# Analysis Age and Sex Distribution Data

United Nations Statistics Division

# Method of age and sex distribution data

- Basic graphical tools
  - Graphical analysis
    - Population pyramids
    - Graphical cohort analysis
  - Age and sex ratios
  - Summary indices of error in age-sex data
    - Whipple's index
    - Myers' Blended Method
- Uses of consecutive censuses
- Median age and dependency ratios

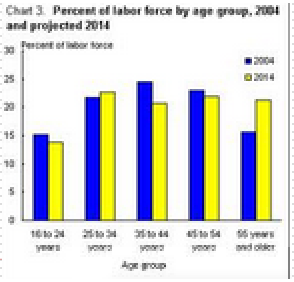
Focus of the presentation

# Importance of age-sex structures

- Insight on quality of census enumeration


- Determined by fertility, mortality and migration, and follows fairly recognizable patterns


- Having strong relation with social and economic characteristics of population



# What to look for at the evaluation

- Possible data errors in the age-sex structure, including
  - Age misreporting (age heaping and/or age exaggeration)
  - Coverage errors – net underenumeration(by age or sex)
- Significant discrepancies in age-sex structure due to extraordinary events
  - High migration, war, famine, HIV/AIDS epidemic etc.

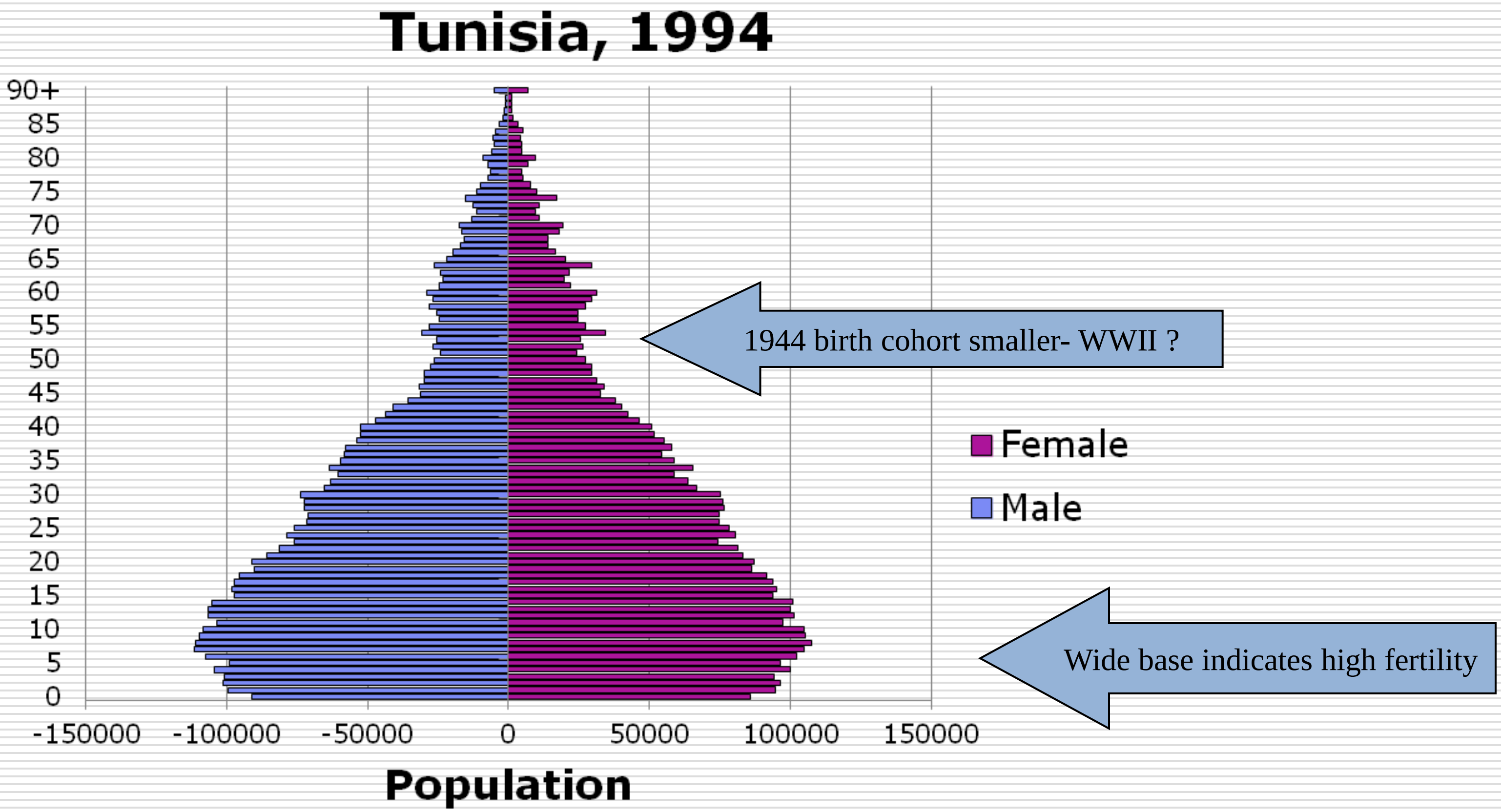
# Approaches to collecting age and its impact on quality

- Age - the interval of time between the date of birth and the date of the census, expressed in completed solar years
- Two approaches
  - The date of birth (year, month and day) - more precise information and is preferred
  - Completed age (age at the individual’s last birthday) – less accurate
    - Misunderstanding: the last, the next or the nearest birthday?
    - Rounding to nearest age ending in 0 or 5 (age heaping)
    - Children under 1 - may be reported as 1 year of age

# Basic graphical methods - Population Pyramid

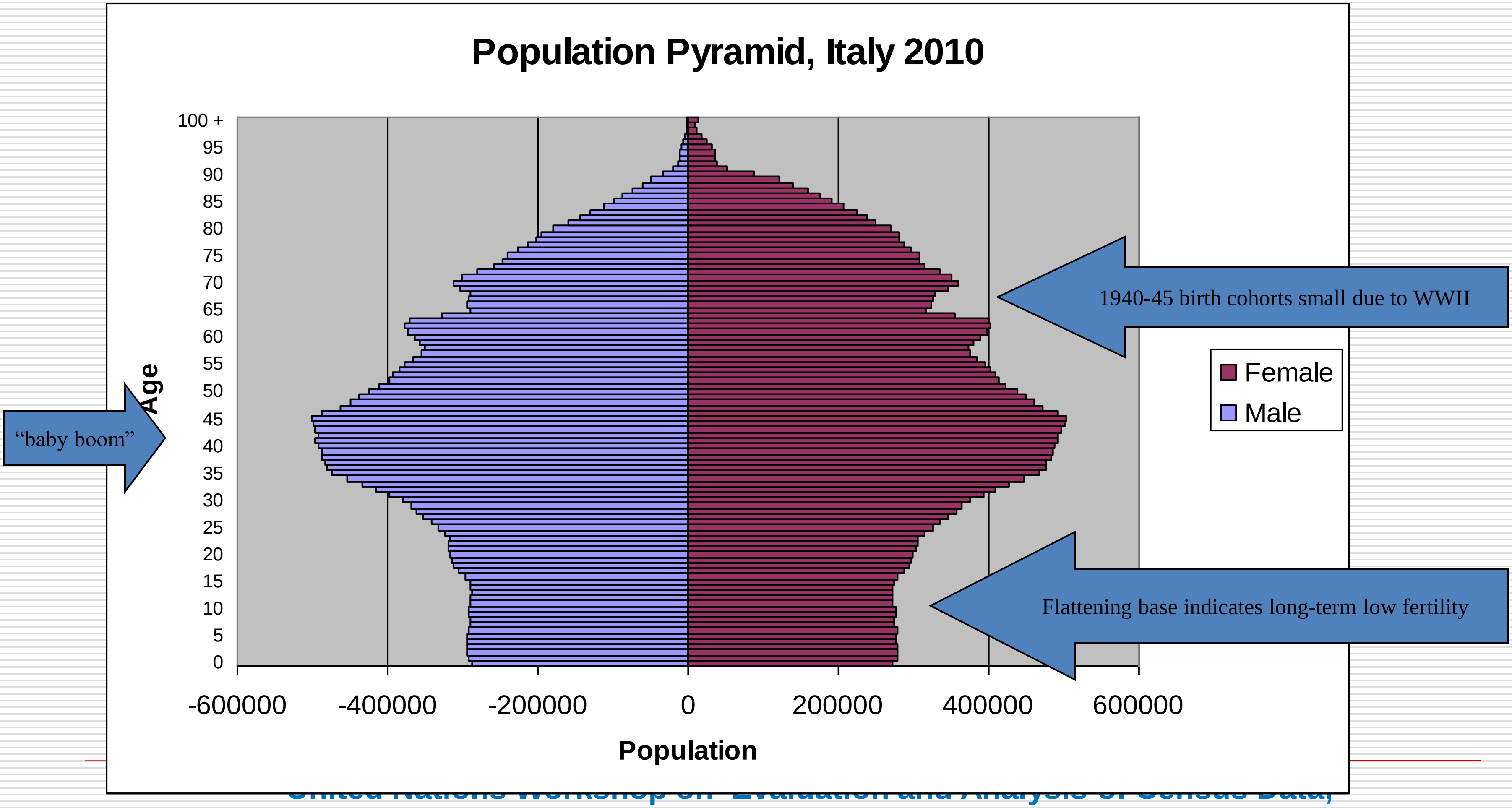
- Basic procedure for assessing the quality of census data on age and sex
- Displays the size of population enumerated in each age group (or cohort) by sex
- The base of the pyramid is mainly determined by the level of fertility in the population, while how fast it converges to peak is determined by previous levels of mortality and fertility
- The levels of migration by age and sex also affect the shape of the pyramid

