of 350 homes in a city of 5 million inhabitants. Data and analysis have to be put into context to be correctly understood and to reach their full potential in terms of interpretation, understanding and messaging.

Context is essentially information that locates data and analysis in a time and a place. Context can be provided by such things as background information, knowledge and experience, triangulation, lessons learnt and judgement. Context should be incorporated in the analyst's way of thinking and working, and used as part of the analysis when sharing and reporting conclusions.

### MAINSTREAMING CONTEXT

The following table[118] provides a list of practices for incorporating context in the interpretation of data and analysis.

| TIPS FOR INCORPORATING CONTEXT IN INTERPRETATION | |
|---|---|
| Consider **past experience and lessons learnt** (knowledge), and have an understanding of the **theoretical basis**, and a grasp of what is known, in connection with the topic at hand | ■ Ensure that conclusions follow the same logic.<br>■ Know what outcomes one can reasonably expect (or not) from a particular phenomenon or process. |
| **Relate new information** to what is already known. | ■ Use data attributes (p-codes, location, time, group, etc.) to determine how the new observation relates to existing and historical observations. |
| Recognize when **two items are the same or different**. | ■ Look for uniqueness in pieces of data (e.g. are these two different records or the same record reported twice?).<br>■ Be 'sensitive' to subtle differences in the local context (e.g. the definition of 'family' in some contexts may extend beyond blood relatives) |
| Analyse data points as they come in and ensure that **such analyses are informed and complemented by the big picture**. The aim is to understand how larger pieces of information potentially fit together and what the whole set of information depicts. | ■ Use debriefing, brainstorming or interpretation sessions with all relevant data and information and knowledge brokers to think through the big picture.<br>■ Incorporate monitoring in your daily work plan. |
| **Be aware of new observations that have changed earlier assertions,** or revise or rectify invalidated assertions accordingly In an iterative way. | ■ Develop new versions of and update existing analysis, and be transparent with regard to changes in the analysis as new information becomes available. |
| **Let disagreement and conflict coexist** in collected data; otherwise, emerging trends will not have a chance to add up to anything when interpreted. | ■ Don't 'over clean' the data available, keep original field questionnaires, and be open to the value of low-quality or non-coded data. |

118 Adapted from ACAPS, 2013.

### PEER REVIEW

**Peer review** is essential for interpreting end analyses. Large-scale research often incorporates formal consultation in the process. Peer review can start even before a report is written, during the analysis or when the process of drawing conclusions is just getting under way. For larger exercises, it is useful to plan an interpretation session that includes everyone involved in the data collection and analysis, other specialists in the field and all pertinent disinterested parties.

Sitting down with the field teams after crunching all the numbers, looking through them together and reflecting on their meaning can be a revelatory experience.

## COMMUNICATION

The final stage (until the next exercise) is communicating the findings. This is usually done by means of a written report, but may also include visuals, presentations, interviews, etc. Communication involves determining not only the information that must be reported, but also the order in which that has to be done, and the most effective means of doing so. A written report of data collection and analysis should always include at least the following:

- the objectives of the exercise
- the theories, models or frameworks used to guide the analysis
- the methods for data collection and analysis
- the tools used in data collection and analysis
- the limitations of the analysis
- the results and their validity
- the conclusions
- the errors and inconsistencies, and the lessons learnt along the way.

Visualizations or excerpts should always include the sources of the data, the dates on which the data were collected, and all significant constraints to their use. The EcoSec Assessment Report template can be used as a checklist for what to include in a complete assessment report, and the EcoSec Executive Brief template for shorter briefing papers. Both are available at the EcoSec Resource Centre under Reporting Templates.

# REFERENCES

ACAPS, *Compared to What? Analytical Thinking and Needs Assessments*, August 2013. Available at: http://www.acaps.org/img/documents/c-160806-tb-compared-to-what-final.pdf.

ACAPS, *How Sure are You? Judging Quality and Usability of Data Collected during Rapid Needs Assessments*, August 2013. Available at: http://www.acaps.org/img/documents/h-130827-tb-how-sure-are-you-final.pdf.

ACAPS, *Quantitative and Qualitative Research Techniques*, May 2012. Available at: http://www.acaps.org/img/documents/q-qualitative-and-quantitative-research.pdf.

ICRC, *Household Economy Assessment in the Gaza Strip*, May-June 2008.

Rosenwasser, David and Stephen, Jill, *Writing Analytically – with Readings*, 2nd ed., Nelson College Indigenous, 2012.

Trochim, William M.K, *Research Methods: Knowledge Base*, Web Centre for Social Research Methods, 2006. Available at: http://www.socialresearchmethods.net/kb/qualval.php.

# CHAPTER 10
# VISUALIZATION

## WHY VISUALIZE?

**Visualization** is a term for any technique used to create a graphic, image or piece of animation to explore, interpret and/or communicate a piece of data, concept, event and/or message. Graphs/charts, maps, information graphics, diagrams, photographs and videos are visualizations that are often employed in economic security analysis and reporting. They can all be used to explore or communicate both concrete and abstract ideas.

*We visualize to:*

**THINK & LEARN**

Explore, find stories and make sense of your data

**COMMUNICATE**

Tell stories and explain your findings

This chapter focuses on **data visualization** and **visual diagrams,** and on the use of graphics in assessments, activities, and monitoring and evaluation exercises. It is meant to be of help to staff who need to develop compelling stories and arguments with the means available to them. It is not intended to provide training in graphic design. That being said, given the tools that are readily available today, and if certain general steps and principles are followed, everyone can design solid visuals .

## DEVELOPING VISUALS

Developing visuals, in the context of this guide, starts after data have already been collected and cleaned.[119] It involves three main steps: **exploring, designing and building**.

Data exploration uses visuals to *think and learn* during the analytical process and, later, while developing visuals. Designing and building are steps on the way to developing suitable visuals to 'communicate' what was learnt.

### EXPLORING

**Data exploration** is the process of getting to know the data and its subtleties. It aims to uncover patterns, trends and outliers in order to make sense of the data and identify the stories it can tell. The process is guided, first, by the analyst's ability to **know what to look for in the data**, and second, by **the use of the appropriate analytical techniques**, including the appropriate visualizations.

Getting to know data includes becoming 'intimate' with the data and 'sensitive' to what it is trying to say. Data intimacy requires the analyst to know not only the data (How was the variable collected? What is the unit of measurement? What is the response rate?), but also their sources (Who collected the data? Where did they get the data? Who shared it? Who is included in the sample? When was it collected?). It is vitally important for the analyst to have been involved in designing the analysis and in collating and treating data, and/or to work closely with the data collectors, data sources and/or other analysts when proceeding with the analysis.

119 Development of visuals may be part of an analysis plan before data are collected, but keep in mind the fact that data analysis can and will reveal unforeseen patterns, trends and stories. Additional data collection, extraction (particularly in the case of qualitative data) and manipulation may be required after the first data exploration and/or during the design phase.

**EXAMPLE**

Data collected from a secondary source may require the analyst – to fully understand the data and to correctly interpret the findings – to follow up with the source or with relevant knowledge-brokers (e.g. people with experience in the context or on the subject) when patterns and trends emerge. You can test the truth of this by downloading data from an open source humanitarian data portal, such as the UNHCR population statistics featured in the section on visual principles (http://popstats.unhcr.org), Humanitarian Data Exchange (https://data.hdx.rwlabs.org/), World Bank Open Data (http://data.worldbank.org) or EM-DAT (http://www.emdat.be/), and then developing graphics with these data. See if you can fully understand the stories the data are telling without consulting others.

Developing visuals to uncover patterns, trends and outliers essentially involves creating a number of graphics. This is done simultaneously with other forms of analysis: calculation of statistics, generation of tables, extraction of keywords, etc. The choice of visual will depend on:

- the type of variable (categorical, numerical, nominal, ordinal, etc.)
- the number of variables (univariate, bivariate, multivariate, etc.[120])
- the number of subjects (categories, groups, etc.)
- the number of records (households interviewed, beneficiaries, villages, etc.)
- the analysis (quantitative versus qualitative, composition, comparison, etc.).

