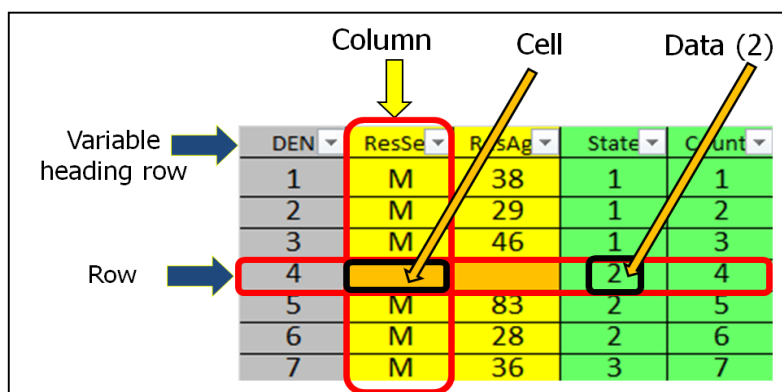


FAO Standard Seed Security Assessment

CREATING MS EXCEL DATABASE

When you open a Microsoft Excel programme, a new file (book1) appears on your screen. This file normally consist of three work sheets (new work sheet can always be added). Every work sheet consists of **Columns** and **Rows**, and the intersections between the columns and rows are the **Cells** (Fig 4.1). Cells are points within a sheet where data (variables) are entered. Questionnaires normally contain a number of variables with volumes of data which are entered in to cells under each variable column.



The diagram illustrates an MS Excel sheet with the following structure:

Variable heading row	DEN	ResSe	RsAg	State	Count
1		M	38	1	1
2		M	29	1	2
3		M	46	1	3
4				2	4
5		M	83	2	5
6		M	28	2	6
7		M	36	3	7

Annotations in the diagram:

- Column:** Points to the 'ResSe' column header.
- Cell:** Points to the cell containing 'M' in row 1, column 'ResSe'.
- Data (2):** Points to the cell containing '2' in row 4, column 'State'.
- Variable heading row:** Points to the first row of the table (row 1).
- Row:** Points to the fourth row of the table (row 4).

Fig. 4.1 . MS Excel sheet

In designing a database one row is normally dedicated for defining the variable headings, where each variable is defined in only one column and the cells are below this are used for entering data corresponding to the variable headings. On the other hand, one column (normally the beginning on) is used to define data entry number. As a general rule of thumb;

- Column** - Variables are normally coded as headings of columns in variable heading row. A column is normally used for only one variable eg. Sex, Age, Income etc.
- Row (s)** – All rows below the variable headings row could be used for entering data. The information from a single questionnaire is entered into row(s) **corresponding** to the questionnaire **data entry number (DEN)**. A single questionnaire should never be assigned more than one data entry number, also no two or more questionnaires should have same data entry number assigned to them. One or two rows above the variable heading row could be used to enter the question numbers corresponding to the variables (see Fig. 4.2).
- Cells** - One cell, one response

	Qn.1.1	Qn.1.2	Qn.2.1	Qn.2.2	Qn.2.3	3.1	3.2
DEN	County	Sub-coun	Age.grou	Res.Gend	Educatio	HH_Gend	HH_Size
1	Kitui	Kitui Central	2	m	3	m	3
2	Kitui	Kitui Central	3	f	2	m	7
3	Kitui	Kitui Central	3	f	2	f	6
4	Kitui	Katulani	3	f	2	f	5
5	Kitui	Katulani	1	f	3	m	8
6	Kitui	Katulani	4	f	1	f	3
7	Makueni	Mbooni	4	m	1	m	13
8	Makueni	Mbooni	2	f	3	m	7
9	Makueni	Mbooni	3	m	3	m	4
10	Makueni	Mbooni	3	m	4	m	5
11	Makueni	Kibwezi	3	f	2	m	5
12	Makueni	Kibwezi	4	f	1	f	5

Fig. 4.2

4.6.1 Defining variables and coding responses

As mentioned earlier, variables from a questionnaire are defined in the **variable heading rows**. When defining variable headings, only continuous characters are used i.e. no space between characters. For a single response such as sex of head of households (HH) the variable could be defined as **Sex.HH** or **Sex_HH** and not Sex HH.

For multiple response questions such as – What crops did you plant last season? Here, each possible response (e.g. sorghum, maize, beans) is a variable – and thus occupies a column - within which a value is typed, in this case yes (1) or No (0).

Responses in a questionnaires are sometimes given code for example reason for planting less area of land – 1=lack of access to land; 2=lack of seed; 3= sickens ----- and others (specified) in the questionnaires. Those others have to be given codes as well.

SSA data entry clerks are strongly advised to make use of the database structure which has been designed by the FAO SSA development team.

4.6.2 Cell validation and data entry

To minimize errors during data entry, certain variable columns should be validated to restrict entering unexpected data or certain characters and/or range (Read more MS Excel application).

Before any data entry begins a questionnaire is given a number and no two questionnaires should have the same number. The existence of a questionnaire number makes strong link between the computer (soft) copy and the paper form and will be useful in the data cleaning stage. Different variables have different data types depending on the type of variable.