# Population pyramid (1) – high population growth

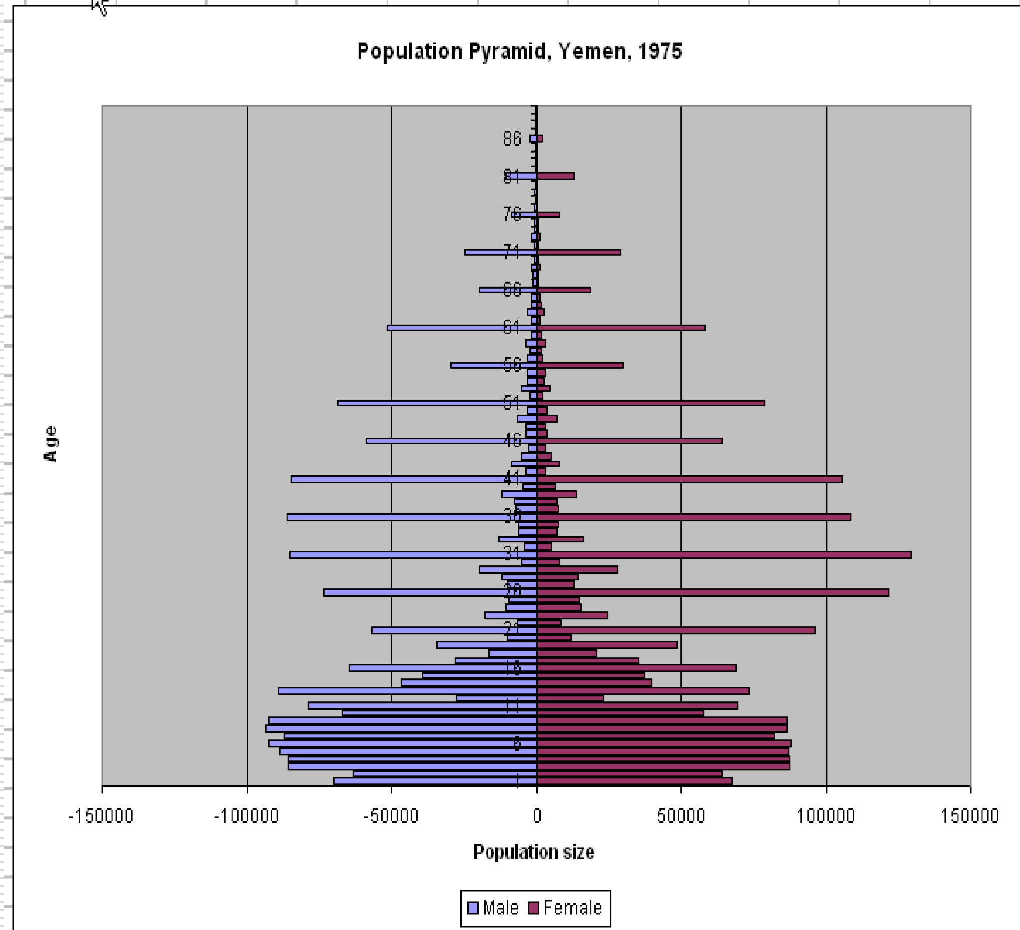


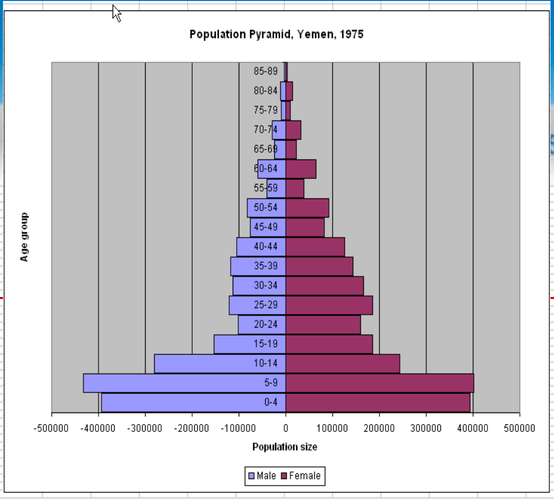

Source: Tabulated using data from *United Nations Demographic Yearbook*

# Population pyramid (2) – low population growth



# Population pyramid (3) - detecting errors

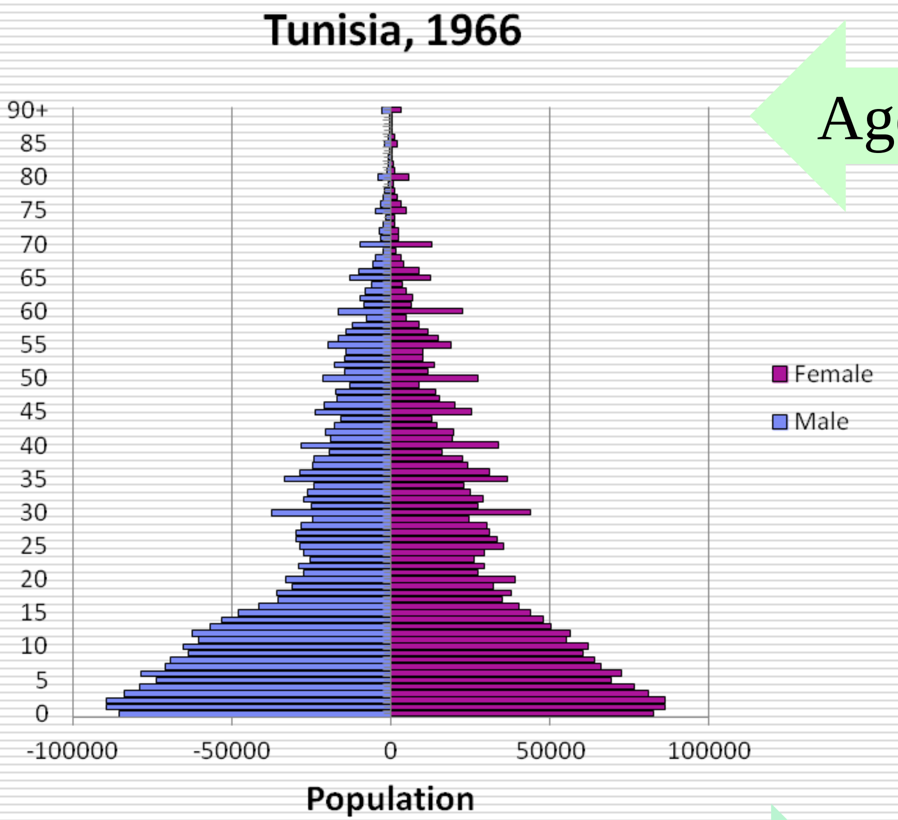


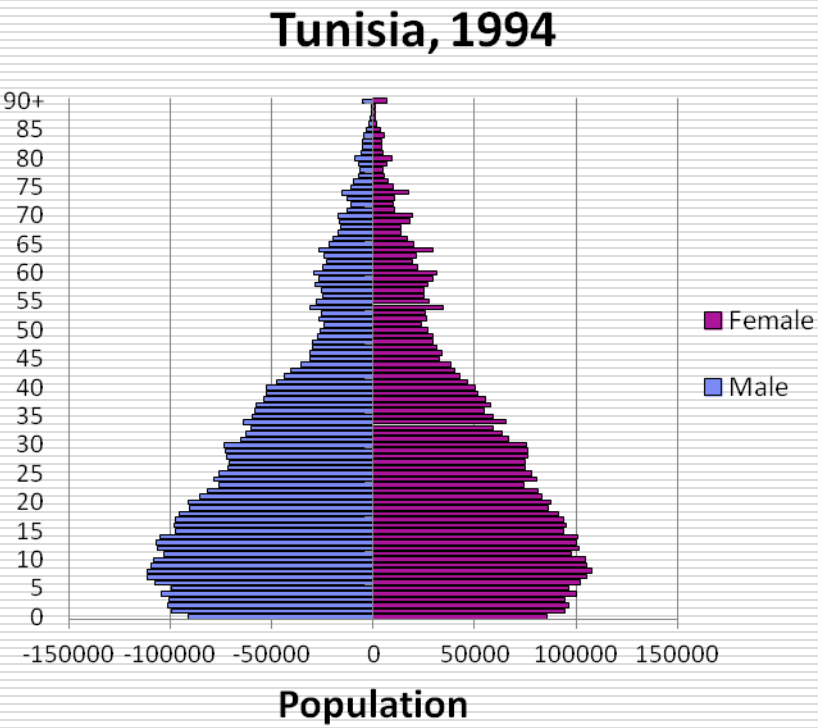

- Under enumeration of young children (< age 2)
- Age misreporting errors (heaping) among adults
- High fertility level
- Smaller population in 20-24 age group – extraordinary events in 1950-55?
- Smaller males relative to females in 20 – 44 - labor out-migration?

