March 2013

TECHNICAL BRIEF

How to approach a dataset

Part 2: Analysis

Aldo Benini

A note for ACAPS

How to approach a dataset -

Part 2: Analysis

Version 19 March 2013

Table of Contents

Figures

Acknowledgement

This note was written with the very active support of Sandie Walton-Ellery, ACAPS, Benoit Munsch, Solidarités International, both in Dhaka, Bangladesh as well as Emese Csete, ACAPS Nairobi, and Patrice Chataigner, ACAPS Geneva. Walton-Ellery and Munsch supplied the raw dataset, interpretation of select findings, and the description of the 2011 assessment context that the reader may find in the main body of this note. Csete and Chataigner helped with the alignment between the text of this note and the flow of the demonstration workbook.

Acronyms and expressions

ACAPS	Assessment Capacities Project
GIS	Geographic Information System
MCA	Multiple correspondence analysis
R	A statistical program
SPSS	A statistical program
STATA	A statistical program
VBA	Visual Basic of Applications

1. Summary

This is the second of two notes on *how to approach a dataset* that emerges from a needs assessment following a disaster. It addresses typical challenges of data analysis. The companion note that precedes it looks at the data preparation phase.

The note speaks to typical analysis forms that assessment analysts may consider when data collection and entry were done without the benefit of a pre-existing analysis plan. It demonstrates the analysis forms with data from a real assessment in Bangladesh.

The note is developed on two lines. A first, more technical part demonstrates the processing of typical data constellations, specifically:

- Overview by key grouping variables
- Summary of all binary variables
- Statistics of preferences measured ordinally
- Tabulation of multiple variables with similar or overlapping category sets
- Multiple-response data tabulation
- Transforming text variables into sets of indicator or ranking variables

The second, more substantively motivated part works through three analysis steps for which the Bangladesh data offers the right kinds of variables, and which hold a generic interest:

- Rapid overview of sites in terms of severity
- Variations of needs as a function of severity
- An index of distress, formed of focus group information, as an alternative to the severity ratings by assessment teams.

We also show how to compare priorities between different sources, such as from separate interviews with men and women. The degree of agreement can be interpreted as a measure of data reliability or of convergent interests and perceptions.

The accompanying Excel workbook demonstrates these analysis types. In addition, for rare situations when multiple response and priority data are entered in text form, it provides formulae to transform the data "from text to numbers".

2. Introduction

2.1 What this is about

This note speaks to data analysts in needs assessment teams in post-disaster situations. It provides guidance for how to analyse some typical data constellations.

The companion note describes the preceding data preparation; its accompanying workbook demonstrates steps to take the demonstration dataset to a shape efficient for analysis work. This note and its associated workbook take it up from that end point.

For background on the dataset, which is an extract from a Bangladesh assessment, refer to the data preparation note. In a generic perspective, we assume:

- The assessment was designed with a common-sense view to what would likely prove important to know. *An analysis plan was not drawn up prior to data collection.*

- Affected communities were selected based on a purposive sample. The purpose was to observe a diversity of situations and needs in severely and less severely affected communities. The sample communities are diverse with regards to some of the factors that likely determine the needs of the affected population. *Ordinary survey estimation with probability weights is not feasible, but descriptive statistics are needed.*

- In assessment datasets, various data types occur. There may be text variables, numeric variables, response sets to multiple-choice questions. Numeric variables may differ in levels of measurement: nominal, ordinal, interval or ratio. *The Bangladesh data contain numeric as well as categorical text variables. We also demonstrate, with simulated data, how to extract indicators from text fields.*

- Ordinal variables are used to cover two data situations: interval estimates of counts or intensities, and preference or priority measures for problems, needs, or types of relief. *We treat the first the same way as categorical variables, the second, if they express mutually exclusive priorities, by a procedure known as Borda counts.*

Our goal is to provide a rational and time-saving approach to confronting a new dataset as it reaches the analyst after data preparation. Our objective, in dealing with the Bangladesh data, is to demonstrate, with the help of a small number of MS Excel features, how to compute the statistical tables needed in analysis and reporting.

This note details the *what, how, and why* chiefly by data type. It also demonstrates some analysis forms that the specific Bangladesh data suggests, but it is obvious that these do not generalize to all assessment situations. Some technical matter is discussed in sidebars, which the hurried reader may initially skip, to return there at leisure.

[Sidebar:] Is there an analysis plan or not?

Analysis progresses in phases. As data collection, field editing and data entry move in parallel, and latest when the data preparation work has been completed, analysis work may begin. In practice, some members of assessment teams may be busy with analysis work already while some of the logically preceding operations are ongoing. Some facets - exploring partial data sets, defining priority outputs, GIS - need not wait. Also, the border between data management and analysis is not always clearly marked; in any event both functions may be discharged by the same team members, who mix them pragmatically.
By and large, however, the analysis follows a similar sequence of phases. After initial checks and documentation, the analyst - as an individual or in a team with its own division of labor - produces and interprets descriptive statistics, most of which are of single variables. This is universally valuable. However, for the following steps, it makes a big difference as to whether the assessment was designed on the basis of a detailed requirement list and accompanying analysis plan:

- If there is a pre-existing plan, the desired analytic output types are most of them known in advance, and the default assumption is that the data will be good enough to deliver them. Some of the data will disappoint - which itself is an analytic result -, but (hopefully) not to the point of dislodging the major thrust of the assessment.

- In the absence of such a plan, the variables to be collected were chosen using common sense as well as country and sectoral knowledge, available secondary data or even from a broad shotgun philosophy trying to cover the unknown, yet potentially interesting.

In this case, a layer of assumptions about relevant relationships may intercede between descriptive statistics and analysis. The assumptions - more appreciatively, one might call them mental models - will be driven by ideas of disaster behaviour, response planning and - importantly! - by first impressions of which variables performed well, and which are defective or trivial.

It is not possible to give universal guidance for the creation of appropriate mental models, except to emphasize that they need to relate to the basic dimensions of the assessment - to matrices of regions, sectors (needs areas) and social or affected groups. Also, after inspecting descriptive statistics, analysts will intuitively focus attention on a smaller set of variables whose distributions and associations they feel will be key to characterizing the situations of interest. At this point, it is important to deliberately neglect the analysis of variables of lesser immediate interest, assuming that they too have been minimally cleaned and documented during the data management phase and are readily available should they be of interest again.

This note, as well as the two accompanying workbooks, leans more towards the second scenario, particularly in the analysis part. However, the tools offered for data management and for the analysis of particular data types should be of help also to those who are guided by a clear list of analysis requirements.

3. Overlap with existing ACAPS resources

This note overlaps with these related documents published in the ACAPS resource repository:

The afore-mentioned:

- *How to approach a dataset - Part 1: Data preparation*

  is the companion note that concerns operations preceding analysis.

- *Severity rating: Analysis note* [1]

  Exemplified with data from an assessment in Yemen, this document offers guidance on formal devices (tagging variables, Pivot table templates, semi-automated recoding).

- *Heat maps as tools to summarize priorities expressed in needs assessments*[2]

  A demonstration, using a segment of the Bangladesh assessment data, of how priority data (with unique first, second .. n^{th} choices from a menu of more than n options) can be handled for group comparisons and visualized in heat maps.

These documents are not necessary for the proper understanding of what follows here. They do elaborate on some of the topics. For example, *"Severity rating: Analysis note"* includes an extensive discussion of recoding issues.

4. Excel techniques

We manage and analyse data in the spread sheet application MS Excel. We assume that readers have mid-level skills. They are familiar, or can familiarize using Excel's Help function, with the following features:

- R1C1 reference style
- Absolute, relative and mixed-typed cell references
- Named ranges
- The general workings of Excel functions
- Conditional formatting.

The analysis forms that we describe make use of these specific functions (in alphabetical order): AVERAGE, AVERAGEIF, CONCATENATE, COUNT, COUNTA, COUNTBLANK, COUNTIF, IF, INDIRECT, ISERROR, LEFT, MATCH, MAX, MEDIAN, MIN, SUM, TEXT.

[1] January 2012 [This is the working title on the ACAPS Web page; the formal title on the note is "Data analysis in needs assessments"];
http://www.acaps.org/resourcescats/download/severity_rating_analysis_note/88.
[2] October 2011;
http://www.acaps.org/resourcescats/download/heat_maps_as_tools_to_summarise_priorities/69
.

All operations can be achieved with those features. However, in order to demonstrate efficient tabulation of multiple variables with similar, but not identical sets of categories, we offer two user-written functions (UNION and UNIQUE). These features are optional. They assume that users can access such functions (in the formula bar). Users do not have to write code, but if these features are to work in their own data files, the code stored in the demo workbook should be copied to a VBA module of their own.

To offer these user-defined functions, the demo workbook has been saved in macro-enabled mode (.xlsm).

5. Data and workbook

The demonstration data for this note is from the *"Joint Humanitarian Needs Assessment of Water-logged Areas of South West Bangladesh 2011"* conducted by a group of NGOs and assisted by ACAPS.

The demo workbook is called *HowToApproachADataset_Analysis_130318AB.xlsm* and can be found at *http://www.acaps.org/en/resources*.

6. Analysis phase

Backup and file name advice apply, as in the data preparation phase - see the companion note.

6.1 Data and documentation

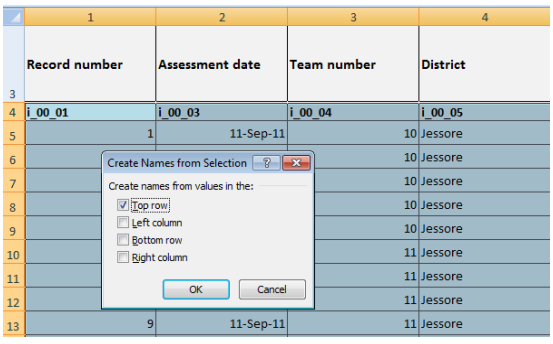
What

We received a dataset that is clean and well formatted. We make a work copy, in which - if it has not yet been done - we create named ranges for all of the data columns as well as for the entire data range. If the variables have not yet been documented, we do so in a new worksheet, where we also calculate their descriptive statistics.

