Get
every one
in the picture

# Data quality 2: Tabulated data

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

# Data quality of tabulated data

- **Consistency**
- **Coverage (or scope)**
- **Representativeness**
- **Completeness**
- **Validity**
- **Reliability**
- **Bias**

# Data quality of tabulated data

- Once our unit record data is as good as it can be, we will make judgments about the quality and reliability of our tabulated data.
- Identify the most important sources of error and provide quantitative measures where possible or qualitative descriptions otherwise.
- Report specific sources of error or bias to inform readers.

# Data quality of tabulated data - consistency

- **Consistency -** a description of the data over time, and whether it follows a similar pattern from year to year (month to month) etc;
  - Are there significant gaps or peaks in our data set?
  - Best assessed from tabulations or graphs of the total number of events by year or month.
  - Good practice to examine sub-regions, as well, to identify any reporting problems in the data.

# Data quality of tabulated data - consistency

- **Consistency -** also known as ***Comparability over time***
  - it may be appropriate to discuss comparability of the same activity for a previous reference period, especially if there has been a change in methodology, concepts or definitions.

- The effects of benchmarking or revisions on comparability over time should be described
  - (for example, with cause of death data, a change between ICD 9 and ICD 10 would be expected to have an impact on the data consistency when tabulated by cause).

# Data quality of tabulated data - Coverage

- **Coverage** - describes the area or population that the data set includes, noting any groups of events that may be missing.

- For example, coverage of a CRVS system may be national, or it may in practice exclude remote and rural areas, births or deaths overseas, or events related to foreign nationals or non-residents.

# Data quality of tabulated data - **Representativeness**

- **Representativeness** is how well the data you collected reflects the broader population for which you want to use the results.
  - Related to coverage, was the entire population included?
  - Very important for demographic surveillance sites and for survey design.
    - For example, if you interviewed only school teachers on their nutrition and eating habits, this well educated, employed group of people would not necessarily be very representative of nutrition and eating habits at a national level.

# Data quality of tabulated data - **Completeness**

- **Completeness** assesses what proportion of the events that we intended to capture did we manage to collect data for.
  - For example, if our area of coverage is births in the national hospital, what proportion of the births in the hospital were recorded in our data set?

- Generally 80% completeness for CRVS can be used for analysis without adjustment (although the completeness should be reported for context).

# Data quality of tabulated data - **Completeness**

- **Completeness** may also be used to refer to the completeness of key fields within the data set.
    - For example, the proportion of births where the mother's age was reported.

- The effect of ***editing and imputation*** on the quality of data should be assessed and described.

# Data quality of tabulated data - **Validity**

- **Validity -** the plausibility of our raw data (in terms of number of events) and of calculated measurements.
  - For example, it is generally implausible that the infant mortality rate would be substantially different for males and females.

- One way of assessing validity is through ***Comparability with other data sources:*** if similar data from other sources exist they should be identified.
  - Do your results match those from the census? From household surveys? How different are they?
  - Where appropriate, a reconciliation should be attempted describing how the data sets differ and the reasons for these differences.

# Data quality of tabulated data - **Reliability**

- **Reliability –** Is your CRVS system able to produce results of a similar quality over time?

# Data quality of tabulated data - **Bias**

- **Bias** is a systematic effect on a statistic or measurement rather than a stochastic or random one.

- Generally related to some aspect of the data collection which results in us being more likely to see particular answers in the data set over others.
  - Some individuals being more likely to be recorded than others
  - Some COD are easier to diagnose than others
- Reporting bias – making some types of data more available than others, highlighting certain findings over others – hard to avoid!

# Re-distributing deaths by age

- When death records are missing age, we need to estimate age at time of death. (This also applies to births where age of mother is unknown.)
- What would happen if we did not assign ages to each birth and death record? Would our fertility rates be accurate?
- Use the age distribution of deaths with known ages to determine how many of our unknown aged deaths should end up in each age group.
- As age patterns are different for males and females, the re-distribution of these deaths should be done separately by sex.

# Re-distributing deaths by age

- Start by setting up a table of deaths by age group and sex for the year(s) where data are missing.
- For the deaths for which age is known, calculate the percent distribution of these deaths by age group for each sex separately
- Multiply the percent for each age group from this distribution to the total number of deaths (including deaths of unknown age) to get the revised number of deaths by age.
- Round your results to the nearest whole person (after all – we don't get part of a person dying!).
- This method can also be used to re-distribute births by age of mother.

For the deaths for which age is known, calculate the percent distribution of these deaths by age group for each sex separately

| Age | Total deaths | | Percentage of total excluding unknown ages (%) | | Re-distributed deaths by age | |
|---|---|---|---|---|---|---|
| | M | F | M | F | M | F |
| <1 year | 14 | 12 | 3.8 | 3.9 | 15 | 13 |
| 1-4 | 6 | 4 | 1.6 | 1.3 | 7 | 4 |
| 5-9 | 2 | 1 | 0.5 | 0.3 | 2 | 1 |
| 10-14 | 1 | 4 | 0.3 | 1.3 | 1 | 4 |
| 15-19 | 5 | 6 | 1.4 | 2.0 | 5 | 6 |
| 20-24 | 9 | 13 | 2.5 | 4.3 | 10 | 14 |
| 25-29 | 16 | 12 | 4.4 | 3.9 | 17 | 13 |
| 30-34 | 23 | 12 | 6.3 | 3.9 | 25 | 13 |
| 35-39 | 25 | 14 | 6.8 | 4.6 | 27 | 15 |
| 40-44 | 22 | 15 | 6.0 | 4.9 | 24 | 16 |
| 45-49 | 26 | 22 | 7.1 | 7.2 | 28 | 24 |
| 50-54 | 35 | 26 | 9.6 | 8.5 | 38 | 28 |
| 55-59 | 38 | 28 | 10.4 | 9.2 | 41 | 30 |
| 60-64 | 48 | 32 | 13.1 | 10.5 | 52 | 35 |
| 65-69 | 58 | 44 | 15.8 | 14.4 | 63 | 47 |
| 70-74 | 36 | 36 | 9.8 | 11.8 | 39 | 39 |
| 75+ | 2 | 24 | 0.5 | 7.9 | 2 | 26 |
| Unknown | 33 | 24 | | | | |
| TOTAL | 399 | 329 | 100.0 | 100.0 | 399 | 329 |

Apply this percentage to all deaths (including deaths with unknown age)

$$= \frac{14 \text{ deaths}}{(399-33) \text{ deaths}} \times 100$$

$$= \frac{14 \text{ deaths}}{366 \text{ deaths}} \times 100$$

= 3.8%

$$= \frac{3.8 \times 399}{100}$$

=15.16 deaths to males aged <1

# Data Quality Lab Part 2

- Perform the age redistribution exercise for your test death data, assign each record an age group.
- Create pivot tables of the number of events by month

- Repeat this exercise with your country data
  - Assign each record an age group for mothers' age and for deaths
  - Create pivot tables for number of events by year
- Look at the results to check that the number of events is not too widely different month to month or year to year, and that no years are missing from your data set.
- If one year has a very low number of events you may be missing data and will need to check with the original source.
- Have a look to see if the data is consistent over time.
- Create a new worksheet for each pivot so that you can refer back to you workings.