Source: Tabulated using data from U.S. Census Bureau, *Evaluating Censuses of Population and Housing*

# Population pyramid (4)- detecting errors



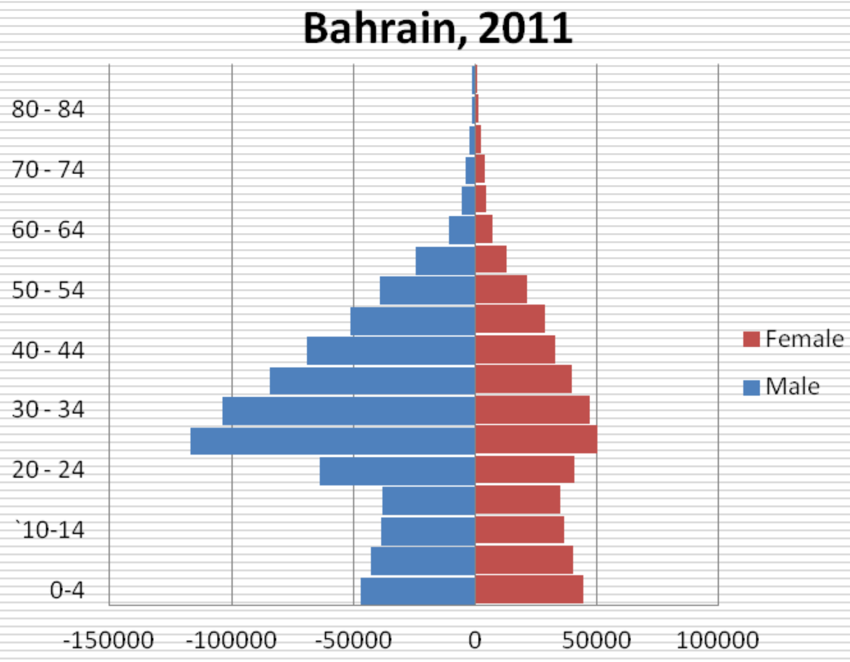

Age heaping

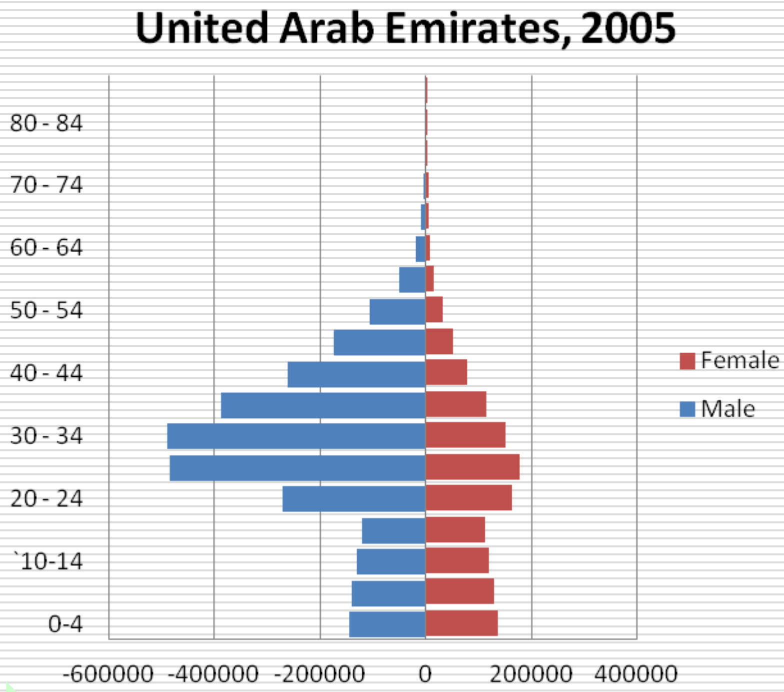



Declining fertility

United Nations Statistics Division

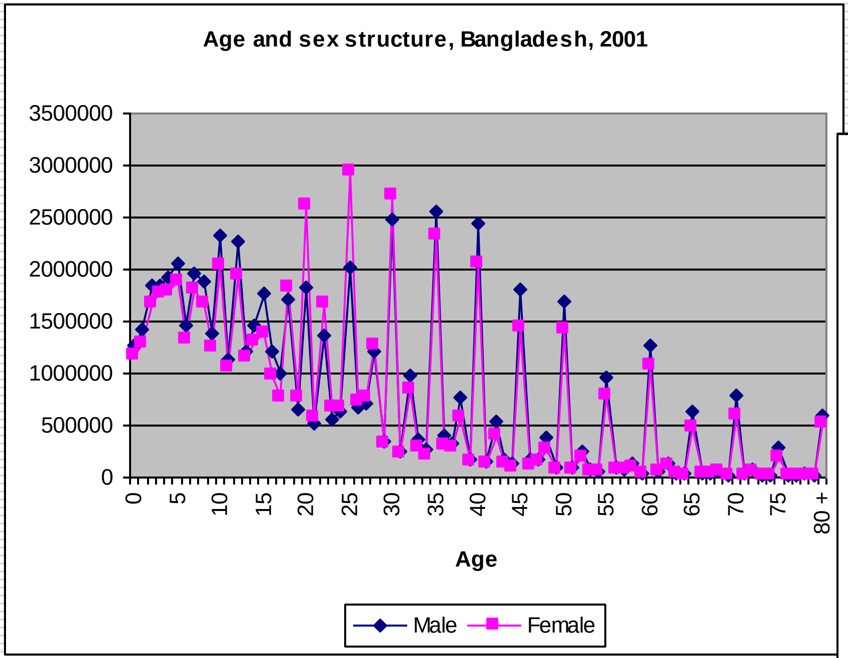
# Population pyramid (4)- detecting errors

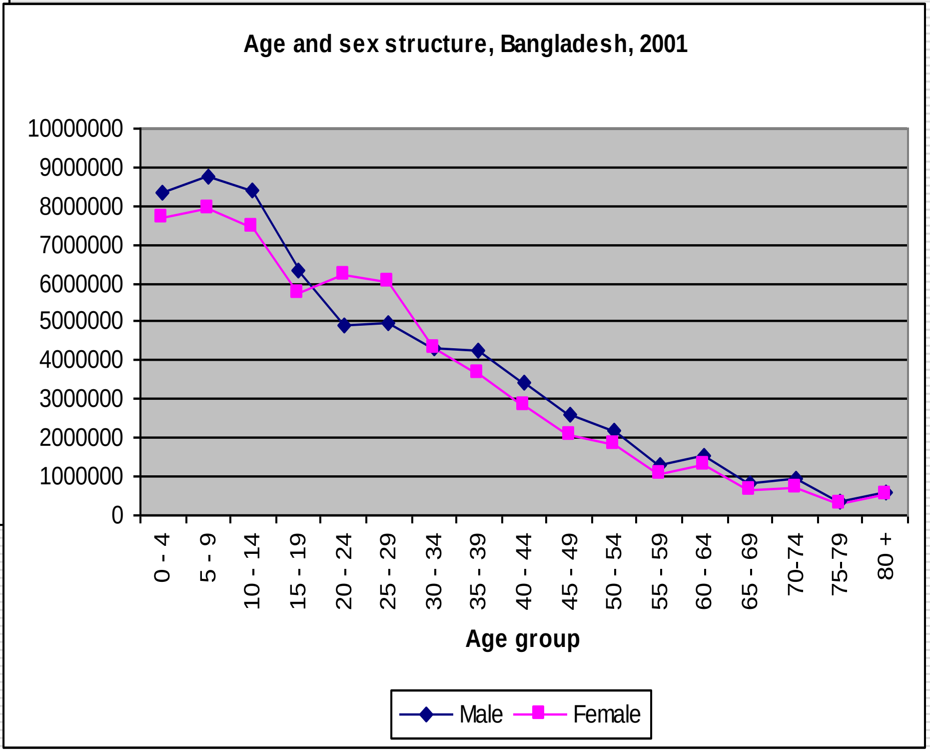



Labour in-migration

# Population pyramid (5) - line instead of bars




Data source: Tabulated using data from *United Nations Demographic Yearbook*

# Basic graphical methods - Graphical cohort analysis

- Tracking actual cohorts over multiple censuses
- The size of each cohort should decline over each census due to mortality, with no significant international migration
- The age structure (the lines) for censuses should follow the same pattern in the absence of census errors
- An important advantage - possible to evaluate the effects of extraordinary events and other distorting factors by following actual cohorts over time

