Get every one in the picture

Data quality 1: Individual records

Data analysis and Report writing workshop for Civil registration and vital statistics data.

UNITED NATIONS ESCAP
Economic and Social Commission for Asia and the Pacific

Adapted from Pacific Community's
Data analysis and report writing
Workshop for the North Pacific

The importance of data cleaning

- Poor data can lead to misleading analysis and subsequently poor decisions and policies
- If we produce data that is not reliable – decision makers will not want to use our data, even if it is available
- Need to establish **TRUST** in our data
  - That does not mean it needs to be perfect.
  - It does mean that it should be the **best of what we have available**, and that we **need to be honest about its limitations**
- Information about data quality should be published alongside our data and findings.

Reviewing data quality should be continual

- Review the systems to make sure that they are collecting data in the best possible way
- Review individual records (unit record data)
- Review tabulated data before further analysis
- Review the plausibility of calculated measures
- Compare the measures to other sources of information
- Publish findings for scrutiny by others

Reviewing unit record data

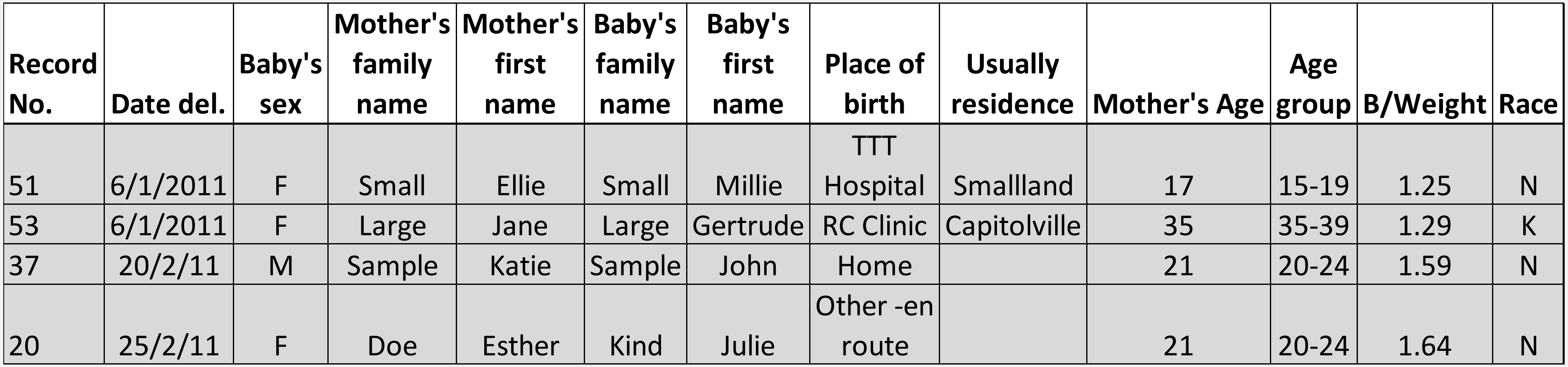
- Checks to ensure:
  - required data fields have been carried over into our working spreadsheet
  - records have been carried over into our working spreadsheet
  - duplicate records have been removed
  - inappropriate records have been excluded (for example, still births have been removed from your sheet on live births)
  - records use variables which are consistent and can therefore be readily aggregated
  - missing values are minimized wherever possible

Setting up unit record data

one record per row

and

one field per column



Record No.	Date del.	Baby's sex	Mother's family name	Mother's first name	Baby's family name	Baby's first name	Place of birth	Usually residence	Mother's Age	Age group	B/Weight	Race
51	6/1/2011	F	Small	Ellie	Small	Millie	TTT Hospital	Smallland	17	15-19	1.25	N
53	6/1/2011	F	Large	Jane	Large	Gertrude	RC Clinic	Capitolville	35	35-39	1.29	K
37	20/2/11	M	Sample	Katie	Sample	John	Home		21	20-24	1.59	N
20	25/2/11	F	Doe	Esther	Kind	Julie	Other -en route		21	20-24	1.64	N