How

- Open the workbook and, if you are in A1 reference style, switch (for ever!) to R1C1 style (Excel Options - Formulas - Working with formulas - check R1C1 reference style.
- Browse through your clean data set ("StartOfAnalysis_Data") and make a work copy ("Analysis01_Data"). Name the data columns after the short variable names, by selecting R4C1:R67C119, then from the menu: *Formulas - Defined names - Create from selection - check Top row only.*
- Name the entire same range, in the address box or the Name Manager, "Database".
- Add a new sheet "Analysis01_Variables", document the variables and calculate their descriptive statistics by the methods detailed in the Data Preparation note, section 1b.

Figure 1: Create names from selection



Why

If the data come from another source or team member, we can legitimately expect a clean and well-formatted data table. Often, this will not be the case - which is why ACAPS offers guidance on data preparation. Yet, even if the originator cares about data hygiene, the documentation may be rudimentary or non-existent. We thus assume that the reader, in the real world, himself has to create the documentation. The naming of all the data columns is essential; it makes the data accessible for the speedy calculation of descriptive statistics in the Variables sheet (via the function INDIRECT) as well as for several subsequent analysis steps.

[Sidebar:] How Pivot tables remember their databases

We name the data range (including the short variable names) "Database" because this is a convenient name for the range to be passed to Pivot tables.

Pivot tables work with an internal copy of the data range - a copy made at the time when the Pivot table was first created. However, our data table may grow through the addition of calculated variables. This, in fact, has happened with our Analysis01_Data, starting from column 120.

If the "Database" name definition is changed in order to include these newly added variables, Pivot tables created earlier (and any copies made of them) do not reflect the new definition. We will not find the new variables displayed in the Pivot Table Field List - until we refresh the table (right-click on the table - then Refresh).

Occasionally, the Refresh command does not seem to produce this result. A fail-safe way around this problem is to repeatedly, as new variables are added, select the grown data range (again including the variable names!) and give it a name that reflects how far it has grown, e.g. "DatabaseCol121" to refer to a range up to and including column 121.

New Pivot tables can then be based on the new name for the database. The price to pay for this is a larger file size. With successive additions of calculated variables, and Pivot tables built on them, the number of copies of data ranges stored internally too grows. Therefore, first adjusting the old "Database" definition, and then refreshing any Pivot tables that need access to the new variables is preferable - if it works.

6.2 Analysis by data type and method

This chapter consists of several sections each referring to a typical data type and to methods appropriate to dealing with it by the tools of Excel. We present basic operations for these situations:

- Overview by key grouping variables
- Summary of all binary variables
- Statistics of preferences measured ordinally
- Tabulation of multiple variables with similar or overlapping category sets
- Multiple-response data tabulation
- Transforming text variables into sets of indicator ranking variables

6.2.1 Overview by key grouping variables

What

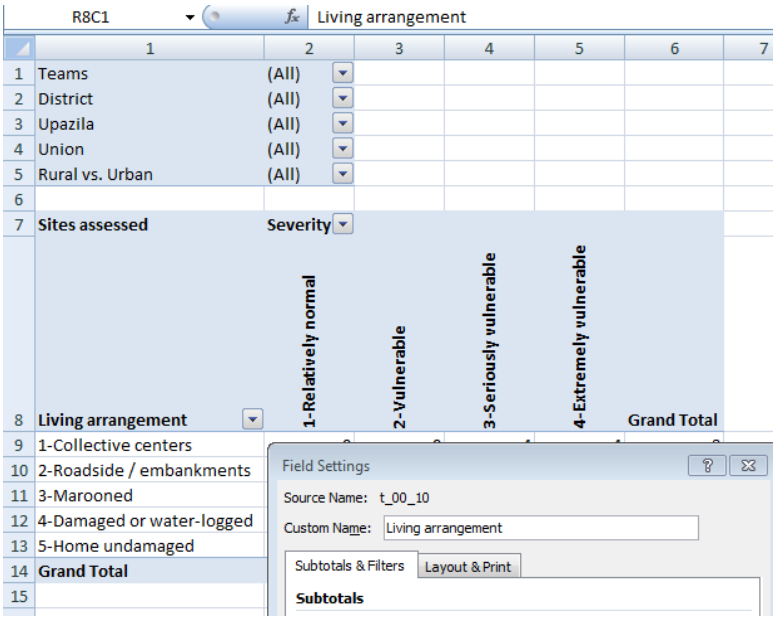
The analyst will want a rapid overview of how the assessed sites are distributed by the most important grouping variables. These are categorical variables that express region, social/affected group, likely major needs (the "sectors"), or other significant ordering criteria. These can be arranged in a number of two-way tables, with filters as appropriate, in the form of Pivot tables. The first of these Pivot tables may serve as the template for several to follow - in other words, the next are created by copying the template to a new sheet, then applying the desired modifications.

In the demonstration data, major grouping variables are "Living arrangement" and (team-assessed) "Severity", plus location (most informatively, perhaps, grouped to sub-districts [Upazila]).

How

- Click the range "Database" in the address box drop-down menu, and create the Pivot table in a new sheet (as in Analysis01_PivotGrouping).
- Optionally, choose the "Classic Pivot Table layout" (under PivotTable Tools - Pivot Table - Options - Display), which enables dragging of fields in the Pivot table grid, and display the column categories vertically, to save space and see the entire table without sliding the view.
- Optionally, rename variables in the Pivot tables by their more intuitive variables labels, e.g. "Living arrangement" instead of "t_00_10", by using the "Custom Name" facility in the Field settings. Field settings can be brought up by right-clicking the variable (in the table or in the drag field areas in the Field List).

Figure 2: Renaming variables in the Pivot table



Why

Sheet "Analysis01_PivotGrouping" illustrates the basic overview challenge. The analyst, from the start of the analysis phase, needs a device that allows for the rapid cross-tabulation of cases by grouping variables that are key to the needs assessment. Here, given the concerns of the

Bangladesh relief community, "living arrangement" and "severity" make important distinctions among the flood-affected communities.

Thus, three basic overview tables may be suggested for this context:

1. Living arrangement x Severity
2. Upazila x Living arrangement
3. Upazila x Severity

In the report filter area, the Pivot table shows administrative tiers (district, sub-district [Upazila], commune [Union]), the assessment teams, and the rural vs. urban status. Subsets defined by the values of report filters can be immediately displayed, and partial data tables with records limited to the defined subset are obtained simply by double-clicking the count, row total or column total of interest.

Fields (in Excel lingo; variables in ours) can be renamed in the Pivot table; for the most often used this may make sense, particularly if the Pivot table serves as a template for others. A custom name can be freely chosen; Excel internally preserves the source name.

6.2.2 Summary of all binary variables

What

For a quick overview, we throw all binary - yes/no type - variables, coded (1 / 0), into the same Pivot table to calculate their frequencies (as fractions, to resist the temptation to talk of percentages when samples are small). If the data table holds too many binary variables, we create subgroups, on substantive criteria, and make several Pivot tables.

How

- Scroll down the Variables sheet (here Analysis01_Variables), determine the binaries of interest, and copy their labels and short names into a new sheet (as in Analysis01_PivotBinaries, R10C1:R30C2).
- Take a copy of the overview Pivot table template and paste it into this sheet, starting at R1C3.
- Modify the Pivot table, as seen here, by moving the template's row and column variables into the Report File, dragging the binaries to the Values area of the Field List, left-clicking each of them, and in the Value Field Settings choose Average as the summarizing function.
- Insert or delete, in columns 1 - 2, some cells in order to align the list of labels and names with the corresponding names in the Pivot table.
- Select the numeric range R10C4:R30C4 and use Conditional Formatting - Color Scales, and select a color ramp that suitably distinguishes between low and high values.

Figure 3: Pivot table collecting frequencies of binary variables (segment)

	1	2	3	4
1			Teams	(All)
2			District	(All)
3			Upazila	(All)
4			Union	(All)
5			Rural vs. Urban	(All)
6			Living arrangement	(All)
7			Severity	(All)
8				
9			**Values**	**Total**
10	Stagnant water: Settlement	t_02_01	Average of t_02_01	0.81
11	Stagnant water: Shelter	t_02_02	Average of t_02_02	0.37
12	Stagnant water: Water points	t_02_03	Average of t_02_03	0.60
13	Stagnant water: Other	t_02_04	Average of t_02_04	0.10
14	Garbage	t_03_00	Average of t_03_00	0.86
15	Disease vectors	t_04_00	Average of t_04_00	0.94

Why

The binary variables collected in this Pivot table have little in common with each other, except that all of them are binary variables, and have been coded as recommended, with 1 for yes, 0 for no (and blank or a text code for missing).

Nevertheless it may make sense to produce their frequencies side by side for all in one table initially, if only for a quick overview of which differentiate well, which may be suspect of over- or underreporting, or which should be looked into for subsets of cases.

These background colors have a purely heuristic value for the visualization of higher and lower frequencies for each variable solely and make no sense for substantive comparisons between them. The table is essentially a rapid lookup device.

Benefits may arise later in the analysis. This table too may be used as a template, this time for small groups of substantively related variables. A copy can be pasted to a new sheet; there it can be reduced to its desired form. Two steps are usually involved: dragging irrelevant variables out of the Values area, and cross-tabulating by dragging a report filter or other variable of interest into the Column Label area.

Figure 4 gives a simple example, of four binary variables related to stagnant water, with district-wise and overall frequencies (columns 2 and 3 are temporarily hidden for easier viewing). We take from this, tentatively, that teams in Jessore and Khulna observed that the sheltered population was much less inconvenienced by stagnant water than those living in their homes. In Shatkira, this was a problem also with half of the shelters.

Figure 4: Frequencies of several binary variables, by a categorical variable

	District			
	Jessore	Kulna	Shatkira	Grand Total
Stagnant water: Settlement	0.96	0.60	0.74	0.81
Stagnant water: Shelter	0.21	0.00	0.53	0.37
Stagnant water: Water points	0.79	0.20	0.53	0.60
Stagnant water: Other	0.13	0.00	0.09	0.10

6.2.3 Statistics of preferences

What

We compute statistics of relative preference for multiple-choice variables with partially ranked items. Assessments generally apply one or the other of two question forms with which to measure preference:

- Respondents hear a list of *n* items and are asked to select the *m* most important ones (mostly *m* < 4). The m selected options are treated equally.
- Or, they are asked to nominate the most important item, second-most important one, etc., usually also for a small number *m* only, with the remainder of the items not ranked. The selected items are coded by their ranks.