# Graphical cohort analysis – Example (1)

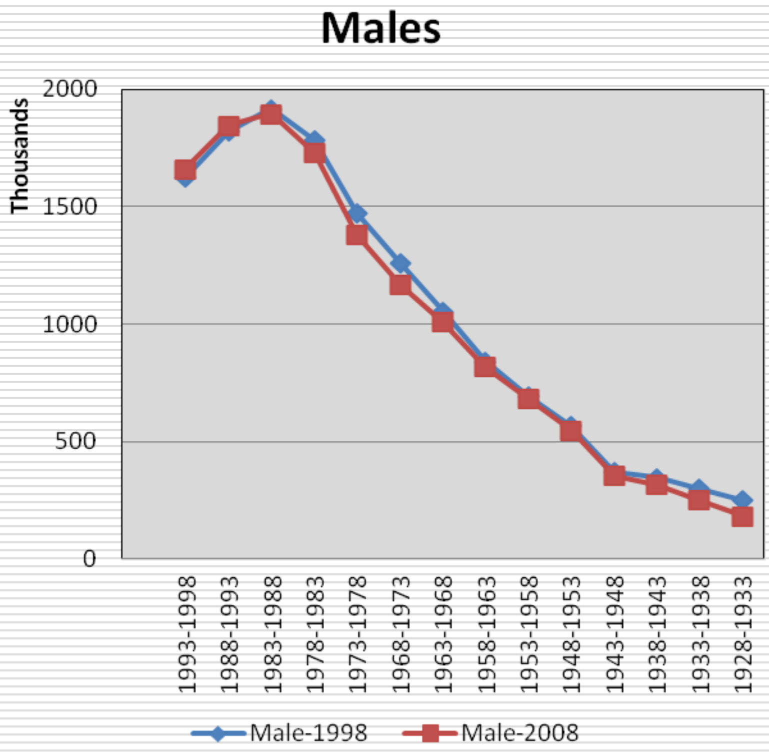
ALGERIA

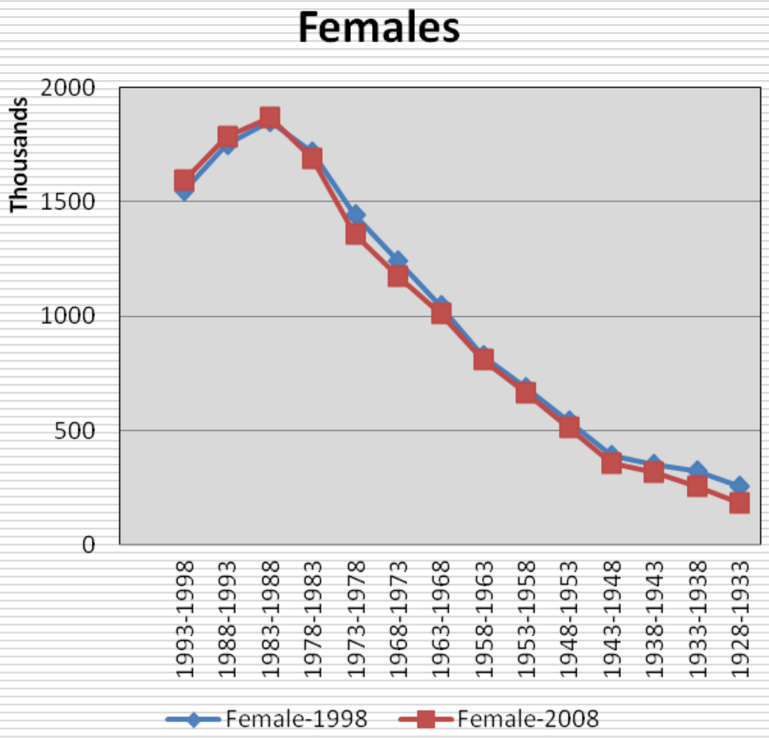
| 1998 | | 2008 | | |
|---|---|---|---|---|
| Age group | Male | Age Group | Male | Birth cohort |
| 0-4 | 1627670 | 0-4 | 1750097 | 2003-2008 |
| `5-9 | 1820858 | `5-9 | 1475674 | 1998-2003 |
| `10-14 | 1918833 | 10-14 | 1662260 | 1993-1998 |
| 15-19 | 1782614 | 15-19 | 1847311 | 1988-1993 |
| 20-24 | 1472255 | 20-24 | 1895704 | 1983-1988 |
| 25-29 | 1259989 | 25-29 | 1730409 | 1978-1983 |
| 30-34 | 1056110 | 30-34 | 1379085 | 1973-1978 |
| 35-39 | 841768 | 35-39 | 1167249 | 1968-1973 |
| 40-44 | 691275 | 40-44 | 1007683 | 1963-1968 |
| 45-49 | 565289 | 45-49 | 817004 | 1958-1963 |
| 50-54 | 371843 | 50-54 | 682357 | 1953-1958 |
| 55-59 | 345318 | 55-59 | 547181 | 1948-1953 |
| 60-64 | 301247 | 60-64 | 354694 | 1943-1948 |
| 65-69 | 252003 | 65-69 | 314958 | 1938-1943 |
| 70-74 | 163292 | 70-74 | 248672 | 1933-1938 |
| 75-79 | 107732 | 75-79 | 181478 | 1928-1933 |
| | | | | 1923-1928 |
| | | | | 1918-1923 |

- **Data is organized by birth cohort**
- **Exclude open-ended age category**
- **People who were born in the same years are compared in the analysis**