Variables or fields required

Births

- First name
- Surname
- Date of Birth
- Sex
- Place of Birth (Hospital, Village, Island)
- Place of Residence (Village, Province, Island)
- Mothers first name
- Mothers surname
- Mothers age
- Live or still birth (or all live births)
- Birth weight (optional)
- Length (optional)
- Weeks gestation (optional)
- Ethnicity (optional)

Deaths

- First name
- Surname
- Date of Birth
- Sex
- Date of death
- Age (use separate fields for days, months, and years)
- Place of Death (Hospital, Village, Province, Island)
- Place of Residence (Village, Province, Island)
- Spouse’s first name
- Spouse’s surname
- Causes of death (by line of death certificate – 1 variable per line)
- Underlying cause of death (more on this in chapter 16)
- External cause (if applicable)
- Occupation (optional)
- Ethnicity (optional)

Check that records have been carried over into your working spreadsheet

- Easy when extracting data that not all records are transferred
- Check the totals against the original source (such as the database)
- Look at the total number of records and make sure that it is within an expected range

Remove Duplicate Records

- Need to find and remove duplicate records
- Questions to ask -
  - How do we know if it is a duplicate? Do all fields have to be an exact match?

Name	Surname	Sex	DOB	DoD	Place of death	Residence	Province
April	Jones	F	April 2013	5/5/2013	Hospital	Noumea	South
Baby	Jones	F	4/4/2013	6/5/2013	Hospital	Noumea	South

Is this the same person?

  - Which record will we use if the data is not exactly the same
  - Be careful when checking for duplicates that you don't remove twins.

Data Matching – example from Tonga

- For deaths to be considered “matched” they must match on 3 of the following criteria (if surname included) or 4 if not.
  - Surname (similar spelling or sound OK)
  - Date or Death/Month of Report (same month)
  - First name (similar spelling or sound OK)
  - Island (place of death or report or residence)
  - Age at Death (within 1 year)

- Some possibility of under-matching when data quality poor (i.e. Insufficient data to match criteria)

Inappropriate records have been excluded

- Stillbirths should be in a different file (not part of live births or deaths)

- These are important events, but should be analysed as a separate category

Has data been consistently recorded?

- Variables should have been entered in a consistent manner – but this is not always the case, especially when using older data
  - In best practice – these should be controlled by your metadata standards
  - Good database design can minimize a lot of problems with different formats being used (through drop down or selection boxes or entry rules)
- Need to be the same format for pivot tables to work (for tabulation)
- Common problems
  - Sex- if we are using M/F for sex, then all records should have one of these values in the field, rather than some having recorded as male, Male, or 1 etc
  - Dates

Missing values are minimized wherever possible

- Is there original data missing?
- Can we obtain this information by looking up another source?
- Can we impute this information from our existing data?
  - for example, if there is no age recorded for the mother in a birth record, but we have her date of birth, an age could be calculated and inputted into the records
  - Is the name provided specific to one sex?
  - Is there an obvious typo?

Derived variables

- To make our analysis easier
- YEAR
  - Tabulations are easier when we have all the years of our data in the same worksheet.
  - But need to make sure we can still separate data
    - (surprising how often date is left out of the records and can only be determined by what file it is in)
  - can be derived from the using the formula **=YEAR( )**
- AGE GROUP
  - Analysis will be done by age group rather than individual years of age

Age groups for collation of deaths

Neonatal deaths (under 28 days)
28 days to <1 year
1 to 4 years
5-9
10-14
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75+
(and unknown)

Basic tabulations

BIRTHS

- Total Number of births by year
- Total Births by year by sex
- Number of births by year, by age of mother (5 year age groups)
- Number of births by geographic sub-region (where relevant) (and potentially by sex and age of mother if there is sufficient data)
- Number of births by major ethnicity (where relevant)