We provide formulas for both approaches. The second leads to a measure known as the Borda count. A sidebar gives background.

Speed is a constant consideration. It may be efficient to throw all response variables to all priority-type questions into one sheet, calculate the measures for all of them there, and create smaller tables, as needed, from copies of this sheet somewhere else.

This has been done for the Bangladesh data, in sheet Analysis01_Priorities. Note that the ranked data has already been coded, *in the data table*, descendingly, meaning: The first priority is valued 3 (where three priorities were elicited), respectively 5 (where five). Items not ranked as priorities are universally coded 0.

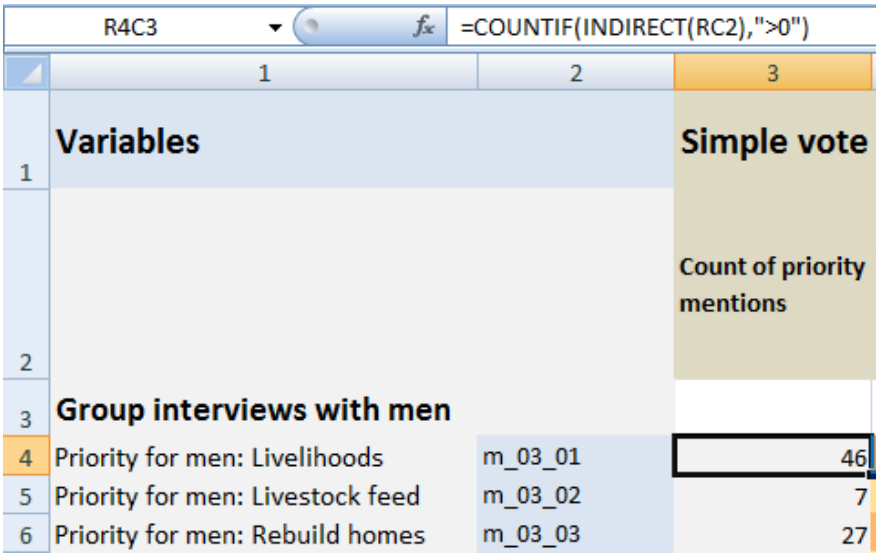
The degree of difficulty of this section is somewhat higher, and it is not feasible to detail the finer steps all in this text. The reader may want to work through the formulas of the demo sheet column by column. Note that we work with functions only; there are no Pivot tables.

How

In columns 1 - 2 of a new worksheet, gather the labels and short names of variables, in blocks by question.

- **With *m* priority items, unranked:**
  - Reference the named data columns and use the functions COUNTIF and INDIRECT to calculate how many times an item was selected as a priority, as in column 3 of the demo sheet Analysis01_Priorities, in formulas like =*COUNTIF(INDIRECT(RC2),">0"):*

Figure 5: Counting simple priority votes



	1	2	3
1	**Variables**		**Simple vote**
2			**Count of priority mentions**
3	**Group interviews with men**		
4	Priority for men: Livelihoods	m_03_01	46
5	Priority for men: Livestock feed	m_03_02	7
6	Priority for men: Rebuild homes	m_03_03	27

- Make the counts comparable by rating them to the number of respondents who ranked any items (and thus do not have missing in this variable), by formulas like:

  =COUNTIF(INDIRECT(RC2),">0") / COUNT(INDIRECT(RC2))

- Copy the formulas downward and use conditional formatting, separately for each question block, as shown in column 4 of the demo sheet.

- **With *m* priority items, ranked - Borda count:**

  - Create a setup as in columns 6 - 12 of the demo sheet.
  - Count priority votes, in columns 6 - 11, by referencing also the priority coding in row 2, with a formula like

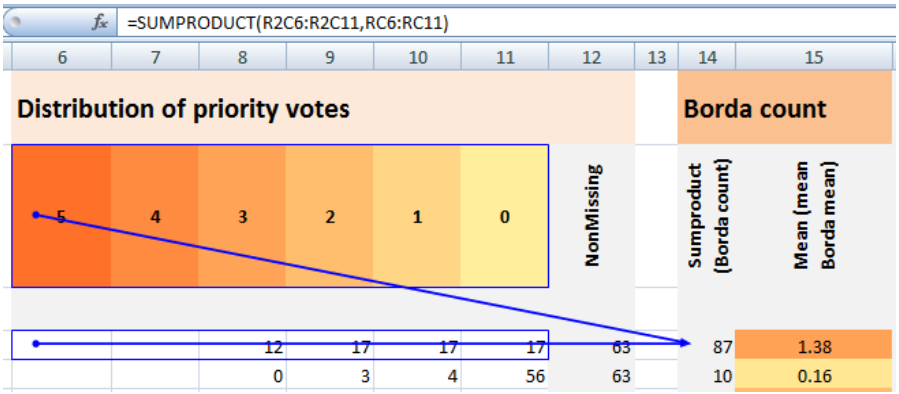
    =COUNTIF(INDIRECT(RC2),R2C)
    [Note the difference in the mixed references!]

  - In column 12, calculate the number of NonMissing as the denominator for the mean Borda count, through *=COUNT(INDIRECT(RC2)).*

  - In column 14, calculate the Borda count as the sum of the products of the priority scores by the votes each priority received, e.g., for livelihoods: 3 * 12 + 2 * 17 + 1 * 17 = 87. Excel has a function for this:

    =SUMPRODUCT(R2C6:R2C11,RC6:RC11)
    [Note the absolute reference to the priority scores, and the mixed reference to the counts, so that the formula can be copied down without modification!]

Figure 6: The Borda count calculated with SUMPRODUCT



- Copy the formulas downward and use conditional formatting, separately for each question block, as shown in column 15.

Why

Responses to multiple-choice questions may be ordered on a preference or priority scale. The respondents may voluntarily reveal a complete or partial order ("We do want to return home, but right now, as we are stuck in these shelters, food aid is more important than house repair materials.") or may be induced, by interviewer requests, to respond with an ordering attempt. In the first case, the set of items may arise from the respondents' statements and may be recorded as such by the assessment team or recoded and combined in situ. In the second situation, interviewers typically make respondents select from a pre-defined list of items, sometimes suggesting that respondents may also express their own, to be recorded under "Other".

Often, priorities (of needs, issues, or other types of items) are elicited in conversation from most important ("first priority") to successively less important ones. The elicitation stops after a small number, typically smaller than the set of items. The remaining items, while formally missing values, are presumed to be of equally low importance.

In data entry, the elicited priorities can be recorded in different formats. Sometimes, they are entered as codes or text in variables that are named "first priority", "second priority", etc. This poses an analytic challenge; in the auxiliary sheet Text02_Priorities we demonstrate the transformation of this kind of priority variables into item-based variables[3].

At other times, the database carries a numeric variable for each item. For unranked priorities, any prioritized items are coded 1, and all non-prioritized items are coded 0. For ranked priorities, the natural coding is "1" for "first priority", "2" for the second, etc., but this is not optimal for analysis. Non-prioritized items are either left blank or filled with zeros. If the coding was "natural", the ACAPS note on "*Heat maps as tools to summarize priorities"* offers data preparation formulas.

Thankfully, the Bangladesh dataset comes with yet coding mode for this type of data, one that is more analysis-friendly. The value recorded for the elicited priority of an item is equal to *"Number*

[3] Whether such transformations should take place already during data preparation is a matter of taste.

of priorities elicited in the instructions plus one minus the conversational priority numerical". In plain English, if the interviewers were instructed to elicit five priorities, the respondent's first priority is scored 5, the second 4, etc. This is regardless of whether in the event the particular respondent produced five priorities or fewer. Non-prioritized items were left blank; these blanks - provided the respondent prioritized at least one item - can legitimately be flushed with zeros.

These data are of the ordinal type. Frequencies, extremes and medians are legitimate statistics. Sums and means are not. However, under conditions of

- Mutually exclusive ranks (which imply preferences between items, not independent intensity ratings)
- Avoidance of items that, by meaning or coverage, are near-duplicates (e.g., presenting respondents with both "livelihoods" and "income")

Priority rankings (with the highest priority getting the highest value, etc.) can be treated in analogy to a voting system. For its simplicity and consensual character, the *Borda count* is a procedure that appears suitable for this data situation.

[Sidebar:] Simple vote counts vs. the Borda count

We demonstrated the calculation of two measures expressing relative item preference, one based on counts of unranked prioritized items (simple votes for an item), the other weighting the counts by the priority score. The latter is known as the Borda count. As a rarely used concept, it deserves some explaining.

Also, we ask whether the insights that the Borda count offers over simple vote counts justify the higher computational effort. Of course, this question ought to be asked already in the design of the assessment, with a view to the interview situation: it takes longer to elicit ranked priorities than unranked ones. Eliciting ranks, on the other hand, may lead the respondents to be more thoughtful, and the response thus more valid.

The Borda count is an election method named after one of its proponents in the eighteenth century (Borda 1781). In its basic form, voters rank candidates. If *n* candidates are running, a vote for the first preference counts as n points, the vote for the second as *(n - 1)* points, and so on. The winner is determined by the highest sum of points that he received from the voters. Over time, numerous modifications away from the basic form have evolved.

The Wikipedia article (Wikipedia 2011) gives an excellent overview. As all electoral systems, the Borda count meets a number of desirable properties and fails on others. Its strong characteristics are its simplicity and consensual character. The latter means that also candidates who are favourites with a minority of voters only stand a chance to win if only they enjoy mid-level preference among a sufficiently large majority. Two disadvantages often held against this method are the possibility that strategically added candidates distort the results (in other words, such a "clone candidate" may let a strategically acting clique-leading candidate win over someone meant to win if the clone had not been added) and, similarly, its lack of robustness to so-called irrelevant alternatives.

The use of the Borda method in social measurement (such as of priority needs among disaster victims) is not new (Lansdowne 1996). Frequently, so many items are presented that respondents manage to rank only a few of them. This may be anticipated; the format will then allow only a small number of ranked items. These are scored such that the score for the item of the highest priority equals the number of allowable rankings (rather than the number of the actually elicited ones). Items of lesser priority are scored descendingly, as the number of allowable rankings -1, then -2, etc. All unranked items are scored 0.