# Graphical cohort analysis – Example (1)

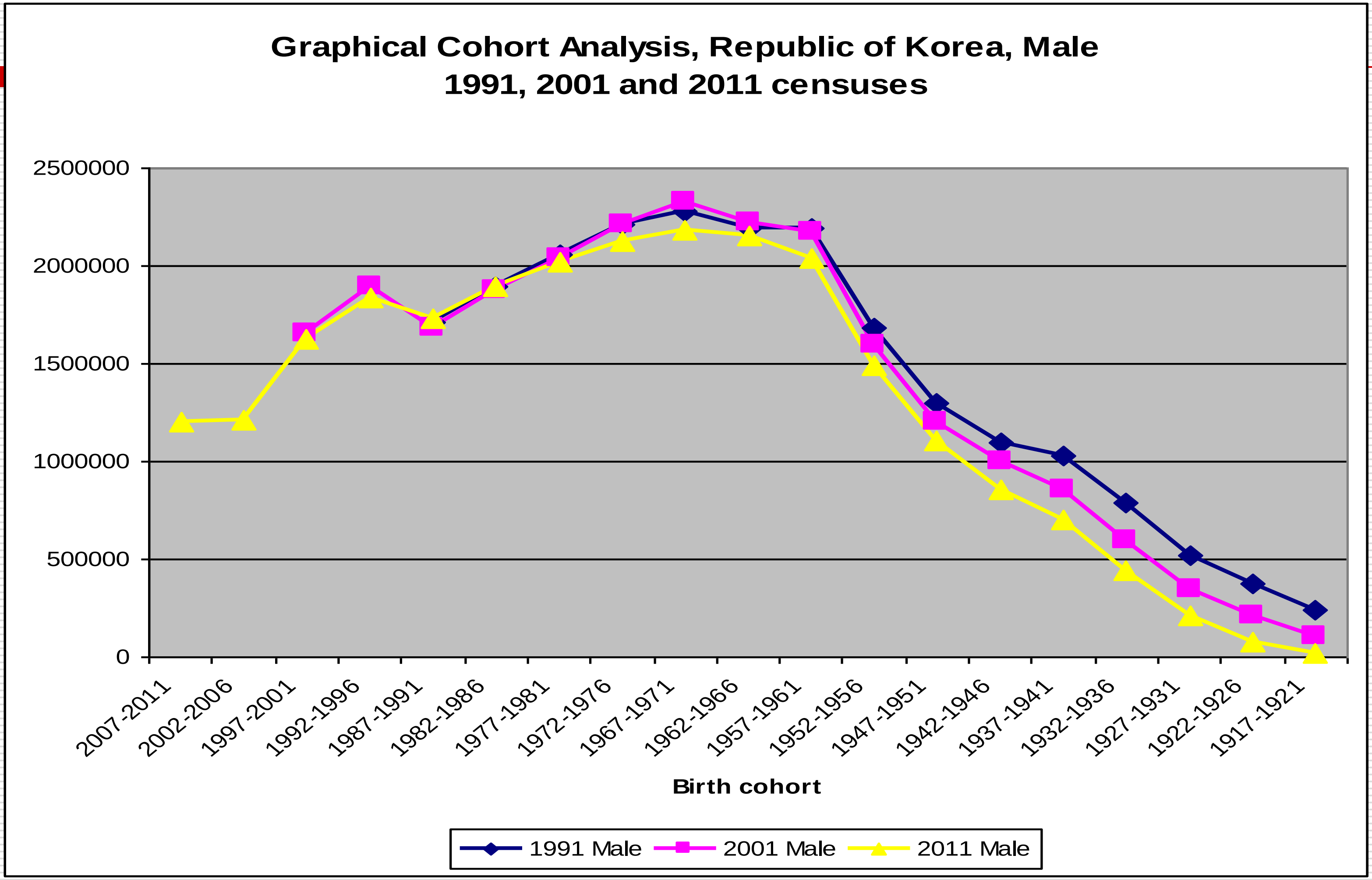
## Algeria, 1998 and 2008 Censuses



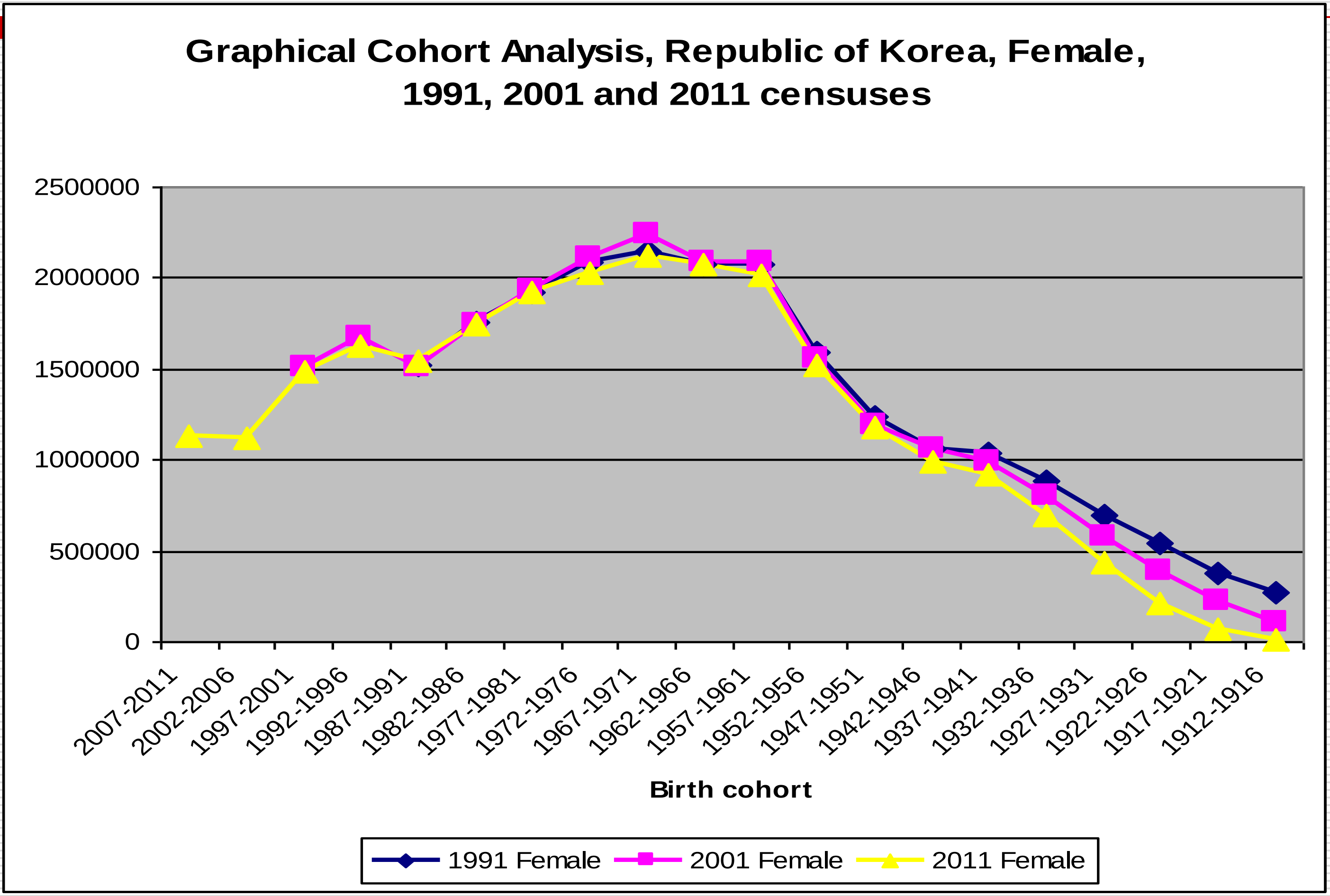



# Graphical cohort analysis – Example (2)





# Graphical cohort analysis – Example (2)



# Age ratios (1)

- 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)
- Significant departures from this "expected" ratio indicate either the presence of census error in the census enumeration or of other factors

# Age ratios (2)

- Age ratio for the age category x to x+4

$$_{5}AR_{x} = \frac{2 * {}_{5}P_{x}}{{}_{5}P_{x-n} + {}_{5}P_{x+n}}$$

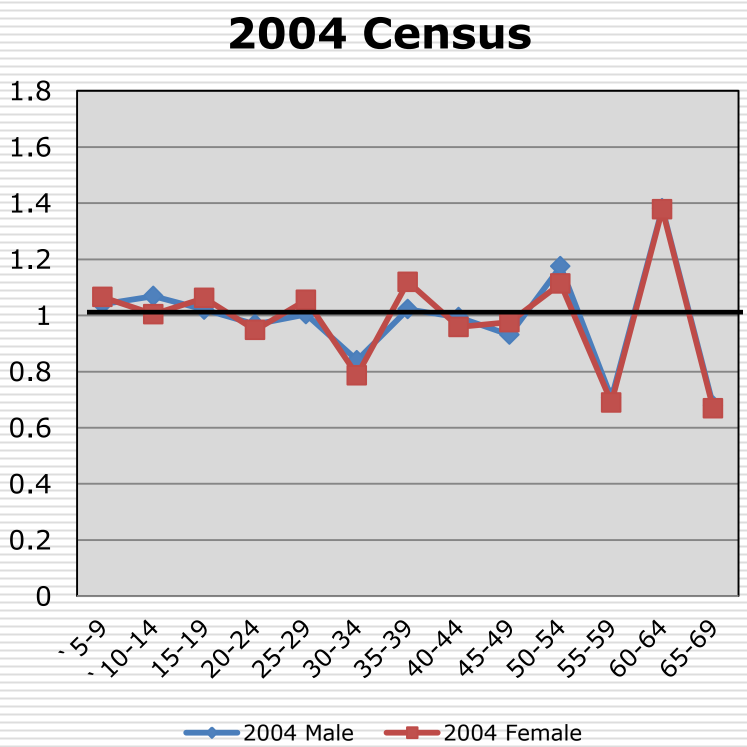
$_{5}AR_{x}$ = The age ratio for the age group x to x+4

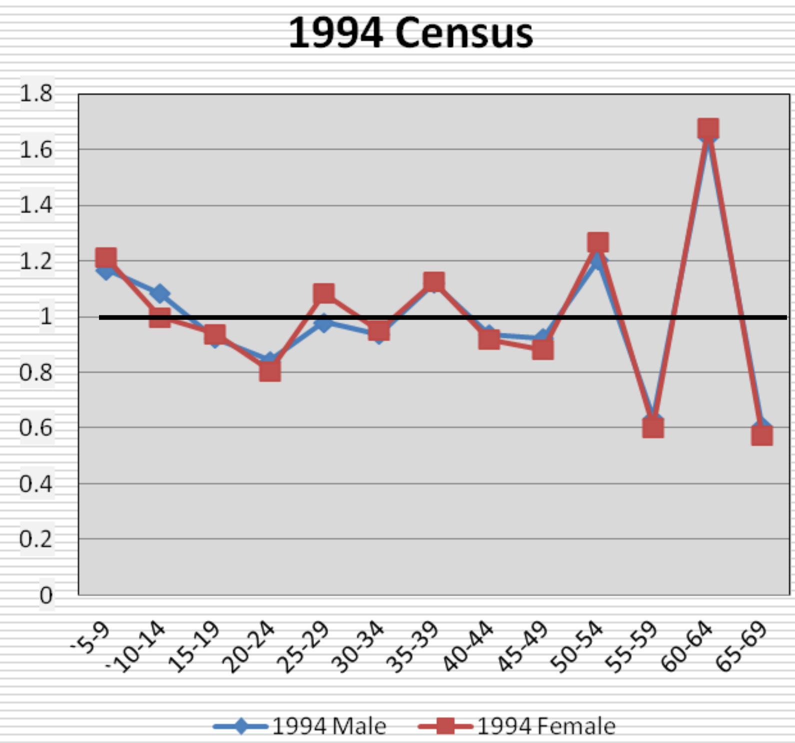
$_{5}P_{x}$ =The enumerated population in the age category x to x+4

$_{5}P_{x-5}$ = The enumerated population in the adjacent lower age category

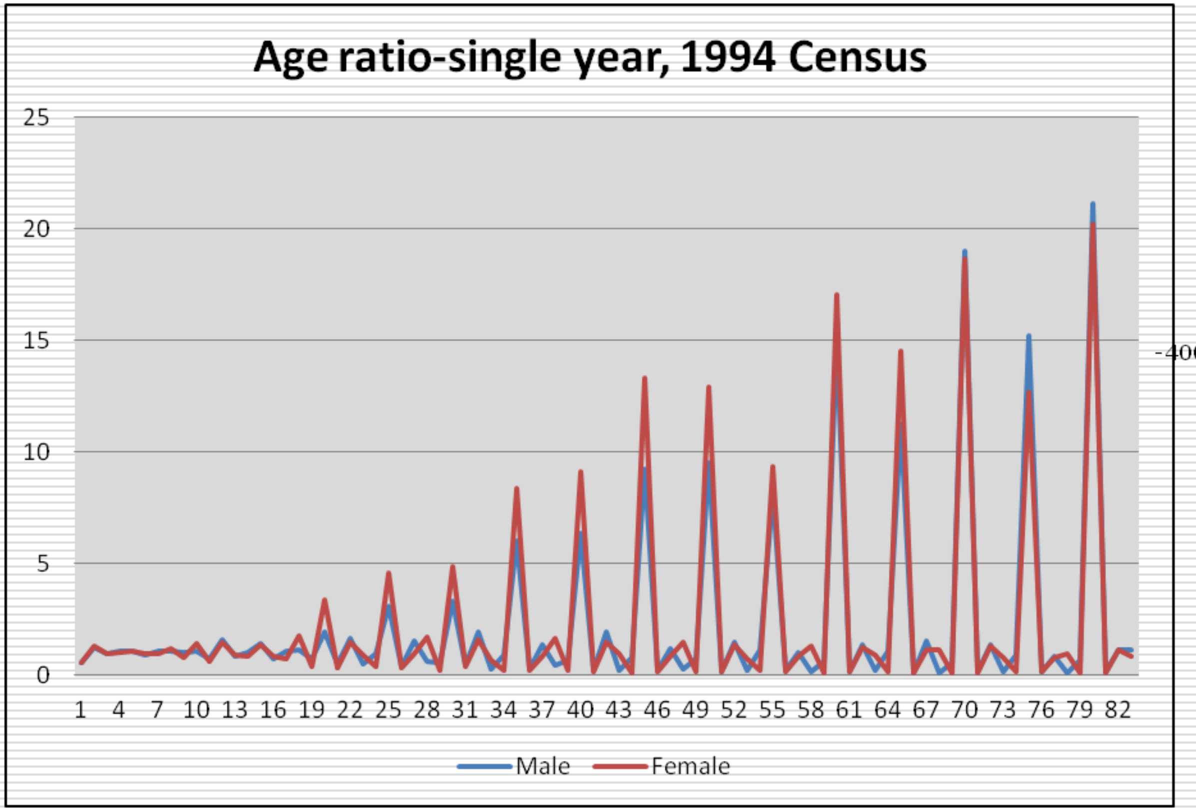
$_{5}P_{x+5}$ = The enumerated population in the adjacent higher age category