It is not always the case that there is *one and only one* visual for an analysis. Different visuals are required for answering different questions, and all have different effects on the user.

**EXAMPLE**

The two tables below were produced to answer this question: What are the sources of staple cereals in 11 different locations? The tables are exactly the same, apart from the technique they use to provide visual comparisons of the different sources of staple cereals. The visual on the left uses horizontal bars and that on the right, choropleth.[121] Both are correct representations of the data; depending on the analyst or the audience, one might appeal or stand out more than the other.

**SOURCE OF STAPLE CEREALS**

% of households in location reporting each source* over 24 hour period

| | Production | Purchase | Gift | Aid |
|---|---|---|---|---|
| Basi Gourma | 12.2 | 66.7 | 7.8 | 15.6 |
| Monzonga | 4.4 | 73.3 | 15.6 | 16.7 |
| Seyna | 21.1 | 74.4 | 11.1 | 13.3 |
| Essouk | 2.6 | 19.2 | 32.1 | 48.7 |
| Kidal | 1.0 | 67.7 | 39.1 | 47.9 |
| Goundam | 8.1 | 94.9 | 3.0 | 3.0 |
| Tilemsi | 5.9 | 100.0 | 2.0 | 0.0 |
| Tonka | 10.8 | 98.0 | 6.8 | 0.7 |
| Koumaria | 22.2 | 92.6 | 13.6 | 0.0 |
| Léré | 17.3 | 94.0 | 28.6 | 3.0 |
| Soumpi | 25.6 | 95.6 | 7.8 | 4.4 |

*Some households may have more than one source

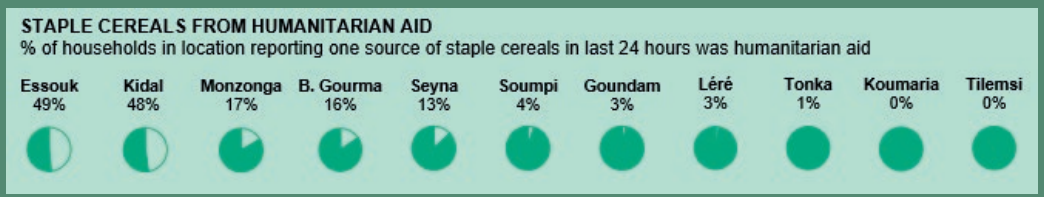
**SOURCE OF STAPLE CEREALS**

% of households in location reporting each source* over 24 hour period

| | Production | Purchase | Gift | Aid |
|---|---|---|---|---|
| Basi Gourma | 12.2 | 66.7 | 7.8 | 15.6 |
| Monzonga | 4.4 | 73.3 | 15.6 | 16.7 |
| Seyna | 21.1 | 74.4 | 11.1 | 13.3 |
| Essouk | 2.6 | 19.2 | 32.1 | 48.7 |
| Kidal | 1.0 | 67.7 | 39.1 | 47.9 |
| Goundam | 8.1 | 94.9 | 3.0 | 3.0 |
| Tilemsi | 5.9 | 100.0 | 2.0 | 0.0 |
| Tonka | 10.8 | 98.0 | 6.8 | 0.7 |
| Koumaria | 22.2 | 92.6 | 13.6 | 0.0 |
| Léré | 17.3 | 94.0 | 28.6 | 3.0 |
| Soumpi | 25.6 | 95.6 | 7.8 | 4.4 |

*Some households may have more than one source

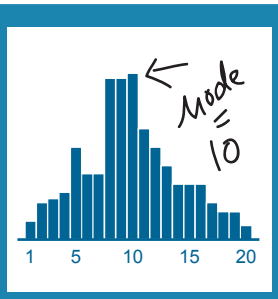
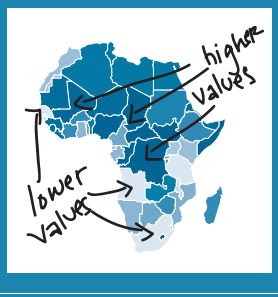
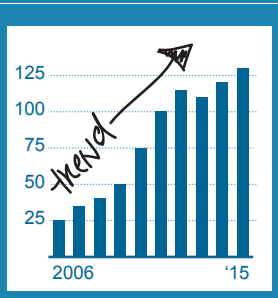
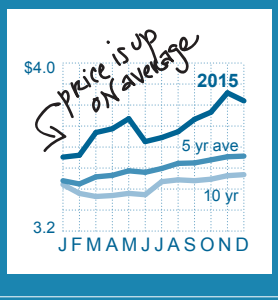
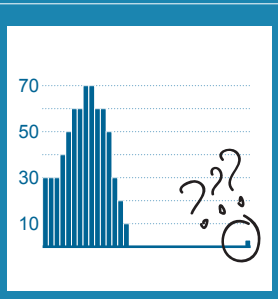
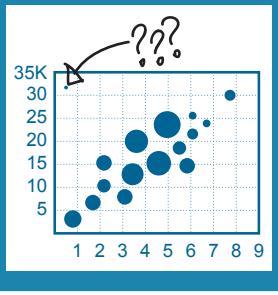
Now say the goal was really to learn what percentage of households used humanitarian aid as their primary source of food. These data can be extracted from the same data as above; but, in this case, a different type of analysis and a different visual – such as the pie charts below – may be more effective in understanding or communicating what the data are trying to tell.




120 Univariate explores one variable at a time, bivariate explores two, multivariate explores more than two, etc.

121 In choropleth visualizations, subjects (geographic locations, categorizations, etc.) are shaded in proportion to the value of the quantitative variable.

The following table provides a list of things to do when exploring data, and a related visual. Annex IV: **Visuals** provides a more detailed list of standard visuals that can be used according to the type of analysis performed (composition, distribution, relationship or comparison).

| | | |
|---|---|---|
| **PATTERNS** | ▪ Explore the data in the variable, including the values/reponses from different subjects<br>▪ Explore the **distribution** of the data, including central tendency, clusters and gaps<br>▪ **Compare** subjects (e.g. categories) and variables with each other and across space |  <br>  |
| **TRENDS** | ▪ **Compare** subjects and variables over time<br>▪ **Compare** values with baselines, indicators with thresholds, etc. |  <br>  |
| **OUTLIERS** | ▪ Throughout the process, **take note of subjects and/or records that are different from the others** and the type and/or value of the difference |  <br>  |

## DESIGNING

Designing visuals for communicating information involves answering three main questions in order:

| | | |
|---|---|---|
| Who is the **AUDIENCE** and what is it that they need to know? | What **INFORMATION** is most effectively communicated through a visual? | What **DESIGN** is the most effective in communicating this piece of information? |

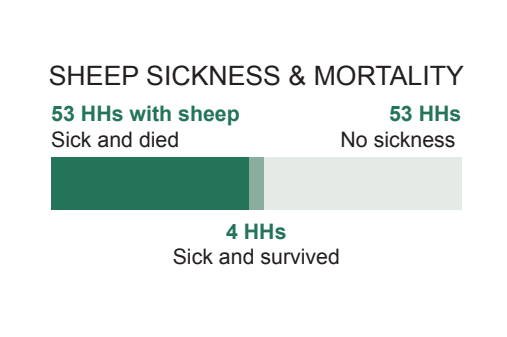
### AUDIENCE

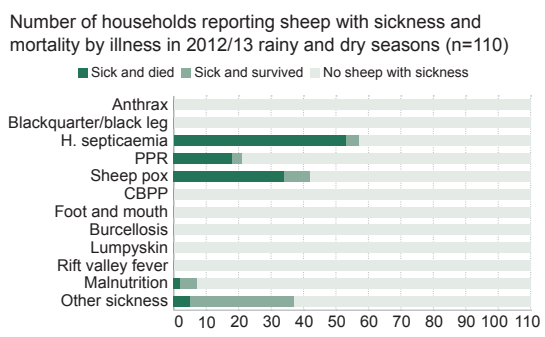
The audience will have three main characteristics that should influence the design of the visual.