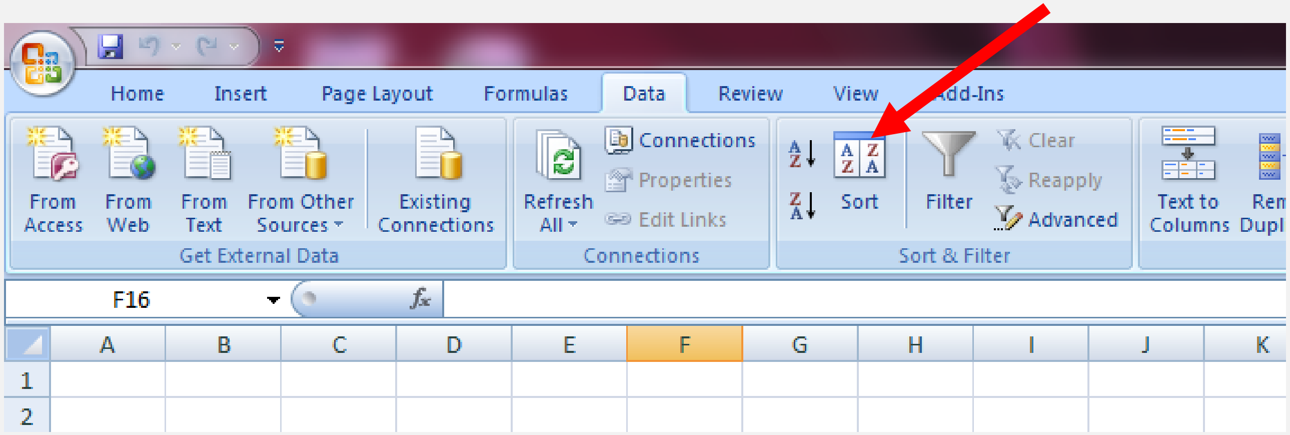
Basic tabulations

DEATHS

- Total number of deaths
- Number of deaths by sex and age groups
- Number of deaths by sex, major ethnicity (where relevant), and age for age group
- Number of deaths by geographic region (where relevant) by sex and age group
- Number of neonatal deaths (deaths in infants aged 28 days or less)
- Number of deaths by age (for ages <1, 1-4, 5-9, 10-14...........65-59, 70-74, 75+ years), sex, and underlying cause of death (according to the ICDv10 103 cause – General Mortality list 1).

Sort function in Excel

- The most important tool when using excel to clean the data is the **sort function** which appears under the **data tab** in Excel 2007. By clicking on the button marked, you can sort highlighted text by any of the fields in your data set.



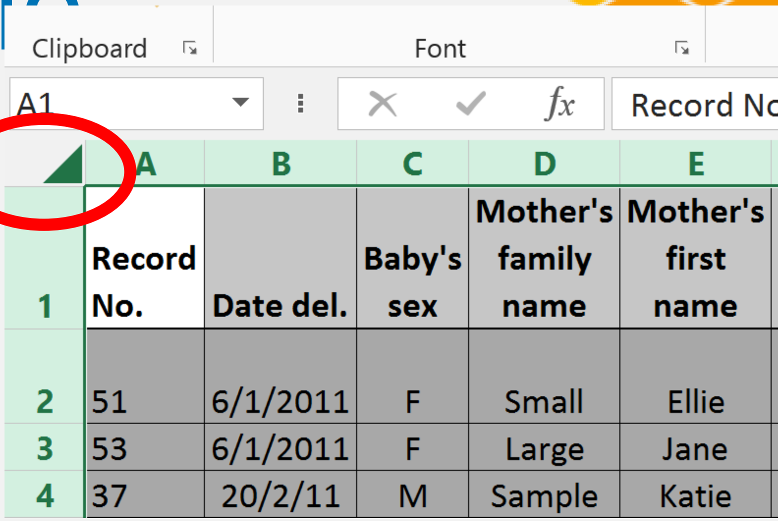
Sort function in Excel

Important points to remember when sorting data:

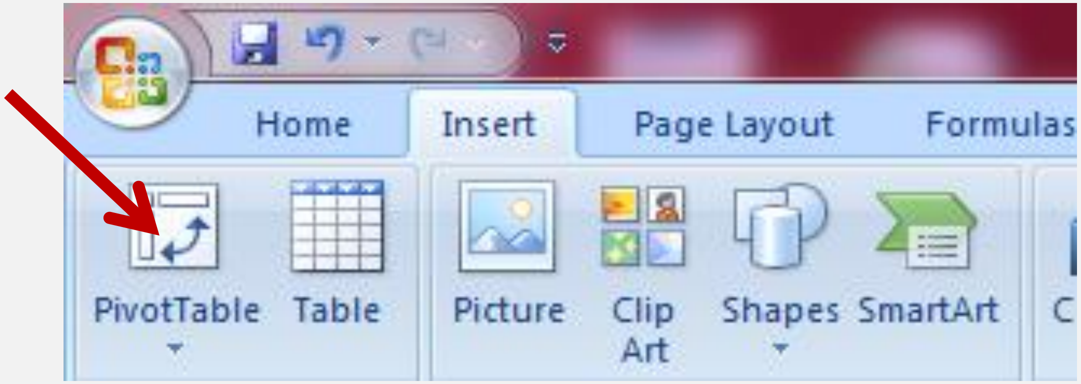
- Always set your data up with **one record per row and one field per column**, otherwise data will become separated and will not be able to be analysed.
- When selecting data to sort – it is good practice to always **select ALL data** by clicking on the arrow in the left upper corner (between the A and 1). Otherwise some columns may be sorted while data in others will stay in the same order, creating chaos in your records.
- Make sure there are **no blank columns or rows** which may interrupt the sort function
- **Label your fields** in the top row and make sure these are not repeated later in the data set. You can then sort by headings by ticking on the box that says “my data has header rows” in the sort dialogue box.

Creating a pivot table

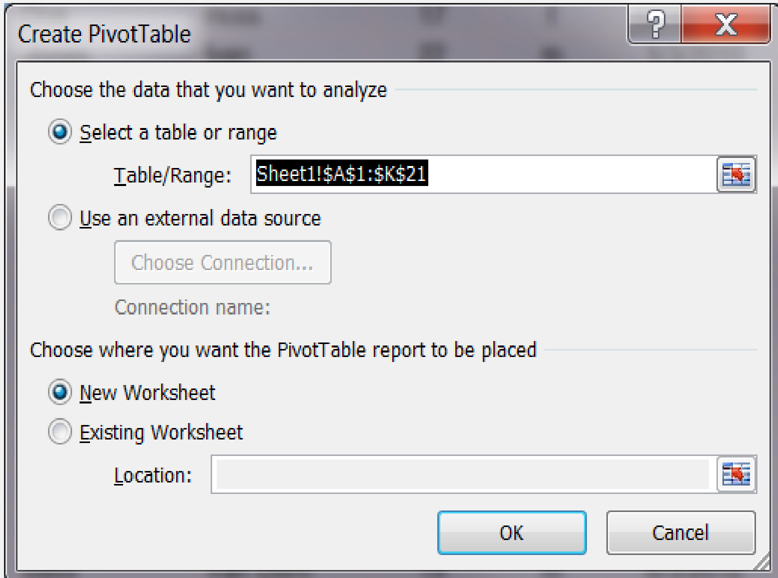
- Highlight the data sheet by clicking in the top left corner (between A and 1)
- Go to the insert tab. On the far left is a button that will allow you to create a pivot table from your selected data.



	A	B	C	D	E
1	Record No.	Date del.	Baby's sex	Mother's family name	Mother's first name
2	51	6/1/2011	F	Small	Ellie
3	53	6/1/2011	F	Large	Jane
4	37	20/2/11	M	Sample	Katie



- Always choose the option of creating the table on a new sheet so as not to confuse your source data.


Create PivotTable

Choose the data that you want to analyze

● Select a table or range

Table/Range: Sheet1!$A$1:$K$21

○ Use an external data source

Choose Connection...