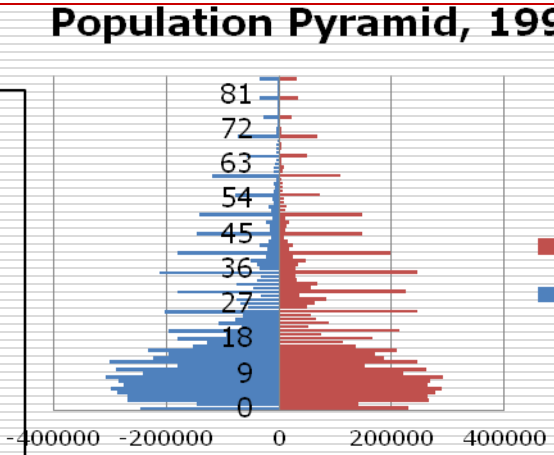
# Age ratios (3) – example –Yemen



# Age ratios (3) – example –Yemen




**Age misreporting increases with ages**

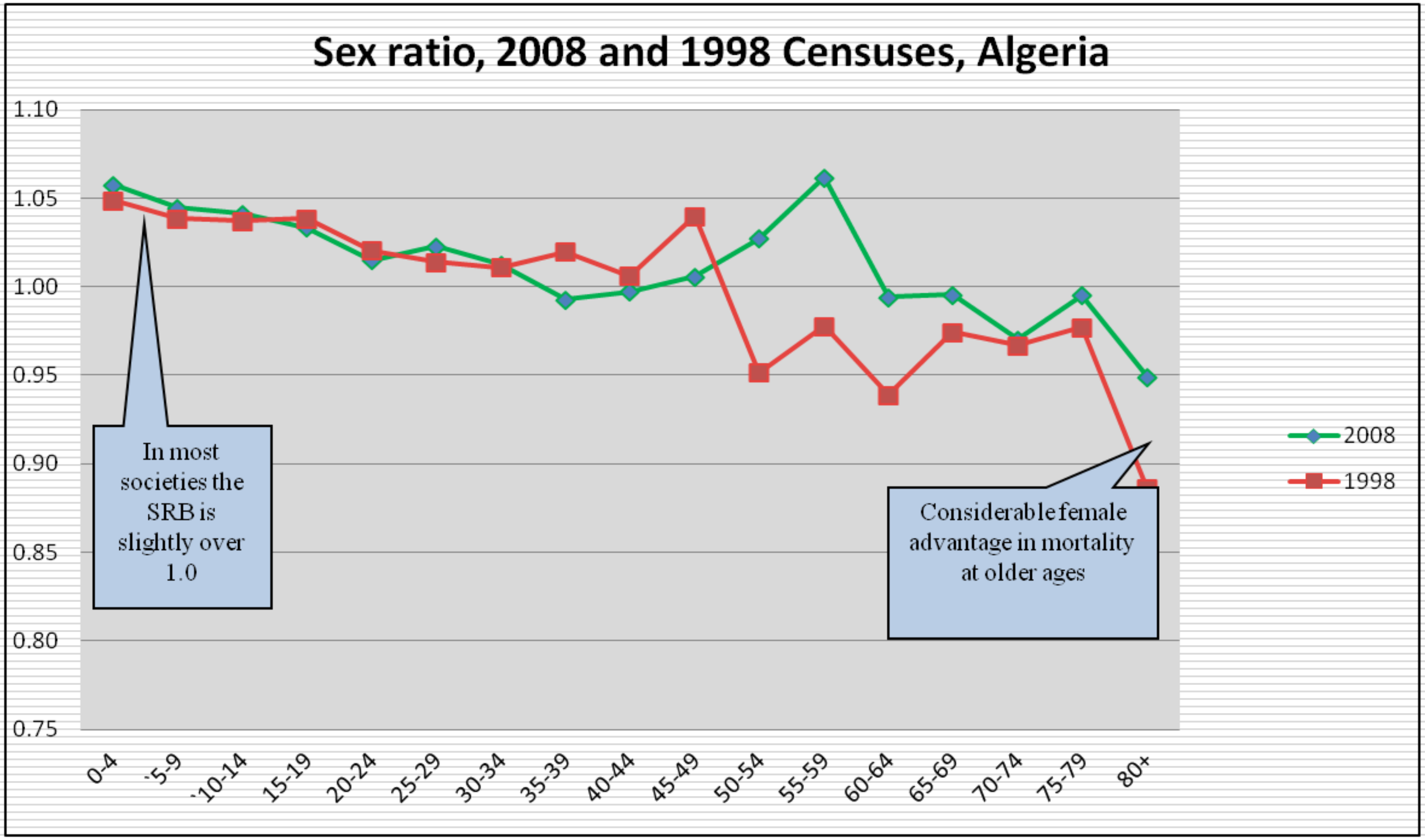
# Sex ratios (1) - calculation

Sex Ratio = $_5M_x \,/\, _5F_x$

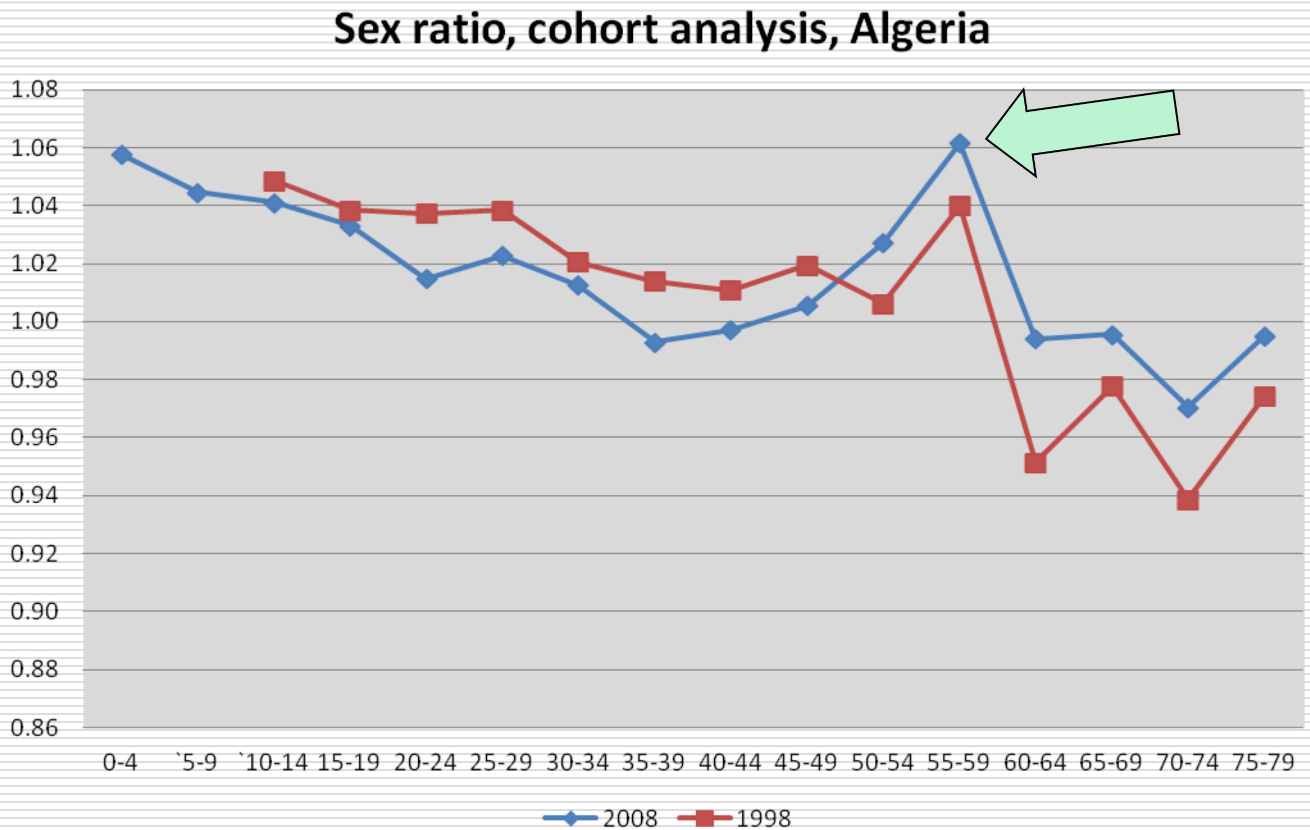
$_5M_x$ = Number of males enumerated in a specific age group