However, two circumstances need to be kept in mind: While elections to office are about clearly distinct candidates, preferences among items relate to concepts that, semantically and institutionally, *overlap* in various degrees. Second, using the Borda method, or for that matter any election method, to establish a ranking of needs implies a *decision*, by some outside agency, that the relative strength of particular needs should be established (and ultimately reflected in policy) under this particular format, and no other.

Two precautions arise:

- The item pool - the set of conceptual stimuli under some thematic heading with which assessment teams present key informants or focus groups - should consist of well distinguished elements. Items with strong conceptual overlap may cause stealing of votes, in sometimes unpredictable ways: either by scattering points about the same "thing" among its similar expressions (and thus depressing the prominence of both), or by diverting attention from "another thing" that, in the absence of conceptual clones, would be more prominent in the revealed needs hierarchy[4].

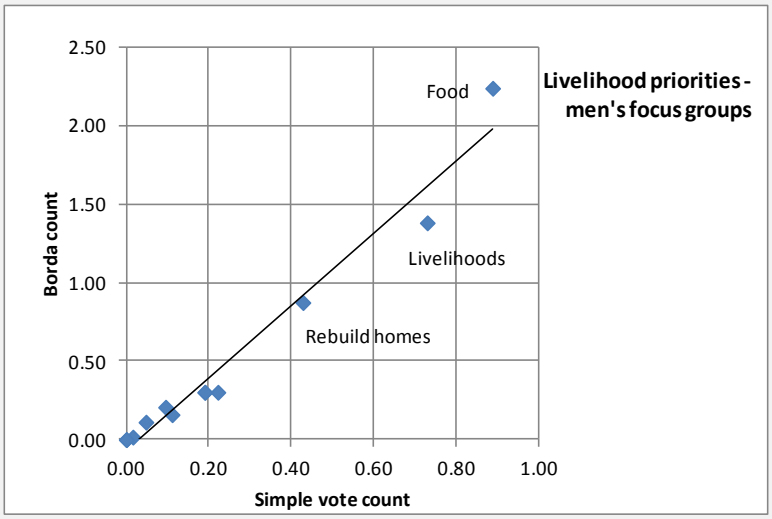
- Second, as an election method, the Borda count is meant to produce winners and losers. Although the point totals for candidates are ratio-level count variables, pragmatically the results are reduced to an ordinal reading: winner with the highest point total, winner with the second-highest, etc., candidate who among the losers had the highest total, etc. However, when we use the Borda count in social measurement, the logic is reversed. We start with preference rankings; these are ordinal. It is the second mental operation - stipulating that these rankings are the elements of this particular election method - that authorizes the calculation of ratio-scale scores (the mean Borda count is the point total for an item divided by the number of voters [sites, community groups, etc.]). The fact remains that the underlying preferences were measured on an ordinal level.

Both precautions together will cause the analyst to interpret with great caution the distribution of Borda counts among items that figure in a needs assessment. Items with very high mean Borda counts plausibly enjoy genuinely high priority. Thus, among the needs expressed by focus groups in Bangladesh, food (2.24 for men, 2.30 for women) scored far above the second priority, livelihoods for men (1.38), rebuilding homes for women (0.98). However, to say that, for women, rebuilding homes was more urgent than the provision of safe drinking water (mean Borda count = 0.68) is questionable. It may be safer to suggest a semi-order: Food is a distinct priority over all others; then there is a set of 2 - 3 needs that are less important than food, but clearly more than the rest of the elicited needs.

[4] In a recent needs assessment in northern Syria, it was found, surprisingly, that "fuel", with its numerous uses, tended to attenuate the importance of several other needs - shelter, food, water.

Is this extra conceptual and computational effort worthwhile? This chart compares the fractions of respondents who voted an item as an (unranked) priority vs. its mean Borda count. In both measures, the majority of the item points huddle near zero - in agreement with common sense that only a few needs are exceedingly strong at any one point of time, and others are basically indistinguishable in their relative urgency.

Figure 7: Comparison of simple vote count and Borda count



The points are aligned fairly closely with the straight trend line. Rank differences are seen only in the little prioritized items, near zero. Yet, there is an important difference: In the simple count, the livelihoods have almost the same priority as food. The fuller information that the Borda count incorporates lets us see that the relative difference between food and livelihoods is more pronounced. This may indicate that in the views of the men's focus groups (and even more of the women's!) the need for food aid is yet as great as to push livelihoods rehabilitation to a distant second. The simple vote count does not catch this nuance.

Ultimately, it is time pressure and analytic comfort that should determine the choice between these two methods. The Borda count formulas are easy to copy and adapt from this demo sheet; they may be harder to sell to assessment consumers.

6.2.4 Tabulation of multiple variables

What

We tabulate the frequencies of multiple variables, side by side. This operation is often called for in situations where the same variable was measured "before" and "after", or was measured separately for particular groups such as men and women.

A small, but vexing problem may arise when the categories actually used are different across variables, and the data table is so large that quick comparison by filter dropdown menus is not feasible.

How

- Use a set-up and the functions COUNTIF and INDIRECT as in this figure, with variable labels and short names arranged in column headings, and the set of all categories ever used as row labels.

Figure 8: Multiple-variable tabulation

R11C3 =COUNTIF(INDIRECT(R10C),RC2)

	2	3	4	5	6	7
8	Predominant defection mode	Men			Women	
9		Before	After		Before	After
10		m_06_01	m_06_02		w_11_01	w_11_02
11	1-Open area	3	20		0	26
12	2-Hanging or open latrine	2	20		14	17
13	3-Sanitary latrines (communal)	15	13		0	10
14	4-Sanitary latrines (household)	42	5		49	6
15	5-Other	1	5		0	4
16	Total	63	63		63	63

- Use formulas like that in the formula bar of the figure, and SUM for the column totals.
  [Note the two forms of mixed-type references in R10C and RC2!]

- In a situation where the set of all categories ever used in these variables is not immediately established, the user-defined functions UNIQUE and UNION can help, as shown in sheet Analysis01_TabCategorical.
- Optionally, use conditional formatting to highlight extremes.

Why

Pivot tables are highly efficient for *multi-way* tabulations, but, to our knowledge, less so for tabulations of *multiple* variables side-by-side. Multi-way tabulations produce statistics for every cell representing a combination of values of several variables. Tabulations of multiple variables are simply frequency distributions, over the same set of categories or numeric values, of several individual variables.

The latter can, of course, be produced by several Pivot tables and the stitching together of copies thereof. However, this is time-consuming. It gets all the more tedious when the actually occurring categories or values across the variables differ. The stitching-together process is slowed down by the need to insert cells in for items with zero frequency.

Sheet "Analysis01_TabCategorical" gives an example of tabulated multiple variables. Looking at the sanitation modes used by women before the disaster, it is obvious that their repertory was smaller than the men's.

Similar challenge can arise when there are subtle spelling differences (e.g. trailing spaces) in the same categories between different variables.

A combination of user-defined functions helps us to quickly generate the union of all category or value sets. User-defined functions and their placement in VBA modules were discussed in the

data preparation note; the two functions needed here - UNIQUE and UNION - are in the modules already attached to this demo workbook.

UNION creates a combined range out of all named data ranges used in the tabulation. UNIQUE lists the distinct values of this super-range. In set-theoretic terms, the two functions together produce the *union of the sets of categories / values* actually used in the variables being tabulated. The syntax is shown in the worksheet. Here two general remarks must suffice:

1. If the union contains variants of the same category (spelling differences, inconsistent prefix, etc.), using the set *as is* to generate the table, and then editing it manually is efficient for one-time use of the variables. If we anticipate using these variables more often, corrections should be made in the data table. The union will update automatically, as long as the formulas are preserved.

2. For numeric variables with many values, particularly continuous ones, computing for each of them a categorical variable (use the function ROUNDDOWN to produce the lower bounds of ranges) in the data table may be the most efficient preparation.

The complexity of tabulations can be multiplied with the help of these functions and of related means, although not as conveniently as Pivot tables let us do for multi-way tabulations. Thus, when tabulations of multiple variables are to be produced on subgroups (say, by district), the function COUNTIFS (ifs - plural!) may handle the task. This is beyond the scope of this note.

6.2.5 Multiple-response data

What

This is about multiple-response data on items that are not ordered (no preferences or priorities are implied). We analyse this type of data the same way as binary variables - in Pivot tables that show the frequency of each item.

However, there is a difference in presentation, about which the analyst (and any tables that he puts in the report) needs to be clear: Frequencies can be based either on the number of *respondents* (notably the number of sites in the needs assessment) or on the total number of *choices* exercised.

We also consider situations where missing data are a problem, or where the response was entered as text in the order in which they were recorded in the interview.

How

- Create a setup similar to the one demonstrated for binary variables, with variable labels and short names listed on the left, and a Pivot table created and aligned with them on the right. The sheet *Analysis01_MultipleR_Pivot* exemplifies the process.
- In a copy of the Pivot table, replace the Pivot table's value field expressions (e.g., Average of m_07_01) by an item description suitably short for tables (e.g., "Shelter not available"). Color the frequencies using conditional formatting, and calculate the mean number of choices exercised, per group (which is equal to the column sums of the frequencies). This is the view by *sites.* For example, 52 percent [if we allow percentages here for a moment!] reported that shelter was not available).

Figure 9: Tabulation of multiple-response data, by sites

By sites	Jessore	Kulna	Shatkira	Grand Total
Shelter not available	0.67	0.80	0.38	0.52
Overcrowded	0.21	0.40	0.15	0.19
Return home impossible	0.67	0.80	0.62	0.65
No repair materials	0.42	0.20	0.35	0.37
No manpower for repairs	0.00	0.20	0.06	0.05
Land issues for men	0.00	0.20	0.03	0.03
Lack of household items	0.17	0.00	0.32	0.24
Lack of privacy	0.17	0.00	0.18	0.16
Exposed to weather	0.08	0.00	0.15	0.11
Other	0.21	0.00	0.09	0.13
Sites	*24*	*5*	*34*	*63*
Mean number of choices	*2.6*	*2.6*	*2.3*	*2.4*

- Make a copy of this table and divide the frequencies by the mean number of choices, in each group. This is the view by *choices* (e.g., 21 percent of all problems formulated concerned the scarcity of shelter)[5].

