Evaluation of Age and Sex Distribution Data

United Nations Statistics Division

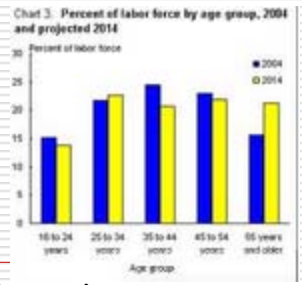
Evaluation method of age and sex distribution data

- Basic 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
- The use of stable population theory
- Uses of consecutive censuses

Focus of the presentation

Importance of age-sex structures

- Planning purposes – health services, sales programs, school, voting, labour supply
- Social science, economist, gender studies
- Studying population dynamics – fertility, mortality, migration
- Insight on quality of census enumeration
- Having strong effect on other characteristics of a population
  - Determined by fertility, mortality and migration, and follows fairly recognizable patterns



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.

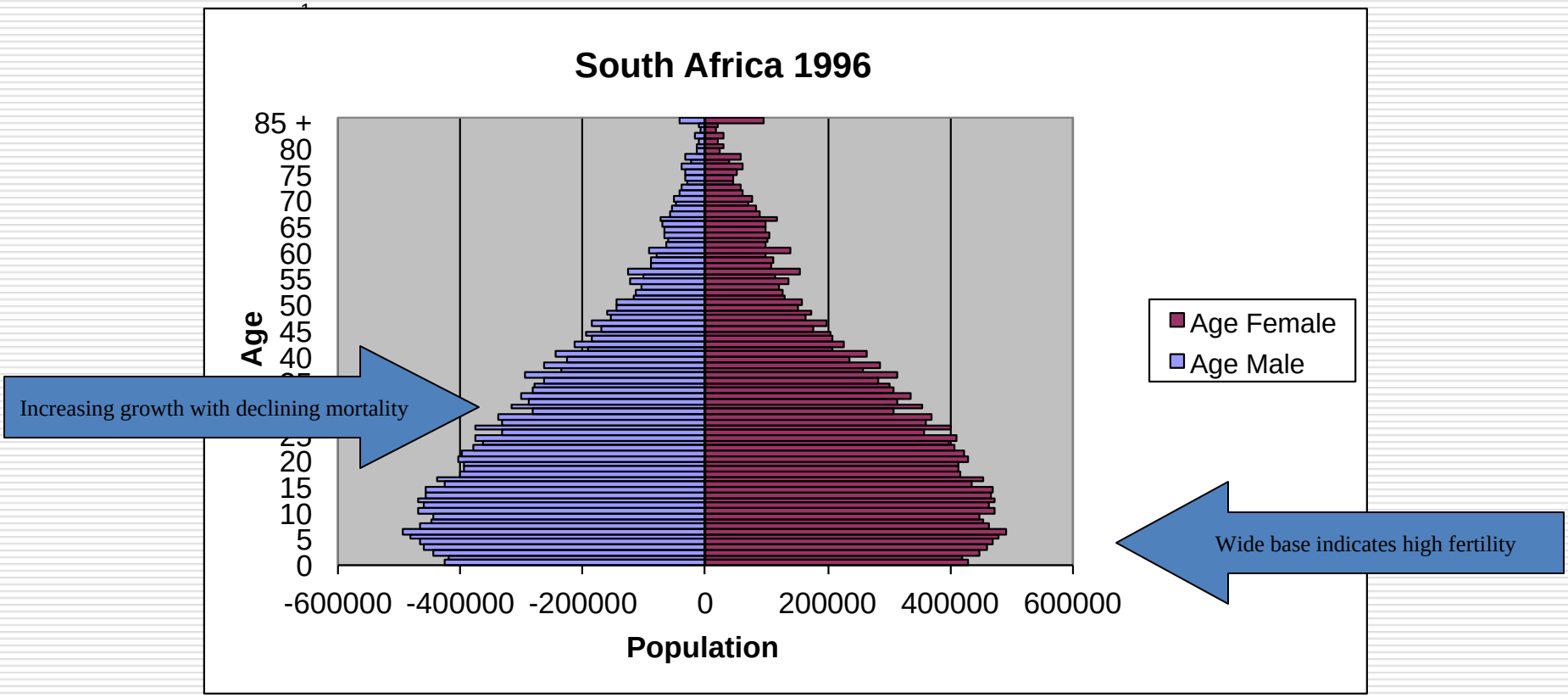
Approaches to collecting age and its impact on quality