- Discontinuous (categorical, classificatory, discrete) variables: variables that cannot be divided into fractions or take finite numbers, e.g. gender (male or female), livestock presence (Yes or No); residential status (Resident, IDP, Refugee, Returnee). These variables can be represented by text e.g. initials (m for male and f for female) or could be assigned numeric codes such as 1=Yes, 0= No; 1=Resident , 2=IDP, 3=Refugee and, 4=Returnee

- Continuous variables – variables that can be divided into fractions or take infinite number of values e.g. Income, temperature, age, area planted, seed quantities, production, yields. These variables are entered as numbers with no unit of measurement attached e.g. for 10kg of seed planted, the quantity of seed planted is normally entered as 10 and not 10kg.
- **Note to the data clerk** - Units for measuring continuous variable may vary from individual to individual or from place to place. This must be standardized before or during data entry.

In order to ensure quality data is entered into the data base, the team leader should put extra efforts to supervise data entry clerks. In controlling the data entry, the team leader should randomly sample questionnaires that have been entered by the data clerk and verify them using the data entry numbers in data base to check if they have been entered correctly. Data entry control is normally done at the end of each day during the data entry process. This process should be done together with the data entry staff in order for him/her sees the mistakes made and build up his/her awareness where to take more care.

Data-entry mistakes and how to correct them:

Codification or simple entry mistakes should be corrected immediately according to the information in the questionnaire. These mistakes are more common in the first days or when the work is done in a rush, but should reduce with time. If a higher frequency of such kind of mistakes is recorded; a higher number of questionnaires should be verified.

Another common data entry mistake is a shift in the columns of the data entries, as either one column was skipped somewhere or entered too early. In these cases, the whole questionnaire should be entered again. One way to minimize column related data entry error is to use different colors for different columns according to the corresponding section of the questionnaire. This allows the data entry clerk to relate the position in the questionnaire to the position in the database.

The importance of accuracy in data entry cannot be overstated, as the correction process can take a lot of work and time.

4.6.2 Data cleaning and verifications

Errors can be introduced during data collection as well as data entry. Before deriving any additional variable or running data analysis, the data manager has to ensure the data is devoid of errors or outliers. In data verification and cleaning, all variables are checked to ensure that there is no error, inconsistent data or outlier. Box 4.3 gives an example which could either be an error or an outlier. Any inconsistent entries or outliers (box 4.3) have to be verified by checking the hard copy of the questionnaire and corrected or appropriate decision made on outliers.

Box 4.3. Example of an error or outliers - The data clerk entered the following quantity (kg) of sesame seed planted by 10 households; 2, 3.5, 45, 2, 4, 1, 1.5, 2, 5.5 and 2.3. There are two possibilities in this data.

- a) An error introduced by the data clerk during data entry where he/she presses key 4 & 5 simultaneously when he wanted to enter 4 or 5, or fails to press a decimal point well when entering 4.5. This can be corrected by checking the hard copy of the questionnaire and correcting the entry.
- b) An outlier – if after checking the questionnaire the number **45** is found to be a correct entry, a confirmation can be made by calling the enumerator (if he can still remember) of the farmers (if the telephone contact is available). If after checking from all these sources and the number (45) is found to be true, then this could be an outlier which could significantly influence the result of the analysis.
 - In the above data set, when analysis of average quantity of seed planted is done with such outlier, the average goes to 6.9, and when such outlier is omitted from the data set, the average comes down to 2.6.

NOTE: Outliers could be correct data but they deviate from the normal distribution. Statistically they significantly influence the result of analysis

Data cleaning is a tedious process that requires patience and time but it should never be skipped. There are two different levels or ways of ensuring the accuracy of the data entered and obtained. The **first level** is filtering the data using the Excel filtering function.

Activation of filter and cleaning of data take the following simple but logical steps after all the data have been entered.

- 1) Highlight all the **Variable Headings**
- 2) Go to **Data** menu and click on **Sort & Filter** icon. Drop down menu will appear on the right side of every variable heading.
- 3) Click on the drop down icon and scan for any inconsistent data or outlier within the list you see.
- 4) Once you identified inconsistent data or outlier, first **De-select All**, and then **Select** the inconsistent or outlier data. Click **OK**. Only selected one(s) will appear on the screen
- 5) Check the **Data Entry Number(s)** corresponding to inconsistent or outlier data identified, Go back to the **Hard Copy** of the questionnaire and **Correct**.
- 6) Where the inconsistent or the outlier is existing in the hard copy, **Consult** the enumerator or team leader for correction.
- 7) In the event that neither the hard copy nor the enumerator/team leader can help, the data manager will have to make judgment to **Omit (Delete)** the inconsistent or outlier data if it will affect the final analysis.

The second level is to verify the consistency in the link between two related variables, for example,

- A household cannot have more land cultivated with the different crops in a specific season than the total available land for the same season
- A household has no cash savings made in a season but the corresponding variable showed the institution where the cash has been saved.

These errors can be avoided by programming failsafe parameters into the database, so that it would be impossible to enter data that is not consistent. Failing this, the most suitable correction process is to verify with the corresponding questionnaire or through logical deduction.

4.6.3 Deriving variables

Certain variables such as yields, seed rates, multiplication rates, animal units are normally not collected directly using the questionnaires but are derived from two or more variables. For example, yield which is the quantity harvested per unit area of land is derived from quantity harvested from a given area planted by the farmers. For any additional derived variable, a column has to be inserted and appropriate calculation be done. The most suitable and easy way for this process is to use formulas. Once the calculations have been made, you have to verify the results, as sometime the results are not correct where data is missing or a number is divided by zero, which is shown with the following symbol: **#DIV/0!** in the database. Those entries have to be deleted before proceeding with the data analysis.