$_5F_x$ = Number of females enumerated in the same age group

# Sex ratios (2) - example

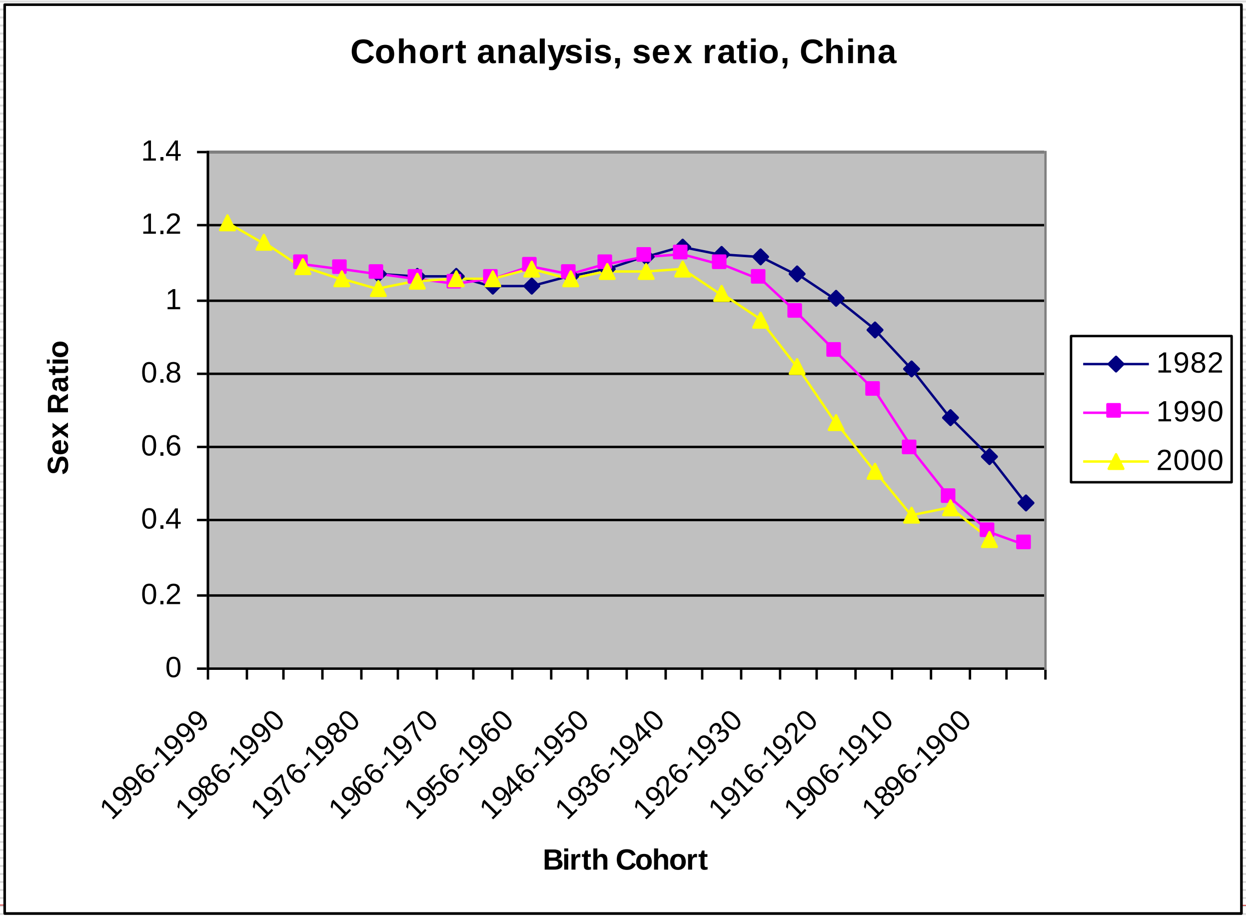


# Sex ratios (1) – cohort analysis




- Fluctuation due to age misreporting – different level for males and females?
- Two censuses indicate an excess of male population at age group 55-59

# Sex ratios (3) – cohort analysis



# Summary indices - Whipple`s Index

- Developed to reflect preference for or avoidance of a particular terminal digit or of each terminal digit
- Ranges between 100, representing no preference for “0” or “5” and 500, indicating that only digits “0” and “5” were reported in the census
- If heaping on terminal digits “0” and “5” is measured;

$$Index = \frac{\sum (P_{25} + P_{30} + ...... + P_{55} + P_{60})}{(1/5)\sum (P_{23} + P_{24} + ....... + P_{60} + P_{61} + P_{62})} \times 100$$



Source: Shryock and Siegel, 1976, *Methods and Materials of Demography*

# Whipple`s Index (2)

- If the heaping on terminal digit "0" is measured;

$$Index = \frac{P_{30} + P_{40} + P_{50} + P_{60}}{(1/10)\sum(P_{23} + P_{24} + ....... + P_{60} + P_{61} + P_{62})} \times 100$$

- The choice of the range 23 to 62 is standard, but largely arbitrary. In computing indexes of heaping, ages during childhood and old age are often excluded because they are more strongly affected by other types of errors of reporting than by preference for specific terminal digits

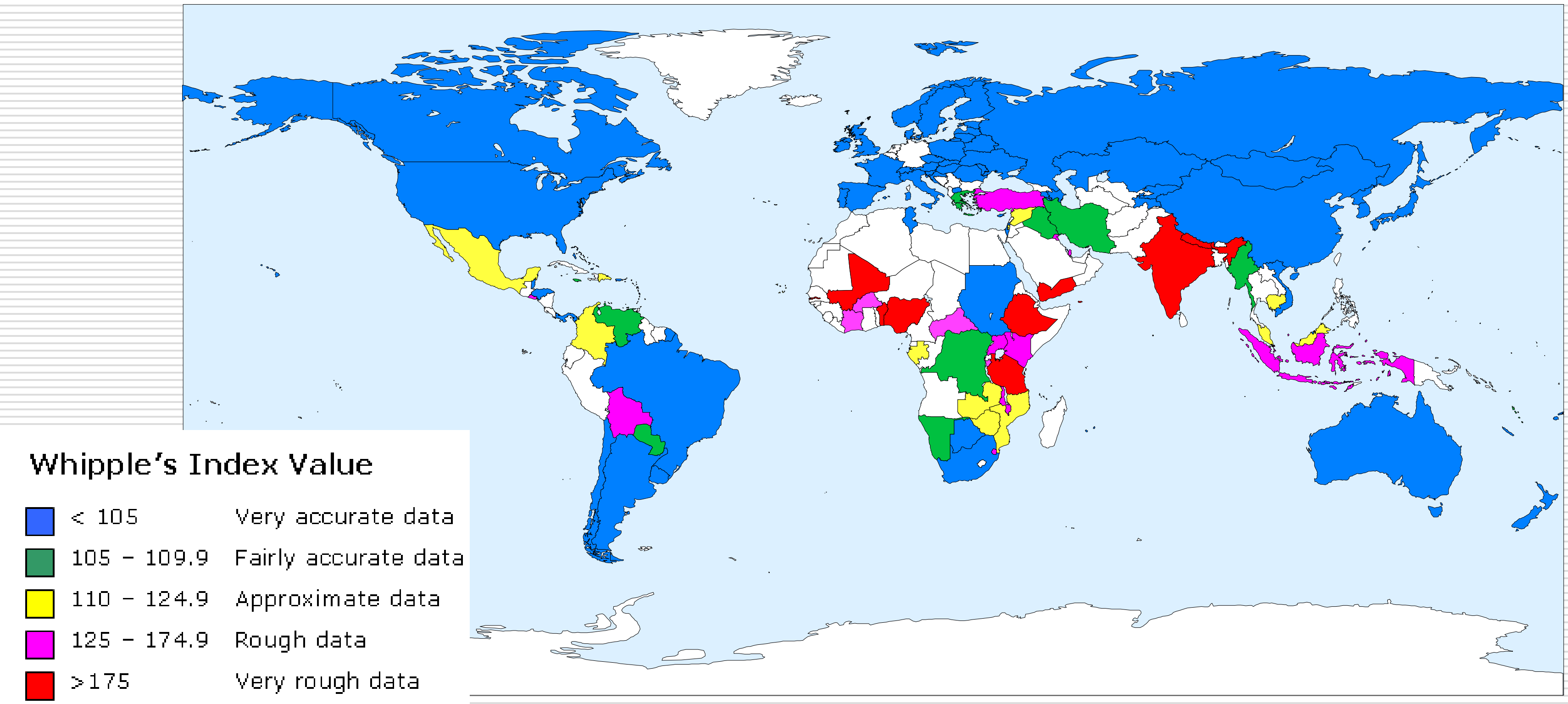
# Whipple`s Index (3)

- The index can be summarized through the following categories:

| | *Value of Whipple's Index* |
|---|---|
| • Highly accurate data | <= 105 |
| • Fairly accurate data | 105 – 109.9 |
| • Approximate data | 110 – 124.9 |
| • Rough data | 125 – 174.9 |
| • Very rough data | >= 175 |

# Whipple's index around the world

- Many of the countries that continue to have high Whipple's Index values are in Sub-Saharan Africa