Connection name:

Choose where you want the PivotTable report to be placed

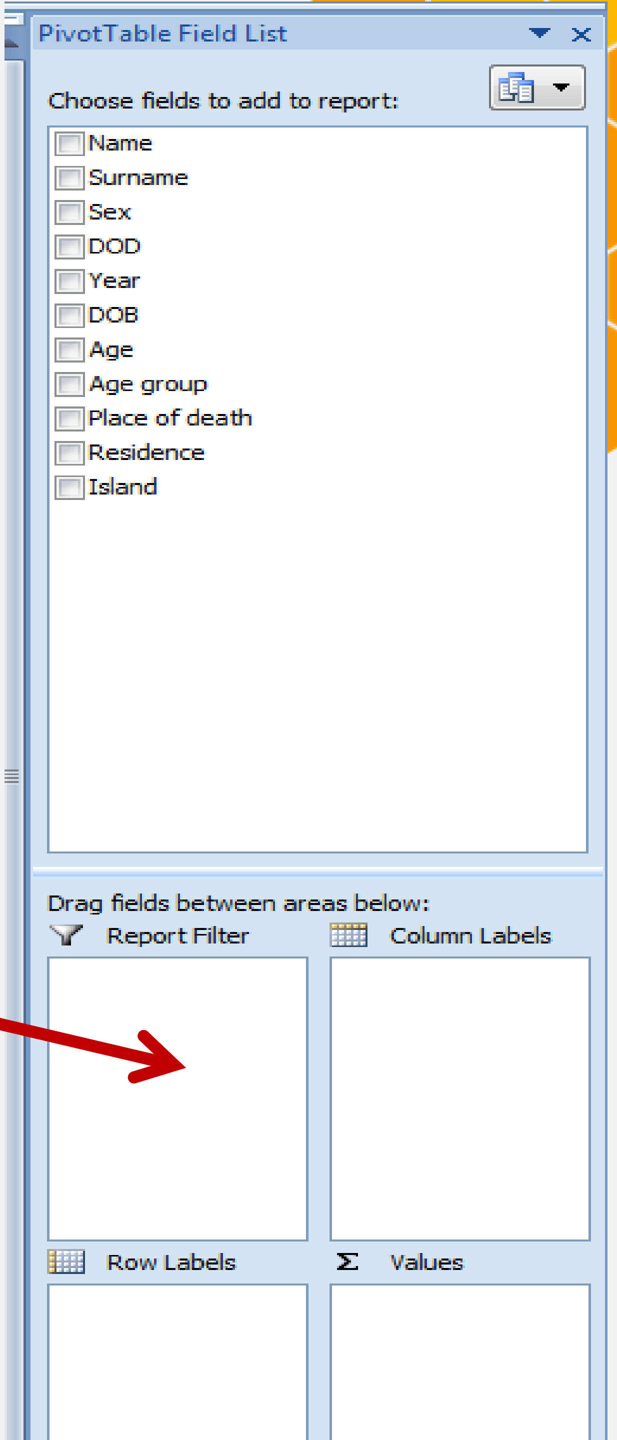
● New Worksheet

○ Existing Worksheet

Location:

Creating a pivot table

- You can then specify which variables to use as columns and rows to tabulate your data by moving them into the appropriate place.
- Use the count of function, and a variable which has no blanks to populate your table.
- Once a table is set up the way you want, copy it and paste it into a new worksheet, as pivot tables cannot be locked.



PivotTable Field List

Choose fields to add to report:

□ Name
□ Surname
□ Sex
□ DOD
□ Year
□ DOB
□ Age
□ Age group
□ Place of death
□ Residence
□ Island

Drag fields between areas below:

Now you try…

- Sort your data by key variables. Do you see any problems with birth data for:
  - Sex of the baby (missing values?)
  - Birth weight
  - Age group of mother
  - Year of birth
- Check your death data for
  - Sex
  - Age group
  - Year
- Correct any erroneous values (except age – we will do this in the next exercise)
- Time permitting, create the basic birth tables as shown in chapter 4 of your VS report template