- **Their use of the information** – The way the information will be used (in a planning meeting, for continual monitoring, to make decisions, to stay briefed, etc.) should guide the choice of visual form (table versus graphic, detailed map versus regional map, etc.) and the level of detail (aggregated information versus disaggregated, exact numbers versus trends, etc.).
- **Their background understanding of the topic** – As with any other form of reporting, it is critical that the information be communicated to the audience in a language they will understand and that the audience have enough information, but not more than required.
- **Their cultural background** – Different cultures appreciate and respond to different visual and verbal styles. Some appreciate more colour and animation while others may respond better to simplicity. Some may prefer casual or informal language while others might be more at ease with elevated diction.

**EXAMPLE**

Two graphics based on the same data are shown below. The one on the left might be appropriate for an audience that needs to know more about the overall sheep mortality rate (e.g. decision-makers), and the one on the right for an audience that needs to know more about the illnesses that are contributing to sheep mortality (e.g. programme staff).




### INFORMATION

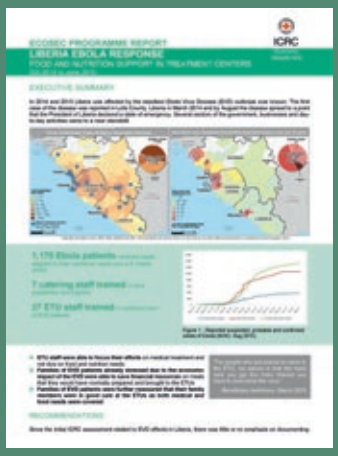
How can one decide what to visualize? Visuals are extremely effective in drawing an audience's eyes to the information and key messages. At the same time, depending on the type of information product where the visual will be featured (report, presentation, video, etc.), too many visuals could be overwhelming and might obscure the key message.

One technique that can be used to identify what to visualize is to create a list of key messages that need to be communicated, and then to identify how each key message could be communicated most effectively. Graphics or complex visuals are good when they are good, but they are not always the most effective way of communicating information. For example, if the key message consists of one statistic (e.g. 40% of the population are food insecure), then communicate it with words and use contrasting fonts (larger font, bold, underline, etc.) to establish visual hierarchy. These words may be supported by a table or small graphic to deepen understanding or contextualization.

**EXAMPLE**
Report with no visuals
→ Risk of key messages being lost in text

**EXAMPLE**
Report with some key messages as visuals and or demarcated text
→ Key messages stand out from main text.

**EXAMPLE**
Report with many visuals
→ Risk of information loss due to visual overload



## DESIGN

For the final visual design, the following must be defined:

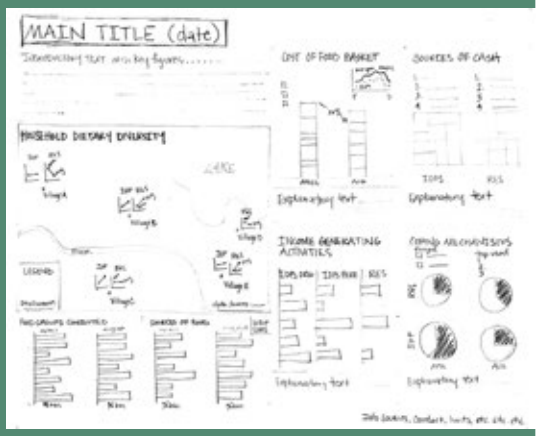
- the medium (static graphic on print, PowerPoint presentation, Web image, interactive Web-based feature, etc.)
- the size (small insert, one-page, half-page, etc.)
- the scope (single graphic or map, information graphic with multiple visuals combined with text, etc.)
- the look (map, bar chart, etc.).

Once these have been determined, proceed to the drawing board. Pick up a pencil and paper and sketch a draft of the visual, paying attention to room for and placement of graphics, text and white space. If it is an interactive design, include functionality.

The most important information should be in the upper left corner, and the less important in the lower right;[122] the key is to tell a story with the data where the reader walks through the story. The goal is that when they leave, they will have learnt something.

**EXAMPLE**

The drawing on the right is an example of a one-page infographic that was drawn on paper before being built in software. It identifies the key topics to address, how to address them (what type of visual) and where each could be placed (making sure that all of them fit on the page). Note that the data are not representative of the real data; that will be the case when the graphs are created electronically. Note also that the drawing may not reflect the final design, because adjustments will most likely be required during the building of the visual. The drawing is, however, an excellent method for getting started and a helpful guide for the analyst during the designing of the visual.



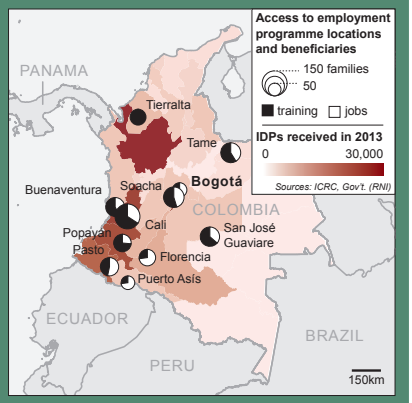

122 This logic is reversed for languages whose scripts are read from right to left.

## BUILDING

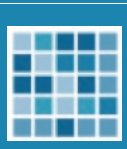
Nowadays, visuals are built mainly on computers, with a broad range of tools and software packages. The choice of tool will depend on the data (e.g. whether the data are interoperable with the tool) and the design (e.g. whether the tool can develop the design). Many different tools will be required sometimes.

**EXAMPLE**

The map on the right was created using three different kinds of software. MS Excel was used to compile, clean and transform tabular data; later, a geographic shapefile in ArcGIS was used to map the distribution of IDPs over administrative units. The result was then exported to Adobe Illustrator, which was used to develop the proportional pie charts on beneficiaries and eventually, to format and build the final design.




The table below lists a number of software packages that can be used to develop the graphics identified in Annex IV: **Visuals**. The list is not exhaustive, and as has already been mentioned, often combinations of these software packages may be required to develop a visual. The tool should not only serve design needs, but must also be interoperable with the database where the data are housed. Analysts should always take into account the amount of work it takes to maintain a visual tool, much of which will be due to updates/corrections to the data and to the master database.[123]

| WHAT YOU WANT TO DEVELOP | SOME SOFTWARE OPTIONS |
|---|---|
| The basics - pie, bar, scatter, line, histogram, box plot, bubble or radar chart  | ■ MS Excel chart tools<br>■ MS Excel conditional formatting<br>■ Adobe Illustrator<br>■ Tableau |
| Matrix with embedded visuals  | ■ MS Excel conditional formatting<br>■ Manually with minimal data in other MS Office tools or Adobe Illustrator<br>■ Tableau |
| Proportional symbols  | ■ MS Excel chart tools (bubble chart) or get creative with bar charts<br>■ Manually[124] with minimal data in other MS Office tools or Adobe Illustrator<br>■ Tableau<br>■ ArcGIS symbology tools (for maps)<br>■ RAW (http://raw.densitydesign.org/) |
| Treemap  | ■ Manually with minimal data in other MS Office tools or Adobe Illustrator<br>■ Tableau<br>■ Google charts<br>■ RAW (http://raw.densitydesign.org/) |

123 See OCHA Graphics Style Book (http://www.unocha.org/about-us/publications/ocha-graphics-style-book), page 27, for information on calculating proportional symbols.