Figure 10: Tabulation of multiple-response data, by choices

By choices	Jessore	Kulna	Shatkira	Grand Total
Shelter not available	0.26	0.31	0.16	0.21
Overcrowded	0.08	0.15	0.06	0.08
Return home impossible	0.26	0.31	0.27	0.27
No repair materials	0.16	0.08	0.15	0.15
No manpower for repairs	0.00	0.08	0.03	0.02
Land issues for men	0.00	0.08	0.01	0.01
Lack of household items	0.06	0.00	0.14	0.10
Lack of privacy	0.06	0.00	0.08	0.06
Exposed to weather	0.03	0.00	0.06	0.05
Other	0.08	0.00	0.04	0.05
Total	*1.00*	*1.00*	*1.00*	*1.00*

[In this example, the differences between the two views are minor because the mean number of choices was similar across the three districts. This is not always so. Which view to favor in the report depends on analytic context and purpose.]

- In rare cases:
  - If missing values are a significant problem, determine the criteria on which to exclude cases with missing. Two situations are typical: 1. the respondent was not exposed to some of the items in the list [for whatever reason]; 2. he did not respond to the question are all. Use a setup as in sheet *Analysis01_MultipleR_IF* with a calculated variable.

[5] Some reader may wonder why we chose the terminology "by sites" vs. "by on choices". This is meant to reflect the typical needs assessment situation. The ordinary statistical usage is "by cases" vs. "by responses".

- If the response data was entered as text in the sequence of recording the choices, create indicator variables, as demonstrated in sheet *Text01_Polytomous*.

Why

The way Pivot tables can be used to analyse multiple-response data tells us nothing new beyond what we have seen in the treatment of binary variables. The difference is that with multiple-response data the results can be presented in two ways, depending on the choice of denominator. We can base the item frequencies on the number of sites responding to the question, or else on the total number of choices recorded.

Two more considerations apply:

1. The Bangladesh data, and our demonstration, is based on the so-called indicator mode of storing the response data. In this, each item has its own variable, where 1 stands for "present" or "yes", and 0 for the negative value. The other storage mode is called polytomous, in which responses are recorded by the order of appearance in conversation, in variables like FirstResponse, SecondResponse, etc.

   Counting instances of items in polytomous data *globally* is feasible as long as the set of items is known - one simply lets COUNTIF suck them up, like a vacuum cleaner, across the multi-column range that stores them. If more subtle and repeated calculations are needed, it is efficient to first transform the polytomous variables into indicators, which we show in the section "Text variables" below.

2. If missing data are a significant problem, they have to be taken into account. Missing values may arise from two situations. A respondent may not have been presented with a subset of the choices available, or the subset may not be available to him for some (other than interview-related) reason. Second, he may not have responded to the question at all (again, either because he was not asked, or because it did not apply). In neither case was it the respondent himself who excluded the items from consideration. Therefore treating them as absent - flushing the cells with zeros - would not be correct. Again we show how to deal with such situations in an auxiliary sheet.

Dealing with many multiple-response questions that were recorded polytomously or have missing-value problems is time-consuming. This is one of the situations where translating the data to a statistical application like STATA and using a dedicated multiple-response routine is likely to be more efficient (see below, page 35).

6.2.6 Text variables

The Bangladesh data table does not employ text variables more complicated that the single-column categorical variable written out in text. All multiple-response variables were nicely entered as sets of 1 / 0 - indicator variables or, in the case of priority data, as Borda scores.

Not every dataset is as considerately formatted when it comes to the treatment of multiple responses captured in text. We often find sets of text variables, with records that fill some of them, and leave others blank, usually in the order in which the response items were recorded. Similarly, the variables may be filled, from left to right across the set, with the item of highest priority, followed by that of the second priority, etc. A second important distinction is between

response categories that are strictly standardized and free-form text. Categories with spelling variants in dirty tables are also quite common; technically they pose the same problems as free-form text.

This table segment, copied from sheet "Text01_Polytomous", illustrates one of the text data situations - the strictly standardized response categories without priority ranking.

Figure 11: Multiple response in text form

RecordNo	Multiple response data: Shelter issues for men			
	Recorded 1st	**Recorded 2nd**	**Recorded 3rd**	**Recorded 4th**
recno	**ShelterMen1**	**ShelterMen2**	**ShelterMen3**	**ShelterMen4**
1	Overcrowded	Not available		
2	Return home impo	Not available		
3	Return home impo	Not available		
4	Not available	Overcrowded	Return home impossible	
5	No repair material	Overcrowded	Not available	

This type of multiple-response data is known as the *polytomous* form (suggesting multiple choices or branching). The analytic strategy is to convert the polytomously recorded text data into sets of dichotomous indicators (for unordered multiple response) or ordinal variables (for preference-type data). In this shape, the multiple response is amenable to statistical description and to selective cross-tabulation with other variables of interest. This table shows the result of the transformation of the above table segment.

Figure 12: Multiple response in indicator form

RecordNo	Multiple response data: Indicator form			
	Not available	**Overcrowded**	**Return home impossible**	**No repair materials**
recno				
1	1	1	0	0
2	1	0	1	0
3	1	0	1	0
4	1	1	1	0
5	1	1	0	1

Obviously, the key is to define one variable for each standardized category. With free-form text data, that is not possible. Instead, one will form (a limited number of) variables each of which is defined by one key semantic component.

Suppose the free-form response reveals a variety of desires, or comments, about returning home from the flood shelters. Plausibly, then, "home" may be one of the key terms. The indicator variable will take the value 1 if the string "home" is found in any of the text variables in the given record and 0 otherwise.

Also note another difference between standardized-category and free-form response: With standardized categories, the number of positive values in the set of indicator or Borda score variables is strictly equal to the number of polytomous variables used in the given record. This is not the case with free-form text. An entry in one variable may be picked up by several indicator variables because it contains more than one of the key terms. An entry may also go unnoticed if none of the key terms that define the indicator variables is contained in it.

What

Excel formulas are employed to semi-automatically convert text-based multiple response to sets of indicator / Borda score variables.

How

- Define the text data constellation in terms of strictly standardized vs. free-form text as well as of unordered items vs. priority-type data.
- Use the table set-up and the formulas as shown in sheet "Text01_Polytomous" (for unordered items), respectively "Text01_Priorities" (for priority-type data) in order to generate the desired indicator / Borda score variables.

Why

We are dealing with four different situations and, accordingly, with four different formulas. However, the formulas have a common core: They employ mixed-type cell references in order to read in the search term (the column headers) and to reference the polytomous data range in the given record. Plus, they combine, working from inside out, the three functions MATCH, ISERROR, and IF.

The rest are basically just adaptations to the specific problem on hand. Note the concatenation operator (&, ampersand) and the wildcard search term for any number of characters (*, asterisk) in the free-text formulas.

The detailed working of the formulas is explained at the bottom of the two worksheets.

6.3 Analyses by substantive concerns

This section demonstrates a small number of tables that are helpful at various points of an analysis train that follows a substantive logic. By this we mean that the analyst will seek to move from a basic overview - combining major grouping and geographic categories - towards more specific needs, issues, and sector priorities - broken down, as appropriate, by the variables or combinations thereof highlighted in the overview part.

This note cannot develop this substantive logic further, in ways that would cover a broad range of assessment situations. Instead a number of analysis forms that likely respond to this logic at various junctures are illustrated.

6.3.1 Quick severity overview

What

Early on, the assessment process strives to obtain estimates of affected populations by region and major social or affected groups. However, such estimates may not be available at this point. Instead, some type of severity judgment may be attempted for each group or combination of group and location.

When such judgments are provided on an ordered scale, Excel provides limited means to visualize severity for subsamples of sites defined by grouping variables. We demonstrate this for the highest severity rank of any site in a cross-table of sub-districts and living arrangements.

Figure 13: Highest severity level, by sub-district and living arrangement

District	Upazila (Sub-district)	Living arrangements				
		1-Collective centers	2-Roadside / embankments	3-Marooned	4-Damaged or water-logged	5-Home undamaged
Jessore	Abhynagar			3	2	
	Jhikargacha	3	3	3	3	3
	Keshabpur	3	3	3	3	2
	Monirampur		2	3	3	
Kulna	Paikgacha		3	2		1
Shatkira	Assasuni	4	2	4	3	2
	Debhata	4	3	1	4	1
	Kalaroa	3	2	2	1	1
	Shaktira_Sadar	4	4	4	4	4
	Tala	4	4	3	4	2

Legend:
1 1-Relatively normal
2 2-Vulnerable
3 3-Seriously vulnerable
4 4-Extremely vulnerable

How

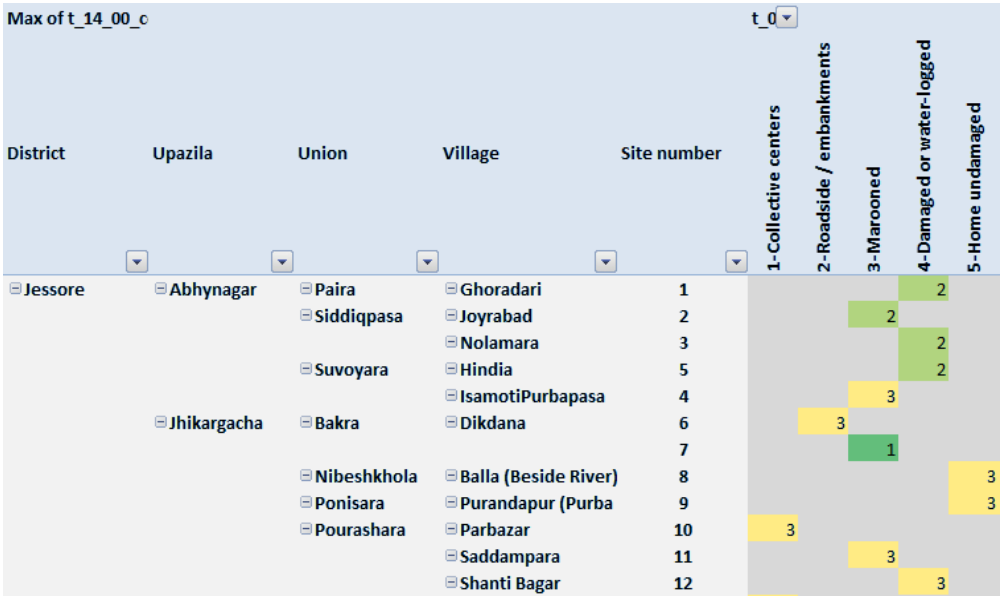
- From the text variable t_14_00, "Site assessment by team", in column 19 of the data table Analysis01_Data, create a numeric ordinal variable by extracting the prefix with the formula

  =VALUE(LEFT(RC19,1))

  where the function LEFT extracts the first character of the string (the prefix), and VALUE tells Excel to treat this as an number so that a statistic of such numbers can appear in the Pivot table data range (where text variables cannot!), as in column 120 of Analysis01_Data.
- Name the new numeric variable something like "t_14_00_code" and label it "Site assessment by team (numeric)".
- Update the database definition (such as by selecting the entire data range up to column 120 and naming it DatabaseCol120).
- Build the Pivot table, as in sheet Analyses02_QuickOverview, setting the summarizing function in the Value Field Settings to maximum.