- 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
    - Use of different calendars in the same country– western, Islamic or Lunar

Basic graphical methods - Population Pyramid

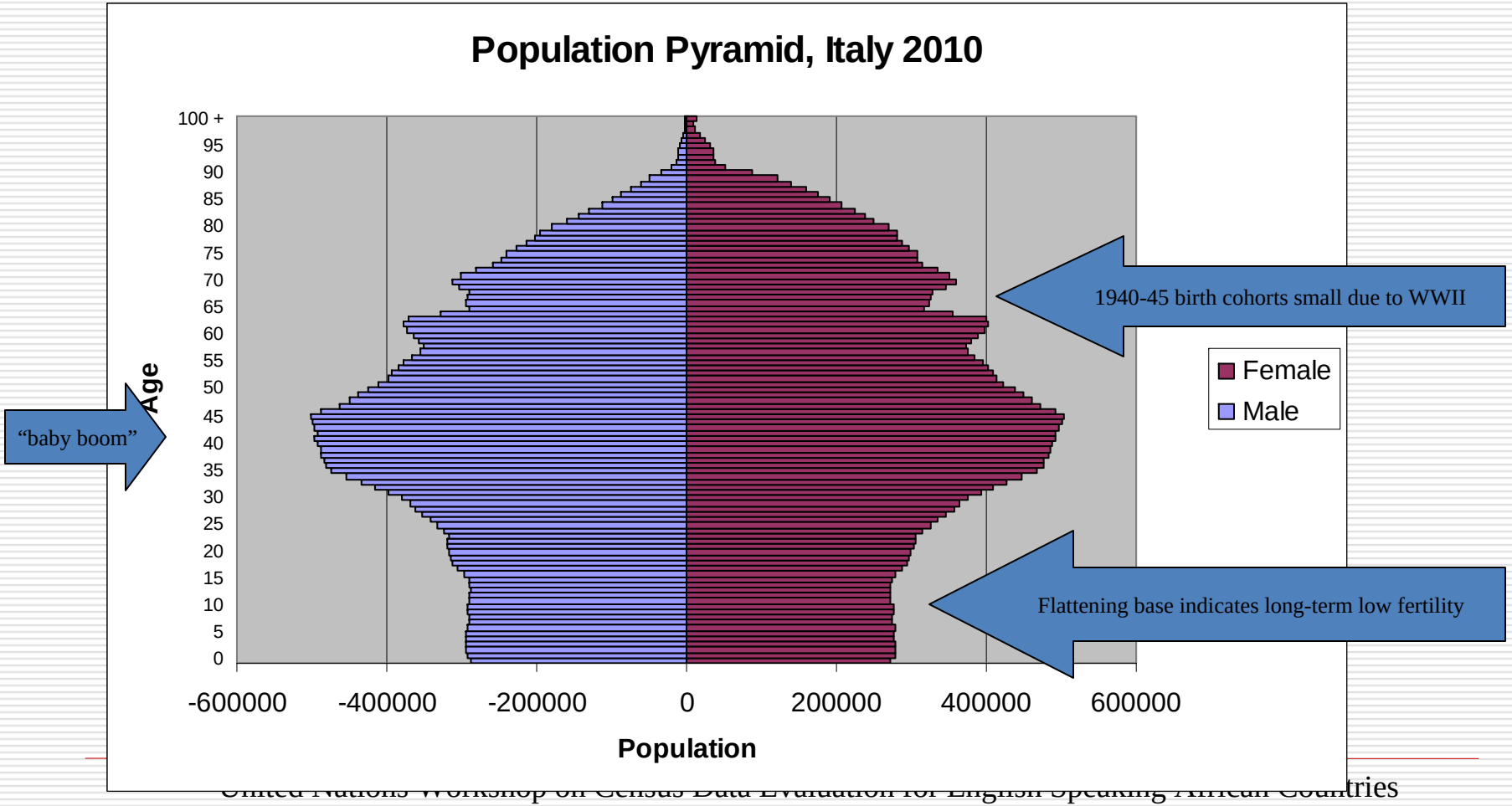
- Basic procedure for assessing the quality of census data on age and sex
- Displays the size of population enumerated in each age group (or cohort) by sex
- The base of the pyramid is mainly determined by the level of fertility in the population, while how fast it converges to peak is determined by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