| WHAT YOU WANT TO DEVELOP | SOME SOFTWARE OPTIONS |
|---|---|
| Choropleth  | ■ MS Excel conditional formatting (for matrix)<br>■ MS Excel Power Map (for maps)<br>■ Tableau<br>■ ArcGIS symbology tools (for maps)<br>■ Manually with minimal data in other MS Office tools or Adobe Illustrator |
| Tag cloud   | ■ Tag Crowd (tagcrowd.com)<br>■ Wordle (www.wordle.net)<br>■ ToCloud (www.ToCloud.com)<br>■ Tagul (tagul.com)<br>■ Tagxedo (www.tagxedo.com) |
| Dendrogram  | ■ RAW (http://raw.densitydesign.org/) |
| Vector map  | ■ MS Excel Power Map<br>■ QGIS<br>■ ArcGIS<br>■ GIS Software + Adobe Illustrator<br>■ GoogleEarth |
| Raster map  | ■ ArcGIS<br>■ ERDAS<br>■ Adobe Photopshop |
| Infographic combining charts, maps, images and text  | ■ Relevant software for each visual + Adobe Illustrator for style/layout<br>■ Relevant software for each visual + MS Word/Excel for style/layout<br>■ MS Excel Power View<br>■ Tableau<br>■ ArcGIS Online (storymaps) |
| Interactive graphic  | ■ Tableau<br>■ ArcGIS server (for maps)<br>■ Google charts (https://developers.google.com/chart/)<br>■ Data-Driven Documents (http://d3js.org/) |
| Venn diagram  | ■ MS Office drawing tools<br>■ Adobe Illustrator |
| Network diagram  | ■ MS Office drawing tools<br>■ MS Visio<br>■ Adobe Illustrator |

All of the software packages mentioned above come with a great deal of support documentation on their use; there is also a large community whose members follow the use of these packages and share their experiences. For support in this regard, work with colleagues who may be familiar with the software, contact the EcoSec information management specialist or Google people who are sharing their experiences.

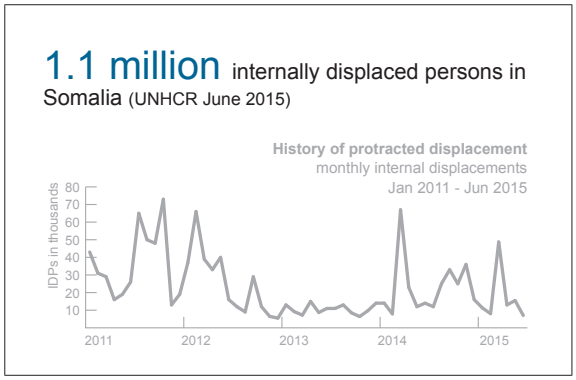
## VISUAL DESIGN PRINCIPLES

The goal in developing a visual for humanitarian planning and programming purposes is to illuminate key message(s) in a credible manner to aid understanding and/or decision-making. Visuals in this context are most effective when they are accurate, clear and concise. Accuracy is achieved by using sound data sets and selecting the appropriate graphics to visualize them (see Annex IV: **Visuals**). Developing clear and concise visuals involves paying attention to the content (what is actually included) and the style (colours, size, shape, etc.).

The following section highlights six key principles with examples of practice and graphics. The data in the visuals were taken from the Afghanistan Geodesy and Cartography Head Office, OCHA Afghanistan, OCHA Somalia, the UNHCR and WFP Yemen. Links to those data sets and/or reports are available in the list of references at the end of the chapter.

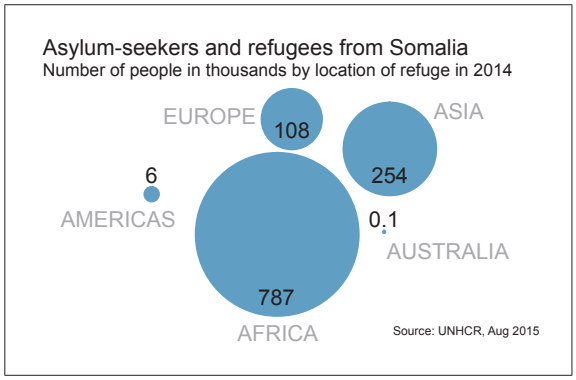
## ESTABLISH A VISUAL HIERARCHY

- Use contrast to emphasize the most important data and de-emphasize the less important or "background" data.
- Exclude details and images that do not serve a purpose in portraying the key message(s).
- Carefully plan the visual layout to ensure that key messages are 'read' first.



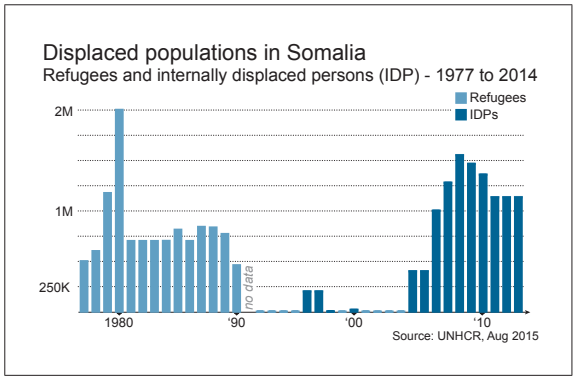

## BALANCE THE LEVEL OF DETAIL

- Communicate numbers as simply as possible (e.g. 1,100,000 can be written as '1.1 million').
- Use graphic visuals (gridlines, scale bars, etc.) as a communication tool, not as decoration.
- Use chart labels if the audience needs them (e.g. Do I need to show the trend or the exact number or both?)



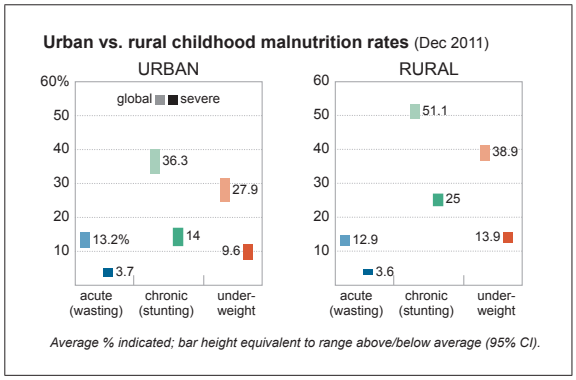

## ESTABLISH HARMONY

- Include every interval (every month, year, etc.) between intervals plotted to ensure flow in the data.
- Use different colours when they correspond to different meanings in the data. Go for monochrome as a default.
- Develop a consistent style using templates, standard colours, symbols, etc. for ease of reading.



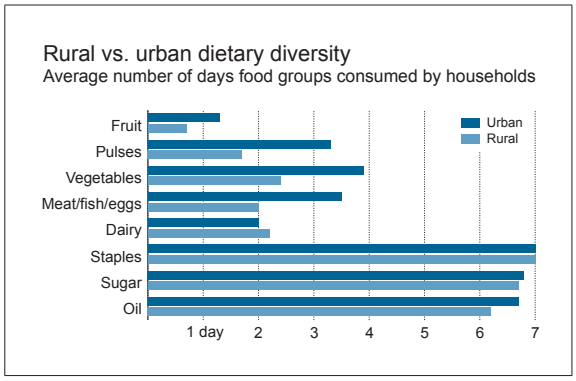

## BE ACCURATE

- Double check the data yourself, then share it with a colleague (or two) for their reactions before publishing.
- Use three-dimensional graphics only to illustrate three-dimensional data.
- Provide information on what the visual reflects (sample size, confidence interval, etc.).



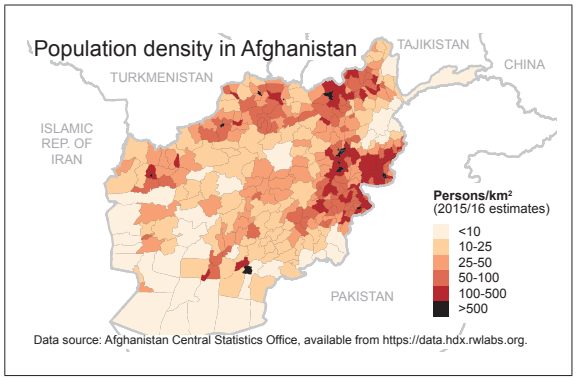

## ENSURE READABILITY

- Ensure that fonts are readable in size (>6 point), colour (dark colours) and shape (resize with same proportions to not stretch fonts).
- Print colour graphics in black and white to ensure that key messages portrayed in colour are not lost.
- Avoid rotating text unless absolutely necessary.