- Use conditional formatting to color the cells by the numeric value of the severity and create a legend because the meanings of these numbers are no longer self-evident.
- Optionally, make a copy of this Pivot table and, in addition to districts and sub districts, include Union, Village and site number as row labels, as in the right half of sheet Analyses02_QuickOverview, in order to get a full listing of sites, with administrative units, living arrangements, and team-assessed severity.

Figure 14: Numeric severity level, by site and living arrangement, as Pivot table (segment)



Max of t_14_00_c					t_0				
District	Upazila	Union	Village	Site number	1-Collective centers	2-Roadside / embankments	3-Marooned	4-Damaged or water-logged	5-Home undamaged
Jessore	Abhynagar	Paira	Ghoradari	1				2	
		Siddiqpasa	Joyrabad	2			2		
			Nolamara	3				2	
		Suvoyara	Hindia	5				2	
			IsamotiPurbapasa	4			3		
	Jhikargacha	Bakra	Dikdana	6		3			
				7			1		
		Nibeshkhola	Balla (Beside River)	8					3
		Ponisara	Purandapur (Purba	9					3
		Pourashara	Parbazar	10	3				
			Saddampara	11			3		
			Shanti Bagar	12				3	

Why

This section is about two topics that happen to combine in this approach, but each of which can be independently relevant:

Substantively, how do we rapidly form an overview in terms of the severity of the situation when vaguely ranked severity judgments are all we can go by? The table should help to quickly discern at least some of the coarser patterns. Thus, a person looking at our sub district table, colored by the maximum severity level, might find:

> *"Sites with extremely vulnerable people occur only in Shatkira District. In Shakira town (=Sadar), this level of vulnerability is found at some sites across all living arrangements, from collective centers to undamaged homes. Moreover, people stranded on roadsides and embankments, traditionally the worst affected after floods, are not rated extremely vulnerable outside Shatkira Sadar and Tala sub-districts."*

Second, *technically*, the Pivot table demonstrates how to summarize an ordinal measure (such as vulnerability with four levels) against one or two grouping variables. The ordinality has two consequences: the measure can be expressed numerically (good!), but the arithmetic mean is

not a legitimate operation, and the median function is not available in Pivot tables (bad!)[6]. Thus the maximum stands in, as the second-best solution. A low maximum in a cell (green in our formatting) indicates that there are no real "hot spots" among its sites.

Another benefit from tackling this problem with a Pivot table is that, as always, such a table is easily copied, and the copy modified with little effort. We obtain a full listing of sites with visualization, by color, of the severity levels, by site and living arrangement. True, by temporarily hiding irrelevant columns, the same overview could be obtained in the data table. Yet, the Pivot table is more effective as a lookup table, particularly if we wish to use report filters at some point.

In the analytic process, such a lookup table has the function to quickly check the robustness of interpretations that we make of summary result. With every site displayed, it lets us see the diversity inside any subset that is the object of a claimed finding. For example, the impression that within Shatkira town, all affected groups were extremely vulnerable is confirmed when we scroll down the individual sites.

6.3.2 Priorities as a function of vulnerability

What

Intuitively, we expect priorities (for types of relief, government action, people's own coping behavior, etc.) to be influenced by how vulnerable or how severely affected groups are. To exemplify, we analyse priorities revealed by men and women at sites rated for severity. The dependent variable is the score that the focus groups assigned to items that they prioritized. For the reasons explained in the sidebar on Borda counts (page 18), we treat the scores as ratio-level variables and summarize by their means.

How

- As in earlier worksheets, start with the usual pattern of arranging variable labels and short names, similarly to sheet Analyses02_PrioritiesByVulnerab.
- Compute the mean scores with the help of the function AVERAGEIF, as in

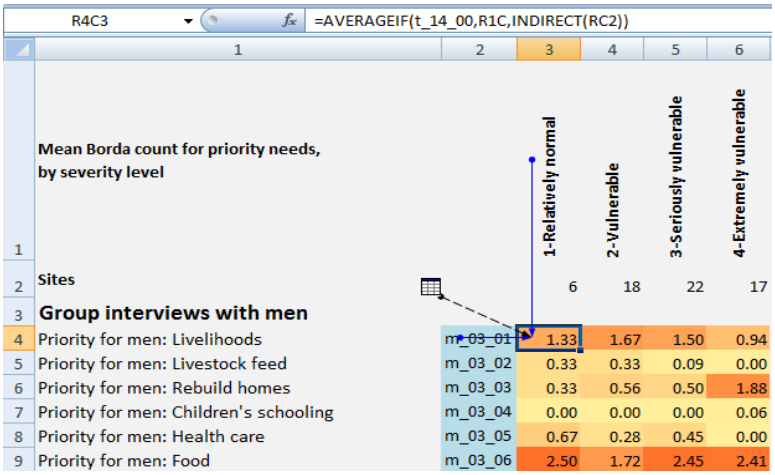
=AVERAGEIF(t_14_00, R1C, INDIRECT(RC2))

Where:

- RC2 references the cell in column2 that holds the name of the priority score data column in point,
- INDIRECT tells Excel to understand the content of this cell as a defined name
- R1C is the mixed-type reference reading in the value of the criterion variable in row 1 (here it is the severity level) applied to the column
- and t_14_00 is the name of the range of the criterion variable in the data table.

[6] Some users may ask whether Excel returns conditional medians through a combination of the functions MEDIAN, AND and IF in array formulas. This works indeed for one criteria (say: subdistricts), but, despite our best attempts, not with several. Readers are welcome to communicate better solutions. As of version 2007, Excel does not offer a MEDIANIFS function, in contrast to AVERAGEIFS.

Figure 15: The function AVERAGEIF



	1	2	3	4	5	6
1	Mean Borda count for priority needs, by severity level		1-Relatively normal	2-Vulnerable	3-Seriously vulnerable	4-Extremely vulnerable
2	Sites		6	18	22	17
3	**Group interviews with men**					
4	Priority for men: Livelihoods	m_03_01	1.33	1.67	1.50	0.94
5	Priority for men: Livestock feed	m_03_02	0.33	0.33	0.09	0.00
6	Priority for men: Rebuild homes	m_03_03	0.33	0.56	0.50	1.88
7	Priority for men: Children's schooling	m_03_04	0.00	0.00	0.00	0.06
8	Priority for men: Health care	m_03_05	0.67	0.28	0.45	0.00
9	Priority for men: Food	m_03_06	2.50	1.72	2.45	2.41

- Compute the mean scores for all sites, with the formula

=AVERAGE(INDIRECT(RC2)).

- Color the output cells, within each question block, using conditional formatting.

Why

As in the previous section, it is helpful to make a difference between substantive and technical points.

Substantively, we explore differences in needs priorities, as formulated by focus groups, as a function of gender and severity (as measured by the team's vulnerability judgments). A reader might note something like:

> *"For both men and women, and at all vulnerability levels, the first priority is food. For the extremely vulnerable, the second priority distinctly is the rebuilding of their homes, which is of lesser priority for most of the less vulnerable groups. The picture of how second and third priorities are distributed is more complex. Of note, among the extremely vulnerable, the distinct third priority for men is livelihoods, while for women this is the provision of safe drinking water."*

Technically, two points may be noted:

- First, this is about conditional means, that is, for each of the priority items, we calculate the mean score over the subsamples defined by categories of some other variable. In Excel, named ranges, mixed-type cell references, the function INDIRECT and the rarely used AVERAGEIF combine for an efficient solution; the same formula can simply be copied to the entire concerned range in the table.
- Second, priority scores are ordinal. However, here we take their arithmetic means. In other words: we treat them as ratio-level. We do so assuming that the focus groups assigning priorities function like voters in a Borda-count election system (see sidebar on the Borda count), In other words: if they choose up to three priorities, their first priority (with score 3) is exactly three times as important to them than the third priority (which is coded with score 1), or 1.5 times as important than the second priority (score 2). Only under this assumption, is the calculation of means in this context justified.

6.3.3 An index of distress

What

So far, the only measure of the overall vulnerability - which we take to be a concept closely allied to severity - is from the assessment teams' personal impressions, not from uniformly evaluating specific types of information given by the interviewed groups. It seems desirable to extract such a measure from the group responses, in a uniform and reproducible formula.

To do so, we form an index combining a number of indicators from several needs areas. Some of the indicators are existing variables in the dataset; for others, we transform existing variables. Since we do not have a sophisticated theory of disaster behavior, we reduce the contributing information to simple binary indicators. In the demo dataset, the measures extracted from women's focus groups are more suitable than those from the men's. For consistency, we use only women's focus group information.

To distinguish our synthetic severity index from the teams' judgments, we call it a "distress index", but the name is almost arbitrary.

How

- Review the variables sheet for variables potentially suited to yield indicators for a distress index. Arrange them in a table and select those that appear to form a balanced mix with regards to several needs areas and/or institutional domains, with, if possible, more than one indicator by domain. Reorganize the table by domain, give each indicator a name that provides a clear substantive interpretation (e.g., "Food is the highest priority"), and note the short name of the variable on which the indicator is based, and reserve a similar short name to be used when we calculate it as a copy or a transformation of the base variable.