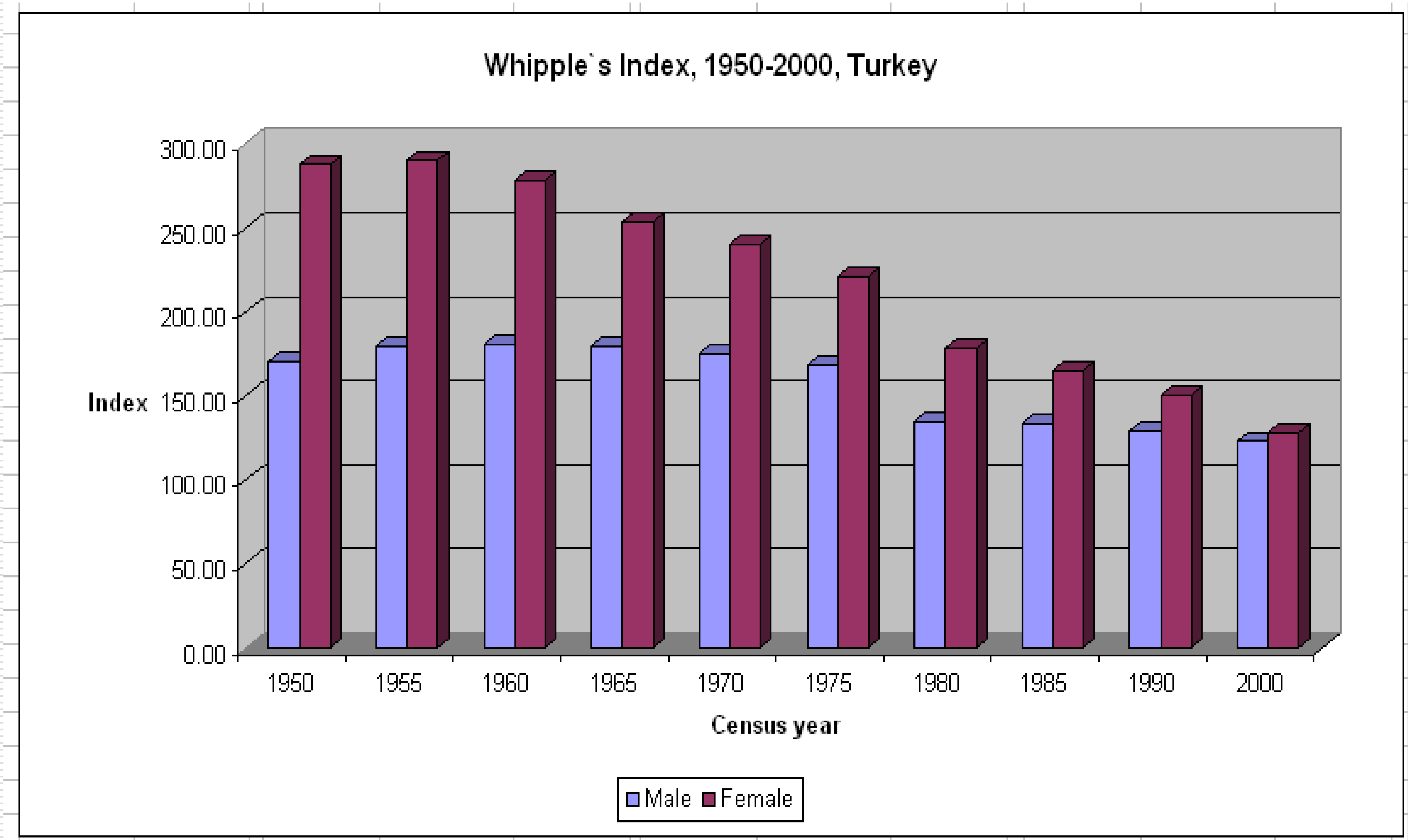





Data source: *Demographic Yearbook* special issue on age heaping: http://unstats.un.org/unsd/demographic/products/dyb/dybcens.htm

United Nations Statistics Division

# Improvement in the accuracy of age reporting over time



# Summary indices – Myers` Blended Index

- It is conceptually similar to Whipple`s index, except that the index considers preference (or avoidance) of age ending in each of the digits 0 to 9 in deriving overall age accuracy score
- The theoretical range of Myers` Index is from 0 to 90, where 0 indicates no age heaping and 90 indicates the extreme case where all recorded ages end in the same digit

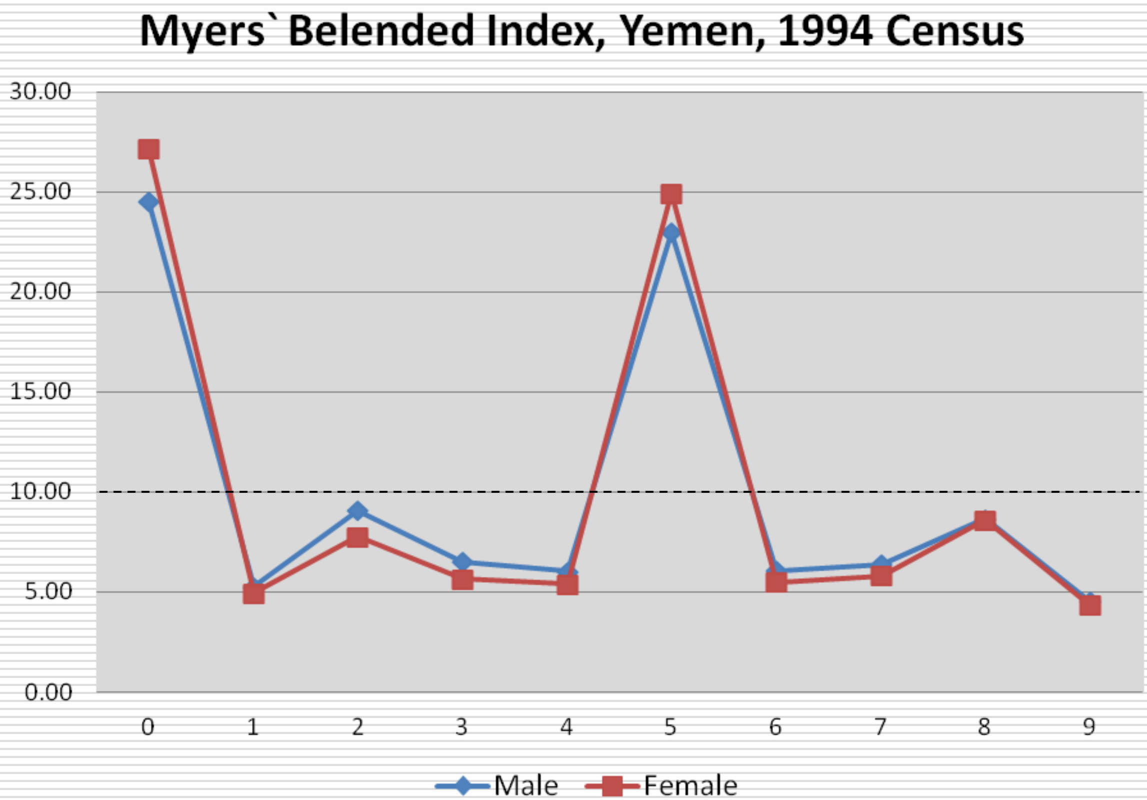
# Summary indices – Myers` Blended Index

| Age | Male |
|---|---|
| 10 | 288878 |
| 11 | 179463 |
| 12 | 299631 |
| 13 | 223022 |
| 14 | 195237 |
| 15 | 231906 |
| 16 | 152000 |
| 17 | 127790 |
| 18 | 180533 |
| 19 | 92898 |
| 20 | 196942 |
| 21 | 66713 |
| 22 | 107949 |
| 23 | 78482 |
| 24 | 64071 |
| 25 | 203117 |
| 26 | 54928 |
| 27 | 67472 |
| 28 | 75971 |
| 29 | 32162 |
| 30 | 179127 |
| 31 | 46012 |
| 32 | 75727 |
| 33 | 37839 |
| 34 | 31961 |
| 35 | 212862 |
| 36 | 33729 |
| 37 | 38564 |

**Myers` Blended Method, Yemen**

| Terminal digit | Sum of population ages eding in terminal digit From 10+x | From 20+x | Weights (given) Col-1 | Col-2 | Blended population (1)*(3)+(2)*(4) | Percent | Deviation from 10% abs(6)-10 |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| 0 | 1,179,546 | 890,668 | 1 | 9 | 9,195,558 | 24.52 | 14.52 |
| 1 | 340,379 | 160,916 | 2 | 8 | 1,968,086 | 5.25 | 4.75 |
| 2 | 549,827 | 250,196 | 3 | 7 | 3,400,853 | 9.07 | 0.93 |
| 3 | 377,435 | 154,413 | 4 | 6 | 2,436,218 | 6.50 | 3.50 |
| 4 | 324,324 | 129,087 | 5 | 5 | 2,267,055 | 6.04 | 3.96 |
| 5 | 955,035 | 723,129 | 6 | 4 | 8,622,726 | 22.99 | 12.99 |
| 6 | 273,101 | 121,101 | 7 | 3 | 2,275,010 | 6.07 | 3.93 |
| 7 | 264,819 | 137,029 | 8 | 2 | 2,392,610 | 6.38 | 3.62 |
| 8 | 342,271 | 161,738 | 9 | 1 | 3,242,177 | 8.65 | 1.35 |
| 9 | 170,273 | 77,375 | 10 | 0 | 1,702,730 | 4.54 | 5.46 |
| **Total** | | | | | **37,503,023** | **100** | **55.02** |
| | | | | | **Index of age prefence=** | | 27.51 |

# Summary indices – Myers` Blended Index




Age misreporting

- Ages ending with 0 and 5 : over-counting
- Ages ending with other digits (particularly with 1, 3, 6, 7 and 9) – under counting

# Conclusion: Uses and limitations

- Assessment of the age and sex structure of the population enumerated in a census is typically the first step taken in evaluating a census by means of demographic methods
- Demographic methods provide:
  - A quick and inexpensive indication of the general quality of data
  - Evidence on the specific segments of the population in which the presence of error is likely
  - "Historical" information which may be useful for interpreting the results of evaluation studies based on other methods, and in determining how the census data should be adjusted for use in demographic analyses

# Conclusion: Uses and limitations

- The major limitation of age and sex structure analysis is that it is not possible to derive separate numerical estimates of the magnitude of coverage and content error on the basis of such analyses alone
- It is often possible to assess particular types of errors which are likely to have affected the census counts for particular segments of the population. Estimates of coverage error from other sources often are required to verify these observations.

# References


- Shryock and Siegel, 1976, *Methods and Materials of Demography*
- IUSSP Tools for Demographic Estimation

  http://demographicestimation.iussp.org/
- PAS-Population Analysis Spreadsheets http://www.census.gov/population/international/software/uscbtoolsdownload.html