## GIVE CREDIT

- If the data used are not the primary source in the report, share the source, the date of the data and the scope of what it reflects.
- Mention any limitations of the analysis or possible areas for misinterpretation.
- Spell out all the acronyms you use.



# REFERENCES

Afghanistan Geodesy and Cartography Head Office (AGCHO), Afghanistan administrative boundaries – geographic shapefiles, 25 September 2012. Available at: https://www.humanitarianresponse.info/en/applications/data/datasets/locations/afghanistan.

OCHA Afghanistan, *Estimated Population of Afghanistan 2015/2016*. Available at: https://data.hdx.rwlabs.org/dataset/estimated-population-of-afghanistan-2015-2016.

OCHA Somalia, *Somalia Humanitarian Dashboard – June 2015*, 30 July 2015. Available at: http://reliefweb.int/sites/reliefweb.int/files/resources/somalia_humanitarian_dashboard_-_june_2015.pdf.

UNHCR, "UNHCR Population Statistics Database", online resource, accessed in August 2015. Available at: http://popstats.unhcr.org/

WFP, *The State of Food Security and Nutrition in Yemen*, 2012. Available at: http://documents.wfp.org/stellent/groups/public/documents/ena/wfp247833.pdf.

# ANNEX I
# TERMS AND DEFINITIONS

The following is an overview of the terms and definitions used throughout the guide. The definitions are within the context of this guide.

### A-E

**Analysis plan** – An outline that sets out, for a specific exercise, the data that have to be collected for a specific exercise, and from where and/or whom; it also describes how the data need to be combined to enable the drawing of conclusions, and the types of analysis to be used.

**Analytical approach** – The overarching way of dealing with or accomplishing the data collection or analysis.

**Association** – In analysis, this refers to the general relationship between two variables.

**Assumptions** – Anything that is accepted, without proof, as true or sure to happen.

**Baseline** – Defined point (measure) that is regarded as the 'base' of comparison. Often the measure before, during normal times, on average, etc.

**Big data** – A general term used to describe data sets from traditional and digital sources that are too large and complex to process and analyse with standard data processing and database management applications.

**Cascading selects** – A feature in electronic data collection/entry that enables researchers or analysts to alter the optional answers to one question based on the response to another.

**Categorization** – The process of grouping data into categories according to specified criteria, which usually means that these data have something in common, such as a similar value (in the case of quantitative data) or a similar feature (in the case of qualitative data).

**Category** – Either text or numbers, but limited to a certain range of specified options, or categories. They are, by definition, discrete. Databases often refer to category lists as "domains".

**Causation** – A measure of the degree to which one or more variables can cause (or predict) the value of another or of others.

**Central tendency** – A measure of central tendency is a central or typical value for a distribution of data.

**Criteria** – Principles or standards against which something may be judged or decided.

**Coding** – A technique use to classify qualitative data into classes – a word or short phrase – in order to discover and reduce data.

**Coefficient of determination** – A statistic used to test how much of the variation in one variable can be explained by its' relationship with the other variable.

**Conceptualization** – **The process of assigning meaning to the concepts stated in the objectives of the study.**

**Consent** – Any freely given, specific and informed indication by which a data subject signals agreement to the processing of personal data relating to him or her.

**Context** – Information that locates data and analysis in a place and a time.

**Comparative analogy** – Comparison of a unit of measurement with something that the audience can more easily understand.

**Comparative statistics** – Statistics that compare two or more subjects, processes or phenomena. Comparative statistics can be descriptive and/or draw inferences.

**Comparison** – In the context of this guide, refers to the act of comparing an indicator or series of indicators between two or more entities (people, households, population groups, animals, institutions, etc.) and/or dimensions (geographic location, time, etc.) in order to learn how they differ according to that entity or dimension.

**Conceptual framework** – A tool for analysts to explain real-world concepts such as processes or phenomena in an abstract or generalized manner.

**Confidence interval** – The range within which a certain percentage of responses or cases would fall. It is used to gauge the reliability of an estimate.

**Control variable** – A variable that may have an influence on other variables, but is not the focus of the study.

**Converging evidence** – Consists of individual pieces of information that do not by themselves support a conclusion, but when combined, constitute a robust body of evidence in support of the conclusion.

**Correlation** – A measure of association: it measures the strength and the direction of the relationship between two variables.

**Corroborating evidence** – Is made up of several pieces of evidence (data or analysis) that support a conclusion.

**Cross-sectional study** – Analyses data on a variable for one given period of time.

**Cross-tabulation** – A table in a matrix format that breaks down responses by discrete factors or categories, and counts the frequency of each in comparison with other variables, categories, factors, etc.

**Crowdseeding** – A method of crowdsourcing in which providers of information are pre-identified and often trained in gathering and sharing information; this makes it possible to have more control over information providers than in crowdsourcing.

**Crowdsourcing** – A method of gathering data and/or information by soliciting contributions from a large group of people; it may be confined to a specific community (eyewitnesses or victims during a crisis, humanitarian workers, technical specialists, etc.) or open to the general public.

**Data** – Raw, unorganized facts or figures that have to be processed and analysed.

**Data protection** – The collective term for the set of basic principles, rights of data subjects, data controllers' obligations (including data security and data integrity) and enforcement measures required to prevent data loss, misuse of data or the breaching of personal rights to data protection and privacy.

**Data security** – The technological and organizational measures required to provide adequate protection for data from any risks to which they may be exposed.

**Data integrity** – Refers to maintaining and ensuring the accuracy and consistency of data over their entire life cycle.

**Database** – An organized collection of data. Databases can be either non-relational or relational.

**Demographic data** – Data on a given population: population statistics, gender, age, ethnicity, etc. Governments can collect demographic data during a census and share them with others through their statistics offices.

**Dependent variable** – A variable that is influenced or 'depends' on another measurable variable. It can be referred to as a 'response' variable.

**Derived variable** – A variable that is produced from another set of variables.

**Descriptive statistics** – Statistics that provide a simplified summary of characteristics – average values, maximums and minimums, proportions, etc. – to help describe data.

**Design effect (DEFF)** – The ratio of variance between the sampled values and actual values due to the natural heterogeneity of data. It is essentially the factor by which the size of a complex sample, such as a cluster sample, would have to be increased in order to produce survey estimates with the same precision as a simple random sample.

**Diagram** – A symbolic representation of information.

**Disaggregation** – Disaggregation in data collection and analysis involves breaking up a data set into two or more different components.

**Discrete quantitative data** – Data in the form of whole numbers: the number of people in a household, the number of visits to a clinic, herd size, etc. Discrete data cannot be a fraction.

**Distribution** – In quantitative analysis, it is the process of arranging all the values of a variable to find the frequency with which they occur.

**Evidence** – Information that helps to demonstrate the truth or falsehood of a given hypothesis or proposition.

**Exploratory data analysis** – It summarizes the main characteristics in the data that may not be implicit in the hypothesis or theories in place before data are collected. The objective of exploratory analysis is to guide the analyst in elaborating hypotheses, theories and other assumptions, which can then be tested either by using the same data or by collecting fresh data and/or information.

### F-J

**Forecasting** – Predicting what could happen, based on past and present evidence.

**Framework** – A matrix of elements based on relevant assumptions, theories, variables and indicators and/or criteria.

**Frequency** – The count of records that have a given value or a value within a specified range (referred to as a group or class).

**Generalization** – The process of making data less detailed.

**GIS data** – Geographically referenced data that can be used in GIS for spatial analysis. GIS integrates hardware, software and data to collect, manage, analyse and display all forms of geographically referenced information.