Figure 16: Distress index - overview of indicators

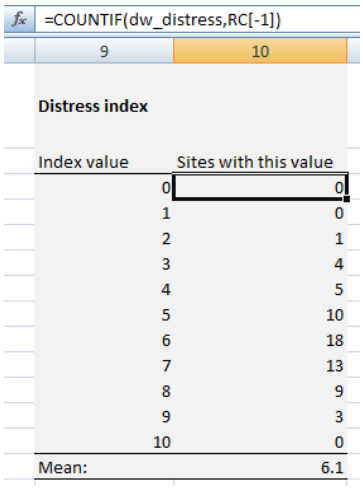
Domain	Indicator	Based on variable	Short name
Survival			
	Food is the highest priority.	w_03_06	dw_03_06
	Livelihoods is the highest priority.	w_03_01	dw_03_01
Health			
	Lack of safe water serious.	w_06_00	dw_06_00
	Diarrhea has increased.	w_26_00	dw_26_00
Safety			
	Lack of safe toilets	w_10_00	dw_10_00
	Women face safety issues.	w_28_01	dw_28_01
Children			
	School attendance is low.	w_18_02	dw_18_02
	Significant child labor.	w_29_00	dw_29_00
Services			
	No electricity	w_04_02	dw_04_02
	Health care facility more than 60 minutes' travel	w_24_02	dw_24_02

- On the right hand side of the data table, mark columns with the labels and short names of the indicators to be created [In our Analysis01_Data, this happens in columns 121 - 130].
- Calculate the indicators, as equal to the existing value of the base variable if no transformation is needed, or by whatever transformation function is appropriate [in the demo, these are listed in column 7 of sheet Analysis02_DistressIndex and, of course, they can be looked up in the indicator variables themselves].

- In another column, mark the label "Distress index" and a suitable short name (we use "dw_distress") and calculate the index as the unweighted sum of the binary indicators (unless there is a strong rationale for weighting them differently) [in column 131 of Analysis01_Data].
- Name the new data columns by the usual "Create from Selection" option and update the Database definition (or create a new name for the grown data table such as "DatabaseCol131" in our case).
- In a new sheet, compute statistics of the distress index, such as in the tables of sheet " Analysis02_DistressIndex":

  ➢ The frequencies of index values (which, theoretically, range from zero to the number of indicators), using the function FREQUENCY or, more simply, COUNTIF.

Figure 17: Distribution of the distress index



=COUNTIF(dw_distress,RC[-1])

	9	10
	Distress index	
	Index value	Sites with this value
	0	0
	1	0
	2	1
	3	4
	4	5
	5	10
	6	18
	7	13
	8	9
	9	3
	10	0
	Mean:	6.1

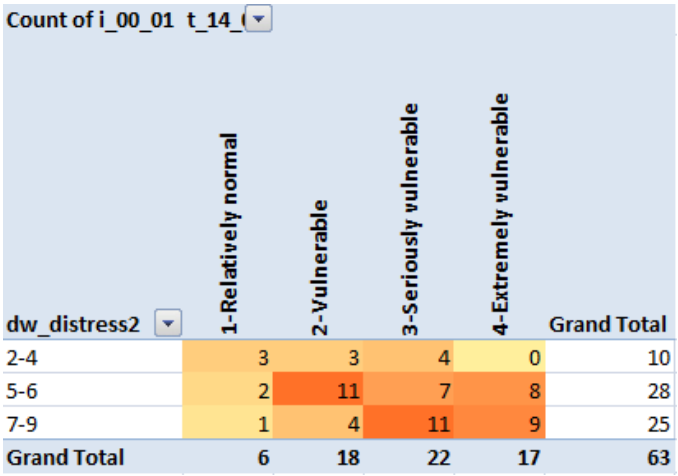
  ➢ In a Pivot table, calculate means by area and living arrangement:

Figure 18: Distress index, by sub-district and living arrangement

Average of dw_	:_00					
District	Upazila (Sub-district)	1-Collective centers	2-Roadside / embankments	3-Marooned	4-Damaged or water-logged	5-Home undamaged	Grand Total
Jessore	Abhynagar			4.0	5.0		4.6
	Jhikargacha	6.0	4.0	6.0	7.0	5.0	5.6
	Keshabpur	7.0	8.0	5.0	6.0	5.3	6.0
	Monirampur		7.0	7.5	6.0		6.8
Kulna	Paikgacha		8.0	6.0		5.0	7.0
Shatkira	Assasuni	6.0	6.0	7.5	7.5	3.0	6.4
	Debhata	6.0	4.0	3.0	6.0	4.0	4.6
	Kalaroa	7.0	6.0	6.0	6.0	3.0	5.6
	Shaktira_Sadar	6.0	7.0	6.0	5.0	7.0	6.3
	Tala	6.5	7.8	7.0	8.0	4.5	6.8
Grand Total		**6.4**	**6.9**	**6.0**	**6.2**	**4.8**	**6.1**

- Cross-tabulate the distress index with the teams' severity judgments, group the index values into meaningful ranges [the short name now appears as "dw_distress2" in the Pivot table], and consider the correlation between the two measures as validation of the ratings by the teams.

Figure 19: Distress index vs. teams' severity ratings



Count of i_00_01 t_14_(

dw_distress2	1-Relatively normal	2-Vulnerable	3-Seriously vulnerable	4-Extremely vulnerable	Grand Total
2-4	3	3	4	0	10
5-6	2	11	7	8	28
7-9	1	4	11	9	25
Grand Total	6	18	22	17	63

Why

We see two motivations to form a distress index or some other functionally similar composite measure:

1. The communities were assessed by different teams. We cannot control for their subjective uses of concepts like vulnerability or severity. An alternative measure, based on information that the people gave, is desirable. Neither this nor the teams' ratings can necessarily claim superior validity, but the strength of correlation between the two measures does say something about the quality of the assessment.
2. A lot of information was collected from the focus groups. Much of it is of sectoral interest. However, intuitively we may assume that, the more severe the situation of a group is, the more likely we will find unmet needs across a wide variety of needs areas. Therefore, statements that the focus groups made in different needs areas have the potential to inform a synthetic severity index (which, to distinguish it from the teams' judgments, we term a "distress index").

As seen in the screenshots above, when forming an index of this kind and for the noted purposes, several calculations are in order:

- **Distribution:** The frequency table for index values relies on COUNTIF.
- **Analysis:** The mean distress score by sub district and living arrangement results from a Pivot table.
- **Validation:** Another Pivot table, grouping the distress scores into three ranges, correlates distress with the team-assessed severity.

We limit ourselves to brief comments on the latter two.

The sub-district table corrects the earlier impression that we formed looking at a table of maximum severity ratings. That one suggested that conditions were worst in Shatkira District. This approach based on direct information from the focus groups paints a more diverse picture. The analyst might note:

> *"The situation of some groups in Jessore and Kulna is similarly bad as in Shatkira. While those living in undamaged homes generally are less distressed, this is not true of such groups inside the town of Shatkira. If we drill down the Pivot table, we will find the record of one community, the Gobordari neighborhood, with seven of its ten indicators checked, among them: food as the highest priority as well as all the health and safety indicators. As is known from many flood disasters, those stranded on roadsides and embankments are, overall, in greatest distress."*

Second, the Pivot table crossing distress score ranges with the team-assessed severity reveals a positive correlation. Excel lets us compute a Pearson correlation (slightly inappropriate because the vulnerability scale is ordinal); the coefficient is 0.37 - a weak correlation, which suggests that the teams synthesized their information in ways that our index does not properly emulate. An alternative explanation is that different teams used different criteria in forming their vulnerability judgments. The fault may lie also with the distress index, whose statistical properties as a scale are difficult to assess with the means of Excel. The best that can be said is that the two measures do not conflict diametrically. Their weak correlation suggests that many ad-hoc measures of disaster impact and of unmet needs may be of limited validity until further study.

6.3.4 Reliability: Concordance between men's and women's priorities

What

By correlating multiple measurements of the same variables, where they exist, we can find out how reliable the concerned data are. However, the idea of the "same variable" is a difficult one, particularly if the measurements are with groups who may apply different interests and perceptions to what the assessment team considers one and the same concept.

This is the case with men and women from the same community, when interviewed separately. Reliability issues mix with genuine differences in the lives of men and women and their impacts on the response to assessments.

Nevertheless, there is merit in correlating or cross-tabulating the response from men and women to the "same" questions for all sites, rather than only comparing overall group averages. We do this for five items in the priority needs question.

How

- In Pivot tables, cross, one item per table, the men's priority scores with the women's, use conditional formatting and observe the degree of concordance. The higher the proportion of sites in cells on or touching the main diagonal, the better men and women agree. An example from sheet Analysis02_Men_Women follows.

Figure 20: Priority scores that male and female groups assigned to "livelihoods"

Sites assessed w_03_01					
m_03_01	0	1	2	3	Grand Total
0	15	1	0	1	17
1	13	3	1	0	17
2	9	4	3	1	17
3	4	3	1	4	12
Grand Total	**41**	**11**	**5**	**6**	**63**

- Create another table comparing the mean priority scores for all items of interest (the values were simply copied from earlier tables in Analysis01_Priorities).

Figure 21: Male-female preference scores for five items

Priorities	Mean Borda count	
	Men	Women
Livelihoods	1.38	0.62
Rebuild homes	0.87	0.98
Food	2.24	2.30
Safe drinking water	0.30	0.68
Short-term financial support	0.30	0.41

- For items with similar mean scores (e.g , rebuild homes, food), check the apparent similarity with the agreement seen in the cross-tables, and interpret.

Why

Separate conversations with men's and women's groups in the same sites offer a reliability test for the measurement of priorities (and for other variables collected in paired occasions). However, as we already stressed, these measurements may not be about the "same thing"; it is difficult to disentangle unreliability from divergence in perception, interests, and position in society.

Nevertheless, such comparisons can be revealing. Take food, for example. As noted earlier, for both men and women, food is the highest priority. In the cross-table for food, we do find that in 24 out of 63 assessed sites, both reported food as their greatest need. But there is little agreement when one or the other gender group assigns a lesser priority. The off-diagonal cells show that men and women in many affected communities diverge in this respect. At the individual site level, the overall degree of agreement, even on a priority need, is not as strong as one might expect.

Such reliability checks in rapid assessments will hardly ever be feasible for all variables. Either multiple measurements are available only for a few variables, or, if most or all have been collected from more than one source, time and interest do not allow it. One would rather do selective checks, on suspicion that some data may be weak, or in hopes that the cross-tabulation will reveal something of substantive interest.

[Sidebar:] Alternating between Excel and STATA

In the preceding note on data preparations, we shared some considerations about using a statistical program in addition to Excel. We exemplified this with the application STATA, which has a similarly strong following as SPSS and R do. Here we share some observations about STATA's value in the analysis phase.

