Training course on demographic evaluation of age and sex data

Part 2(ii)

27 May – 3 June 2022
Contents of Part 2(ii) : Age investigations

- Collection of data on age
- Age investigations
  - Heaping
  - Omission
  - Age exaggeration
  - Measures of age heaping
  - Age ratios
Age & sex structure

• Gives the fundamental insight into the composition of the population
• Generally investigated together
  • Described separately, for clarity
• Investigations provide clear indication of errors relating to
  • Heaping
  • Omission
  • And (possibly) indication of age exaggeration
How are age data collected?

• Two approaches to collecting data on age
  • PREFERABLE: Ask the respondent their date of birth
    • More exact and less prone to age heaping error
    • Requires that most of the population are ‘date aware’
    • May still be errors associated with
      • A preference for ‘round’ years of birth (years ending in 0 or 5)
      • Particular events when there was mass registration of the population (e.g. the 1994 elections in South Africa)
  • LESS PREFERABLE: Ask the respondent their age
    • Usually less accurate as a result of
      • Misunderstanding: last, next, or nearest birthday
      • Rounding to nearest age ending in 0 or 5
      • Infants reported as age 1, not age 0
Age data

• Usually recorded as age last birthday
• Data on age is usually grouped into quinquennial (5-year) bands
  • To avoid the problem of excessive noise
  • To avoid the problem of age heaping
  • But ALWAYS investigate single-year age data first
Age investigations

• Before you start, be alert to
  • How the data on age was collected, or calculated
  • Unreported ages
    • Ignore and report as ‘unknown’?
    • Distribute across known ages?
    • Distribute by other characteristics?
  • Open-ended groups (e.g. ages 85+)
    • Want elements in the open-ended group to be large enough for analysis; but not so large that you lose valuable information
    • No explicit guidance on choice of starting point for the open-ended group; perhaps use what has been used before?
  • The use of ‘error codes’ that may be mistaken for ‘age’
Age investigations

• Don’t draw population pyramids for data evaluation
  • They are useful for public communication of results, but are (VERY) time-consuming to produce in Excel and there are faster and better ways of getting useful information

• Indices of age structure (e.g. Whipple, Myers) are of limited value
  • Again, one gains as much information from visual assessment
  • Covered here for completeness
Age investigations

• The essential investigation is to plot the numbers enumerated at each age and by sex
  • First, by individual age
  • Then, in conventional 5-year age groups

• A useful formula if you don’t have the data grouped into 5-year ages:
  • $AgeGroup = 5.\text{int}(age / 5)$ (e.g. if age = 13, $AgeGroup = 10$)
  • You can then use the Excel $\text{SUMIF}$ function to produce a tabulation by 5-year groups
What do we expect to see in looking at age data?

• Age heaping
• Omission
• Age exaggeration

• Clues – or pointers – to other demographic factors:
  • Recent fertility trends
  • The age pattern of mortality
  • The extent of migration
Age heaping

• Tends to be on ages 0 or 5
• Unless derived from year of birth, in which case heaping may be on ages corresponding to year of birth ending in 0 or 5
• May (rarely) be associated with more complex patterns of heaping
  • E.g. South Africa voter registration in 1994
Omission

- Omission tends to manifest, particularly, at the youngest ages (0 and 1)
  - Might create a misleading impression of recent fertility patterns
    - Omission of infants suggests lower fertility in the year or so before the census/survey
  - Sex-specific omissions may occur in some societies, where (e.g.) women are differentially under-enumerated
    - But be alert for other possible explanations (migration; war; famine; diseases such as HIV that typically affect particular age groups)
Age exaggeration

• Tends to manifest at the oldest ages, particularly in societies where age is venerated
  • Results in there being apparently many more very old people than could plausibly be with knowledge of patterns of mortality
Questions to ask in ANY graphical analysis

• What are we looking at?
  • What does the plot show (axes; quantities)?
  • Does the plot show what you think it does?
• What do we see?
  • What are the characteristics of what we have?
• How might we explain what we see?
• How might we test our explanations?
  • With what data?
Population of Bahrain -- 2010

Population of South Africa -- 2011

WHAT DO YOU SEE?

Whipple’s Index

- A summary measure of age heaping on ages ending in 0 and 5
  - Assumes (unrealistically) that the population within any 5 year age group is rectangular (i.e. equal numbers at each age in the same group)
  - Typically calculated only for ages 23 – 62 to avoid common problems at the youngest and oldest ages
    - The effect of mortality at oldest ages (violates the assumption above)
    - The effect of omission and heaping at younger ages
Whipple’s Index

• Calculated as

\[ WI = \frac{1}{5} \left( \frac{N_{25} + N_{30} + \cdots + N_{55} + N_{60}}{N_{23} + N_{24} + \cdots + N_{61} + N_{62}} \right) \times 100 \]

• Takes values in the range >100-500
  • 100 ▶ rectangular
  • 500 ▶ completely heaped
  • < 100 ▶ aversion to 0, 5

<table>
<thead>
<tr>
<th>Value of WI</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>WI &lt; 105</td>
<td>Highly accurate data</td>
</tr>
<tr>
<td>105 ≤ WI &lt; 110</td>
<td>Fairly accurate data</td>
</tr>
<tr>
<td>110 ≤ WI &lt; 125</td>
<td>Approximate data</td>
</tr>
<tr>
<td>125 ≤ WI &lt; 175</td>
<td>Rough data</td>
</tr>
<tr>
<td>WI ≥ 175</td>
<td>Very rough data</td>
</tr>
</tbody>
</table>
Myers’ Blended Index

• An alternative approach which tries to avoid the problems with Whipple’s Index
• Produces an index of preference for each terminal digit, indicating the deviation from 10% of the proportion of the total population reporting ages with a given terminal digit
• Theoretical range from 0 (no preference) to 90 (complete preference for a single terminal digit)
• Implemented in the Excel workbook that accompanies this session
Index values

- As can be seen from the summary of the index values below – there is next to no useful information contained in these values beyond what was already visible in the data

<table>
<thead>
<tr>
<th>Country</th>
<th>Whipple’s</th>
<th>Myers’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bahrain 2010</td>
<td>100.18</td>
<td>1.10</td>
</tr>
<tr>
<td>Nepal 2001</td>
<td>206.15</td>
<td>18.66</td>
</tr>
<tr>
<td>South Africa 2011</td>
<td>97.61</td>
<td>0.98</td>
</tr>
</tbody>
</table>
Age ratios

- In the absence of sharp changes in fertility or mortality, significant levels of migration, or other distorting factors, the enumerated size of a particular cohort should be approximately equal to the average size of the immediately preceding and following cohorts.
  - The age ratio for a particular cohort to the average of the counts for the adjacent cohorts should be approximately equal to 1 (or 100 if multiplied by a constant of 100).
Age ratios

• Significant departures from this “expected” ratio indicate either the presence of census error in the census enumeration or of other factors

• An age ratio is calculated as:

\[ nAR_x = \frac{2 \cdot nN_x}{nN_{x-5} + nN_{x+5}} \]

• A (wrongly) low value of \( nN_x \) will result in THAT age ratio \( nAR_x \) being low, AND in the two adjacent values of \( nAR_{x-5} \) and \( nAR_{x+5} \) being too high
Reminder!

• If you have any queries or questions on this material, please email Thomas.Moultrie@un.org, copied to escap-crvs@un.org

• I will do my best to answer questions, either by return of email, or in the first plenary session on 27 May 2022