**Geospatial or geographic data** – Data with a geographic or spatial component.

**Histogram** – A graphical representation of the distribution of data: the y-axis shows the frequency and the x-axis the values; the histogram may be accompanied by a table that shows the number of records for each range of values.

**Interpretation** – The process of determining what an analysis means through contextualization, use of experience, and selection of the most important findings – in order to draw conclusions.

**Impact indicators** – Look at the long-term impact of a programme on individual beneficiaries or communities.

**Independent variable** – A variable that is not influenced by other measurable variables.

**Indicator** – A variable that indicates something, such as a change or a trend. Indicators are compiled from data, and measured and interpreted through comparison with standard or context-specific baselines, thresholds or target values.

**Inferential statistics** – Investigates models and hypotheses to make predictions or draw inferences about a population based on observations taken from a sample, or to test the probability of observed differences being true or false (or something in between), with a quantifiable level of confidence, precision and significance.

### K-O

**Linear correlation** – Explores the association between two normally distributed numerical variables thought to have a linear relationship.

**Logical framework** – A framework that demonstrates the causal relationship between the elements that are to be measured.

**Longitudinal study** – Studies the same variable at repeated intervals over time.

**Matrix** – A tool that can be used to look at the intersection of two constants, variables or processes.

**Memoing** – A process for recording data collectors' observations and thoughts as they evolve over the course of the study.

**Metadata** – Data about data; they include but are not limited to information on the source of the data, the date the data were collected and copyright information.

**Methods** – The techniques for collecting and analysing data.

**Mixed-method approach to analysis** – Analytical approach that makes use of both quantitative and qualitative analytical methods.

**Multi-group study** – Examines several groups, with the objective, possibly, of comparing them (inter-group comparison) and doing intra-group analyses of each. When the same variables are measured for each group, using the same methods, this is often called a 'paired design'.

**Narrative research** – A type of qualitative research that uses life-stories collected through journals, interviews, transcripts, etc. as the unit of analysis.

**Non-probability sampling** – Different from probability sampling in that it does not use random selection throughout the process (however, it could at certain stages); every individual does not, therefore, have an equal chance of being selected for the sample.

**Non-relational database** – Two-dimensional arrays of data (normally in the form of a single table) that can be developed with most data-manipulation tools like MS Excel, or statistical software packages such as Statistical Analysis System (SAS), Statistical Package for the Social Sciences (SPSS), etc.

**Objective** – A statement that outlines the results expected from an exercise; it identifies, clearly, what needs to be understood.

**One-group study** – A study that examines one group individually, with the objective of conducting an intra-group analysis.

**Operationalization** – The process after conceptualization, during which broad concepts are made measurable.

**Outcome indicators** – Used to describe the short to medium-term effects of programme activities on beneficiaries' lives.

**Outlier** – An observation that is at an abnormal distance from all other measures.

P-T

**Participatory approach** – An approach in which responsibility – for decision-making and for defining, collecting and analysing pertinent data – is shared by the team leading the exercise, the participant(s) or community involved and any other parties interested. The team leading the exercise will exert some degree of authority, and play the role of mediator, but opinions will be shared openly.

**Peer-review** – Process of reviewing data, analysis and/or reporting with relevant peers.

**P-codes** – The short name for 'place code': a code name given to a unique geographic feature, such as a populated place (village, town, city, etc.) or an administrative unit (province, prefecture, state, commune, etc.).

**Plausibility** – The appearance of having truth or of being credible.

**Population of interest** – The group of people on whom one would like to focus the study – and from whom, potentially, generalize its findings. In sampling, they are the people from whom a sample is drawn via a sample frame.

**Primary data** – Data collected by the person or persons who will make use of them, such as data collected during a survey or an experiment, or by witnessing something.

**Probability sampling** – Any method of sampling that uses some form of random selection in which every individual has an equal chance (probability) of being selected for the sample; and in which the selection of one individual is independent of the selection of another.

**Process diagram** – A type of diagram that maps processes (events, decisions, activities, etc.) and their outcomes.

**Process indicators** – Measure the implementation (the process and the results) of programme activities.

**Proportion** – A number that describes the representation of a key characteristic or value in relation to everything else.

**Qualitative approach to analysis** – Approach to analysis that aims to explore and understand phenomena, and is based on the collection of data in their natural setting through observation and discussion.

**Qualitative data** – Data that are descriptive.

**Quantitative approach to analysis** – An approach to analysis that seeks to confirm specific set hypothesesis by quantifying data and information. It is usually based on fixed, structured formal surveys or measurements of specific variables (food price, body weight, etc.).

**Quantitative data** – Numerical data expressed as statistics, rates, proportions, etc. Quantitative data can be further classified into continuous and discrete data.

**Range** – The difference between the highest and lowest values of a variable.

**Ranking** – A technique used to order data from smallest to largest, or vice versa. It is useful for putting records in context, i.e. for showing where one record falls in relation to another.

**Regression analysis** – A type of analysis that tries to estimate how one dependent (response) variable is influenced by one or many independent (explanatory) variable(s). It can be used to develop models on causal relationships, as part of prediction analysis (such as forecasting), in inferential statistics or to test hypotheses.

**Relational database** – Databases housing a collection of tables and elements, each having some relation to the other through a constant, a variable or a set of criteria.

**Relationship** – In analysis, the correspondence, connection, or link between two or more variables of interest.

**Remote sensing** – A technique used to acquire information from a distance. Remotely sensed data are normally extracted from satellite images (collected from sensors or cameras on satellites), aerial images (collected from manned or unmanned aircraft) or on-the-ground images, which are then processed, classified, analysed and interpreted.

**Results monitoring framework** – A type of logical framework used in monitoring and evaluation exercises that outlines the expected results, so that they can serve as a guide for monitoring (process, outcome and/or impact indicators).

**Results monitoring** – Analysis that normally uses a framework (the results monitoring framework, also called the 'logical framework') that is set up to describe the expected and actual changes after a project or programme. The framework makes it possible to measure progress in terms of results-based indicators.

**Risk analysis** – A type of analysis that looks into the likelihood of an adverse impact in the event of a shock (or threat). It establishes a framework and specific criteria to measure the risk of a particular shock to a certain population, group, geographic region, topographic feature, etc. .As such, it is a projection of possibility or possibilities.

**Sample** – That group of people, households, institutions, etc. which is selected from the sample frame for interviewing or studying. There are two main sampling methods: probability and non-probability.

**Sample frame** – A list of potential sampling units (people, households, institutions, etc.). It is, in fact, an exhaustive list of all the sampling units that have a chance or probability of selection for the sample.

**Sampling bias** – An error that can occur when certain groups or individuals within a population of interest do not have an equal chance of being selected for the sample.

**Sampling distribution** – The distribution of an infinite number of samples of the same size as the sample in the study.

**Sampling units** – Units used to draw the sample: households, individuals, children, etc.

**Scale** – This is a technique used to give social data a specific order (e.g. better, same, worse). Qualitative scales are in effect a type of category where categories have a specific order and a specific relationship to one another.

**Scatter plot** – A simple graphic that is used to explore the relationship between two numerical variables.

**Scenario-building analysis** – Analysis that involves drawing up plausible scenarios – normally two or more – based on past and present evidence.

**Secondary data** – Data collected by someone other than the user of the data: census data, data from national or other international organizations, historical accounts, media reports, etc.

**Severity analysis** – Analysis that establishes a framework and specific criteria to measure the severity of a situation.

**Situation analysis** – Analysis that examines the internal and external, and the direct and indirect, factors that may have some bearing on a situation. It aims to understand the relationship between these various factors, the circumstances that led to the current situation and prospects for the future.

**Skip logic** – A survey feature where the response to one question will determine what the next question will be: for example, whether it will be a follow-up question or whether the data-collector will skip to another set of questions.

**Standard deviation** – The spread of values around the estimate in a single sample, or the variation of the sample values. Denoted by the symbol σ.

**Standard error** – The standard deviation of the sample estimate, and is expressed in percentage points (e.g. +/- 5% or +/- 0.05).