Analysts who are comfortable with STATA may find that they are more productive when they alternate between it and Excel, depending on the type of analysis at hand. There are two motivations for using STATA. One is speed. Particularly for rapid initial exploration, one and two-way frequency tabulations are extremely fast with STATA's ***tab*** (tabulate) command. A second advantage, by no means less important, of working in this application is the availability of statistical analysis forms that Excel, save for installing commercial add-in packages, does not offer. With all advanced analysis forms, however, we have to keep in mind that the data are from a purposive sample, and this in turn limits the validity of probabilistic analyses.

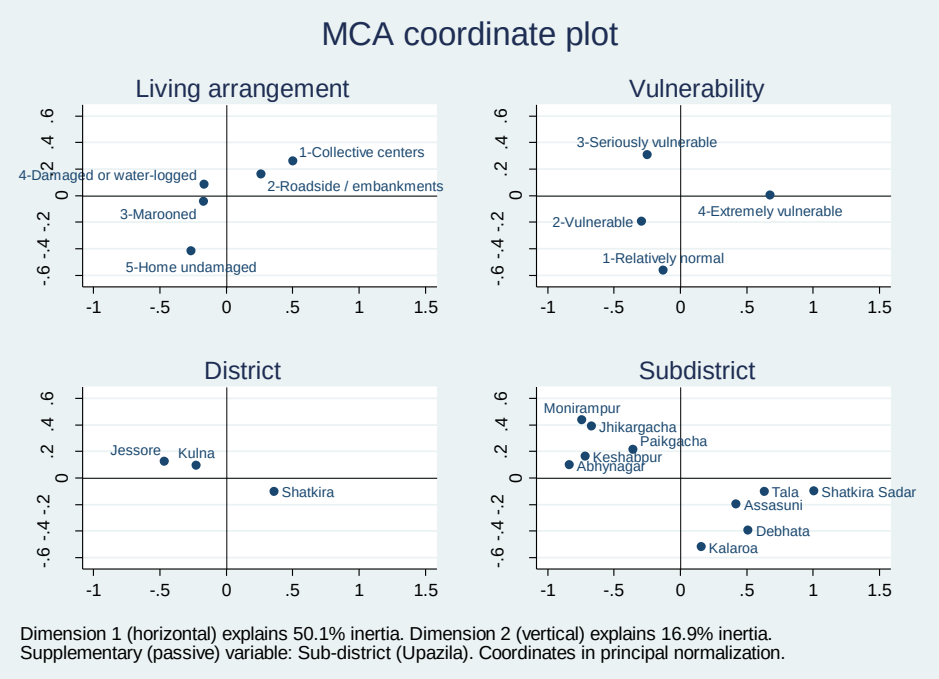
We illustrate the benefit of statistical analyses with one analysis form that may be helpful in many assessments and then mention a few other STATA procedures.

Early on in the analysis, we wish to characterize the sample of assessed communities in some of its basic dimensions, notably region, affected groups, severity of their current situation. These arrive as categorical variables, some unordered (e.g. the regions), others with a built-in order (e.g. severity judgments on an ordinal scale). In Excel, we are limited to inspecting two-way tables (produced as Pivot tables) that circle through the variables of interest. Thus we might study tables for: Regions x affected groups, Affected groups x severity, and Severity x regions. STATA lets us see the interactions of more than two categorical variables at a time.

The procedure that achieves that is known as Multiple Correspondence Analysis (MCA) (Wikipedia 2013a, 2013b). MCA creates an n-dimensional space of continuous variables into which every category of every included variable is projected. It proceeds such that the category locations on the first dimension account for the relatively largest share of the variability in the data (the variability in MCA is known as "inertia"). Those on the second dimension account for the second largest share, etc. Although in this space coordinates can be computed for each observation (assessed site), the major interest is with the proximity of the categories from the different variables. The proximities are usually visualized on the plane defined by the first two dimensions.

The MCA plot below does this for three variables of the Bangladesh dataset - districts, living arrangement, and the team-assessed vulnerability level. For illustration, a plot of the sub-districts has been added; the sub-district coordinates are passive, meaning they are the mean coordinates on the first two dimensions of the concerned sites already calculated from the three active variables.

Figure 22: Multiple correspondence analysis, coordinate plot, in STATA



Dimension 1 (horizontal) explains 50.1% inertia. Dimension 2 (vertical) explains 16.9% inertia.
Supplementary (passive) variable: Sub-district (Upazila). Coordinates in principal normalization.

A number of findings leap to the eye:

1. We basically deal with a **one-dimensional solution** (given 5 + 4 + 3 = 12 categories in three variables, dimension 1 accounts for a handsome 50 percent of the total inertia; moreover, dimension 2 explains only a third of that).

2. There are a number of **sharp distinctions** along the first dimension:

   - Flood victims living in collective centers, on roadsides and on embankments score considerably higher than those in other living arrangements.
   - These two groups are closest to extreme vulnerability. This level is sharply offset from the other three vulnerability levels; on this first dimension it is "extremely vulnerable" vs. the rest. This raises the question whether the assessment teams were capable of meaningfully distinguishing more than two levels of vulnerability.
   - In regional terms, the situation is worst in Shatkira District, which occupies a location close to "collective centers", "roadside / embankments", and "extremely vulnerable". All the sub-districts in the lower right quadrant belong to Shatkira; they are clearly set apart (on both dimensions) from the sub districts of Jessore and Kulna; there is no overlap.

The point to emphasize here is the rapid visualization of interactions among several variables.

STATA's ***mca*** command is only one instance of rapidly executed procedures that can help the hard-pressed analyst explore assessment data. The choice of other procedures depends on data constellations, analytic interest and likely sensitivity to the particular sampling used. Purely descriptive procedures are robust and should be used without hesitation. ***table*** is STATA's closest relative to Excel's Pivot table; it can return also cell medians and percentiles, something which is beyond Excel. ***tabm*** is a user-written command for the tabulation of multiple variables. It achieves the same results as the Excel procedure described on page 20 - it is faster to calculate, but tedious to format for reporting. ***multitab*** does a great job on multiple-response

data, particularly of the inconvenient polytomous kind. It comes with numerous options to address these situations fast and flexibly; analysts eager to study up on the further complexities of processing such data (and visualizing them in graphs) may want to consult this article (Jann 2005).

Finally, we should not forget that even simple charts - e.g. a bar graph of frequencies - can be made faster in STATA than in Excel - barring extensive editing, which can be time-consuming and should be reserved for the few that will make it into the report. But STATA is not in all situations more efficient than Excel. If you expect some data to be corrected pending further checks and inquiries, it may be easier and safer to work in Excel. Late corrections and additions may affect recoding decisions; as long as the formulas are in place, secondary recoding in Excel should be painless.

Analysts employing STATA may gain time and additional insight over mere Excel users. They may not always meet with understanding and acceptance for its outputs and may have to do some extra explaining. Also, in the interest of reproducibility, STATA work should always be documented in log files, and these backed-up and shared as text files.

7. Conclusion

Needs assessments vary greatly in context, design and data. It is challenging to work out generic guidance for how to approach a dataset for analysis. This note has followed two lines of development:

1. A number of data constellations - such as multiple-response item variables - recur in many or most assessments. For the processing of these, we proposed some sets of instructions. These have to be taken with, not one, but two grains of salt. Users may find better solutions for a given problem. Some may even classify these situations differently and create toolboxes altogether different from the set of features and formulas that we marshaled for the purpose.

2. A second approach is more substantive in nature. In this, the analyst, after minimal preparation, looks at the given dataset and asks: "How much music is in it? How much will it tell us? What are the core models that I can execute with these variables? Which variables are critical, which will foreseeably play a minor part?"

This kind of thinking cannot be fully demonstrated with just one dataset. With the Bangladesh data in hand, we identified a small number of analysis forms that might prove of interest elsewhere, substantively and/or technically. Limited to the means of Excel, we demonstrated methods for a rapid overview in terms of severity, the variations of needs as a function of severity, and a synthetic measure of distress that provides an alternative to the judgments that the assessment teams formed on the severity of local situations.

The distress index illustrates a dilemma in assessments. We can approach the data in two philosophical moods. We can take them literally. When people say "Food is our first priority", they mean it - and who are we to doubt it? The literal interpretation of such data is the only feasible approach for sectoral planning. If food aid is to be considered, data on food (or its absence) is needed - in greater detail and by probing deeply, but always with food in mind.

The modest agreement between men and women even on this priority need sows a doubt. Maybe one cannot take all these statements literally. Some may mean something else. Community groups respond to information requests in a hectic atmosphere, on top of the already stressful post-disaster life. They need to agree, in short moments and with little clarity about what will happen to their information, on factual as well as value judgments (priority need statements imply values).

In this view, what the assessment team hears is a sampling of concepts that express underlying dispositions of a broader, and partially hidden, nature. There may be a cluster of deprivation to which food and safe water are closely associated, and another in which house repairs and the safety of property coalesce. The items that the community groups mention may be but varying manifestations of these deeper dispositions.

Purposive sampling, time pressure, the unavailability of better statistical models - these all limit the investigation of a deeper structure. But, supposing it exists, it should make us accept more easily that the observed data are of modest reliability, and the measures we construct on their basis have modest validity. Such humility may make the proposed analysis forms seem more adequate - as long as the readers feel they can handle them and apply them within the time limits of the analysis phase.

8. References

Borda, J. C. (1781). "Mémoire sur les élections au scrutin." Histoire de l'Académie royale des sciences **2**: 85.

Jann, B. (2005). "Tabulation of multiple responses." Stata Journal **5**(1): 92-122.

Lansdowne, Z. F. (1996). "Ordinal ranking methods for multicriterion decision making." Naval Research Logistics (NRL) **43**(5): 613-627.

Wikipedia. (2011)."Borda count." Retrieved 28 March 2011, from http://en.wikipedia.org/wiki/Borda_count.

Wikipedia. (2013a)."Correspondence analysis." Retrieved 19 March 2013, from http://en.wikipedia.org/wiki/Correspondence_analysis.

Wikipedia. (2013b)."Multiple correspondence analysis." Retrieved 19 March 2013, from http://en.wikipedia.org/wiki/Multiple_correspondence_analysis.