**Stratification** – The process of dividing the population of interest into non-overlapping sub-groups (called strata) that share certain characteristics of interest or pertinence to the objectives and the potential outcomes of the study.

**Tag cloud** – A visual representation of text data: here, each piece of text is weighted with its frequency in a data set.

**Theories** – Based on evidence or observation, and can be used to make predictions on the value of a variable, and on relationships, outcomes, trends, etc.

**Threshold** – Predetermined points (measures) that must be crossed to begin producing a specific effect or to elicit a response. The threshold must be a constant.

**Tools** – Instruments that assist in data collection or analysis: questionnaires, mobile phones, etc.

**Trend analysis** – Type of analysis that studies past and current data on a given indicator to try and spot trends, including anomalies (irregular events or tendencies) and patterns (repeated events or tendencies). The aim is to acquire a better understanding of the relationship between two or more entities or dimensions.

**Triangulation** – The process of combining or comparing several sources and/or observations on a given topic, with the aim of increasing confidence in the result by decreasing the bias associated with 'one side of the story'.

**Typology** – A type of categorization that puts subjects into groups, based on certain traits.

### U-Z

**Validity** – The extent to which a concept, conclusion or measurement is well founded and corresponds to the real world.

**Variance** – A measure to quantify how much the responses in a given variable are different from one another.

**Variable** – Any piece of data that can take on different values, and is subject to change. It is the opposite of a constant, or a piece of data that does not change. A variable can be a piece of quantitative or qualitative data.

**Vulnerability analysis** – Vulnerability is the degree to which a person affected lets a given event cause harm. It is a person or population's sensitivity multiplied by their lack of ability to cope or adapt. A vulnerability analysis normally starts by defining vulnerability in a given context (e.g. vulnerable to what?) and then identifies specific criteria (called domains).

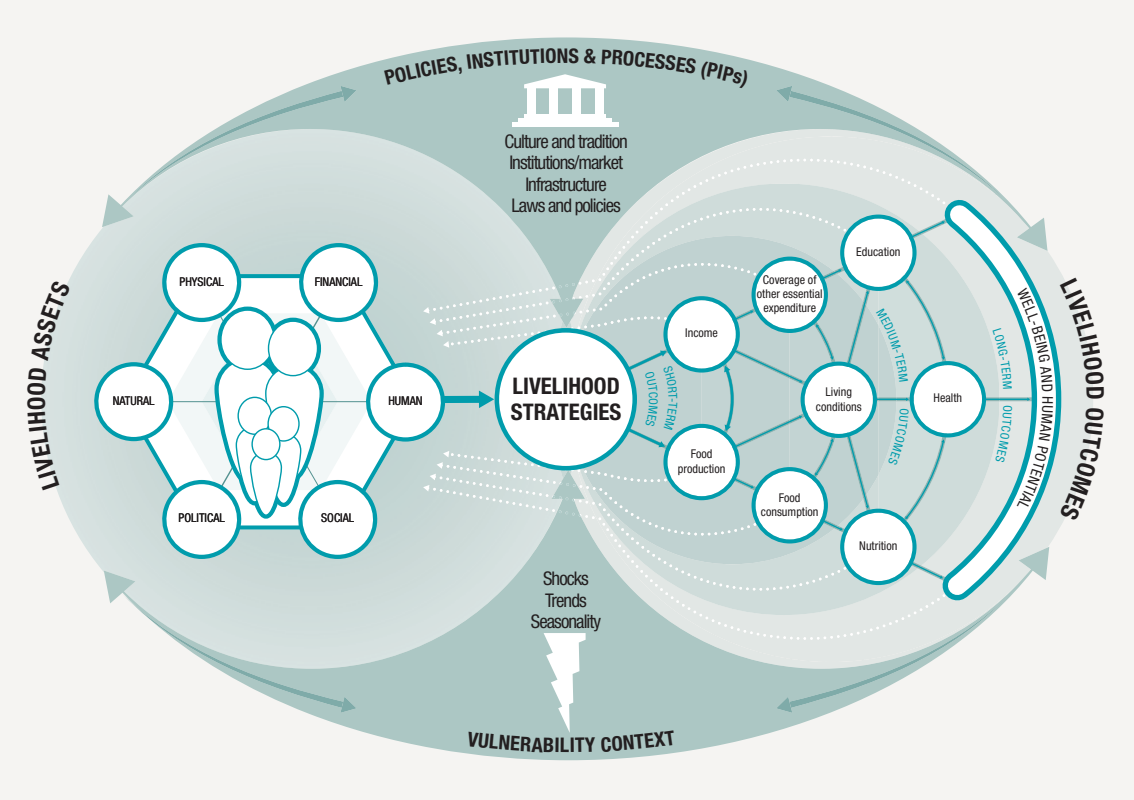
# ANNEX II
# ANALYSIS DESIGN TOOLS

SOME OF THE ANALYSIS DESIGN TOOLS USED IN ECONOMIC SECURITY ANALYSIS ARE DESCRIBED BELOW.

# CONCEPTUAL FRAMEWORKS

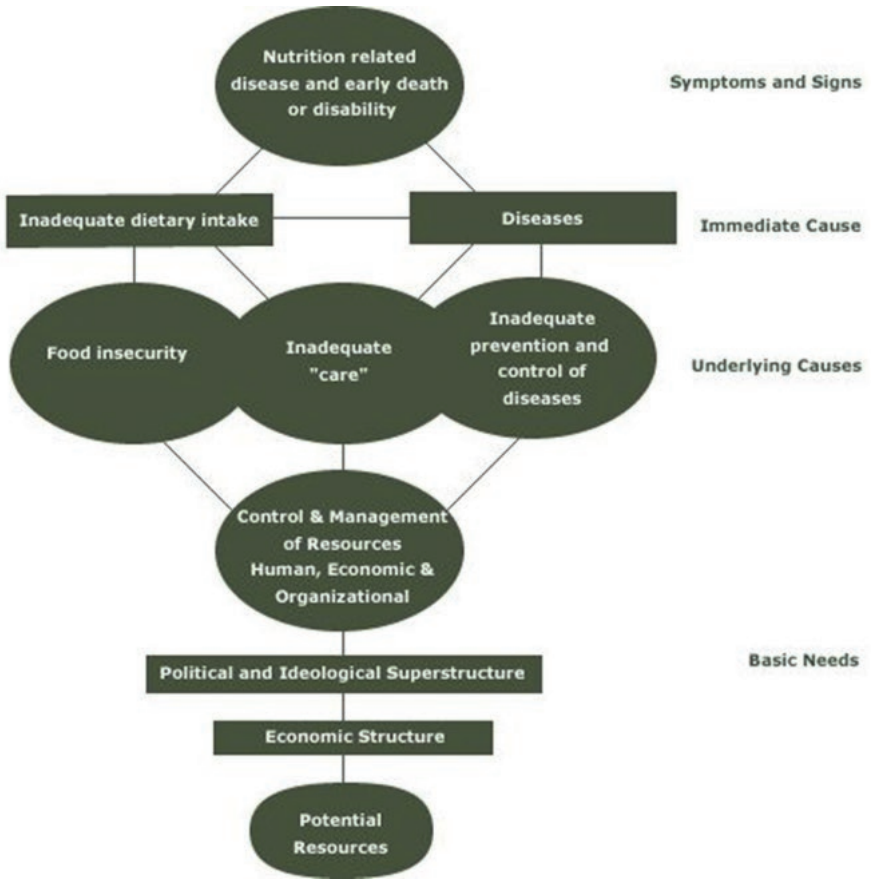
## ECONOMIC SECURITY CONCEPTUAL FRAMEWORK

The Economic Security Conceptual Framework (adopted by EcoSec and adapted from the DFID Sustainable Livelihoods Framework, 1999) describes, in a simplified way, the interaction between livelihood assets, strategies and outcomes, and how there are affected by and influence policies, institutions and processes (PIPs) and the 'vulnerability context'. The framework can be used as a tool for assessing economic security. The full framework is available at the EcoSec Resource Centre, and a more detailed description of it can be found in the EcoSec handbook *Assessing Economic Security* (ICRC, 2016).




## CONCEPTUAL FRAMEWORK OF THE CAUSES OF MALNUTRITION

In 1990, UNICEF developed a conceptual framework for the causes of malnutrition. The framework examines the various causes of malnutrition, including those related to health and dietary intake, and caring practices, living conditions, the local environment and policies, institutions and processes. The framework can be used as a tool for assessing malnutrition. A more detailed description of the framework is available at the EcoSec Resource Centre, on the "Nutrition" page.




## LOGICAL FRAMEWORKS

### ECOSEC REFERENCE FRAMEWORK

The EcoSec reference framework for the civilian population is a summary table that presents, logically and concisely, the objectives (both general and specific objectives) and selected outcome indicators for an ICRC (sub-) programme/core activity. It highlights the intended results (short- and medium-term outcomes) of the (sub-) programme for one or more target populations, and proposes indicators to monitor and evaluate progress towards the achievement of these results. The reference framework also contains information about the main issues and potential causes to which the ICRC is responding through these (sub-) programmes, and an overview of the main activities.

The framework can be used as a guide for developing a results framework for country programmes (such as those developed in the Planning for Response process). Its components, such as short-term outcome indicators and data sources, can also be used to develop analysis plans or frameworks for data collection and analytical exercises.

The latest version of the reference framework is available at the *EcoSec Resource Centre* on the "Results-Based Monitoring" *page.*

## ANALYSIS PLAN TEMPLATE

The following table can serve as a guide for developing an analysis plan for a data-collection and analysis exercise. Chapter 3: Analysis design provides guidance on the content in an analysis plan.

| INFORMATION NEEDS | CONTEXTUAL INFORMATION | INDICATORS AND THRESHOLDS | DATA REQUIRED | DATA SOURCES AND METHODS | ANALYSIS TYPE |
|---|---|---|---|---|---|
| **Objective 1:** | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| **Objective 2:** | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| **Objective 3:** | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |

# ANNEX III
# ADDITIONAL SAMPLING FORMULAS

## BASIC FORMULA FOR MEANS OR TOTALS

For quantitative analysis of continuous variables, where the mean or the total is the focus of measure (average household income level, mean daily per capita calorie consumption, etc.), an alternative to the basic formula for proportions can be used to provide an optimal sample size. This formula uses the standard deviation, the measure of variance of the variable of measure (not the prevalence) and the degree of variance within the population of interest. The following formula is used for populations whose size is unknown:

$$n = \frac{Z^2 \times \sigma^2}{e^2}$$

where:

- n = required sample size
- Z = score associated with desired confidence level (90 to 95% is most commonly accepted; see the table in Chapter 5: Sampling chapter for scores)
- σ = standard deviation of the variable (can be taken from a previous study or through a pilot study)
- e = precision or margin of error (usually 0.05 or 0.10)

## FORMULA FOR COMPARISON SURVEYS USING MEANS OR TOTALS

A formula for variables using means or totals can be used instead of the comparison formula for variables expressed as proportions: for instance, the following formula from FANTA:[124]

$$n = \frac{(Z\alpha+Z\beta)2 *(sd12 + sd22)}{(X2-X1)2}$$

where:

- n = required sample size for each round or group
- $Z_\alpha$ = z-score corresponding to the degree of confidence desired to be able to conclude that an observed difference ($P_2$-$P_1$) would not have occurred by chance (α is statistical significance)
- $Z_\beta$ = z-score corresponding to the degree of confidence desired to be certain of detecting a difference ($P_2$-$P_1$) if one actually occurred (β is statistical power)
- $sd_1$ = *expected* standard deviation for the indicator for a particular survey round or comparison group 1
- $sd_2$ = *expected* standard deviation for the indicator for a particular survey round or comparison group 2
- $x_1$ = *estimated* level of an indicator at the time of the first survey or for the main group/control area
- $x_2$ = *expected* level of the indicator, either at some future date or for the comparison group
- ($x_2$-$x_1$) = the magnitude of the change or comparison-group differences it is desired to be able to detect

124 Robert Magnani, *FANTA: Sampling Guide*, December 1999. Available at: http://www.fantaproject.org/monitoring-and-evaluation/sampling.

# ANNEX IV
# ANALYSIS AND VISUALS

# WHAT DO YOU NEED TO VISUALIZE?

## SOME EXAMPLES OF WHAT YOU COULD USE

### COMPOSITION

Snapshot of a single variable, normally descriptive statistics.

**PIE**

Share of total few subjects

**TREE MAP**

Share of total many or few subjects

**STACK**

Share of total few or many subjects

**STACK & SUB-STACK**

Share of a share

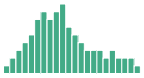
### DISTRIBUTION

Frequency, variance, location, etc. of a single variable.

**SCATTER**

Two variables

**BAR HISTOGRAM**



Single variable with few records*

**LINE HISTOGRAM**

Single variable with many records

**MAP**

Distribution over space

### RELATIONSHIP

Correspondence or connection between two or more differe variables.

**MATRIX**

Between two categorical variables

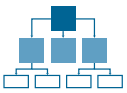
**SCATTER**

Between two numeric variables

**BUBBLE**

Between three numeric variables

**NETWORK**



Normally qualitative or mixed data

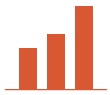
### COMPARISON

The measure of the same variable for different subjects or points in time.

**PROP. SYMBOLS**

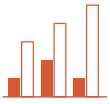
One variable few subjects

**VERTICAL BAR**



One variable, few subjects

**CLUSTERED VERT. BAR**



Many variables, few subjects

**VENN**

Normally qualitative or mixed data

**PROP. SYMBOLS**

One variable many subjects

**HORIZONTAL BAR**

One variable many subjects

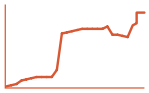
**CLUSTERED HORIZ. BAR**

Many variables, many subjects

**RADAR**

Multiple variables, one or few subjects

**LINE**

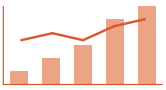


Over time continuous start to now

**STACKED LINE**

Comparison of comparison over time

**BAR & LINE**



Compare comparisons with trend or baseline

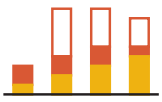
**TAG CLOUD**


Compare counts, qualitative data

### COMBINED ANALYSES

A combination of different types of quantitative analysis displayed on the same graphic. Normally two, possibly three in one graphic as more than that may become crowded and challenging to read.

**STACKED BAR**

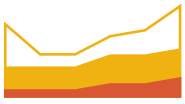


Compare shares & totals

**PROP. SYMBOLS & PIE**

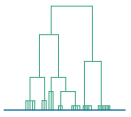
Compare share & totals

**STACKED AREA**



Comparison of composition over time

**DENDROGRAM**



Clustering of & relationship b/w subjects

**MAP & PROP. SYMBOLS**

Compare values over space

**CHOROPLETH MAP**

Compare values over space

**CHOROPLETH MATRIX**

Compare values b/w categorical variables

**BOX PLOT**

Compare range of values of one variable between subjects

***Variable*** is any piece of data that are subject to change, or vary. ***Subject*** in the scope of this graphic may refer to either categories, groups or records.

**MISSION**

The International Committee of the Red Cross (ICRC) is an impartial, neutral and independent organization whose exclusively humanitarian mission is to protect the lives and dignity of victims of armed conflict and other situations of violence and to provide them with assistance. The ICRC also endeavours to prevent suffering by promoting and strengthening humanitarian law and universal humanitarian principles. Established in 1863, the ICRC is at the origin of the Geneva Conventions and the International Red Cross and Red Crescent Movement. It directs and coordinates the international activities conducted by the Movement in armed conflicts and other situations of violence.


ICRC

4276/002 05.2017