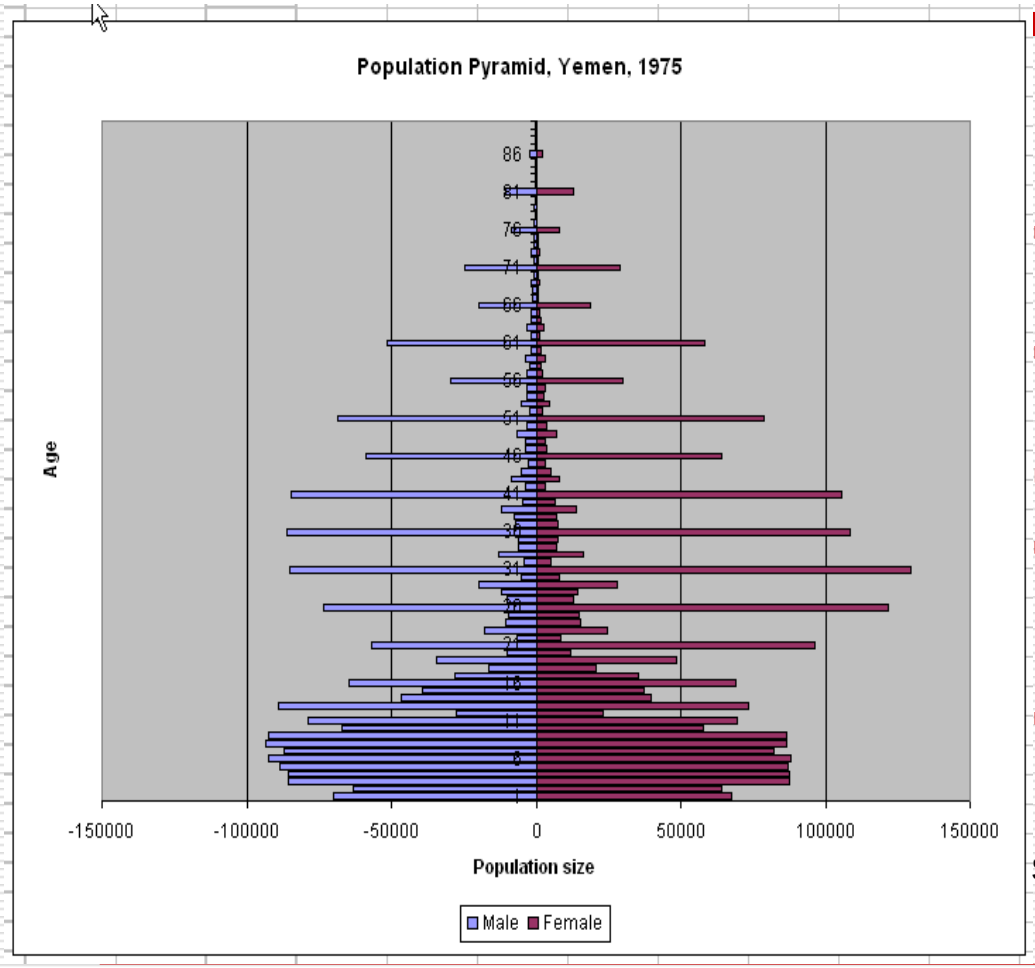


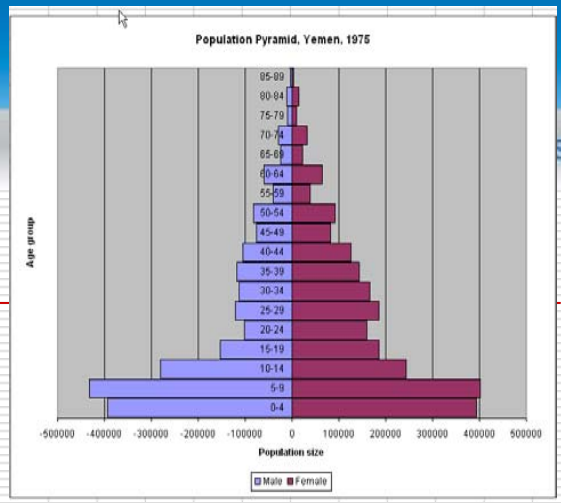
Source: Tabulated using data from *United Nations Demographic Yearbook*

Population pyramid (2) – low population growth



Population pyramid (3) - detecting errors

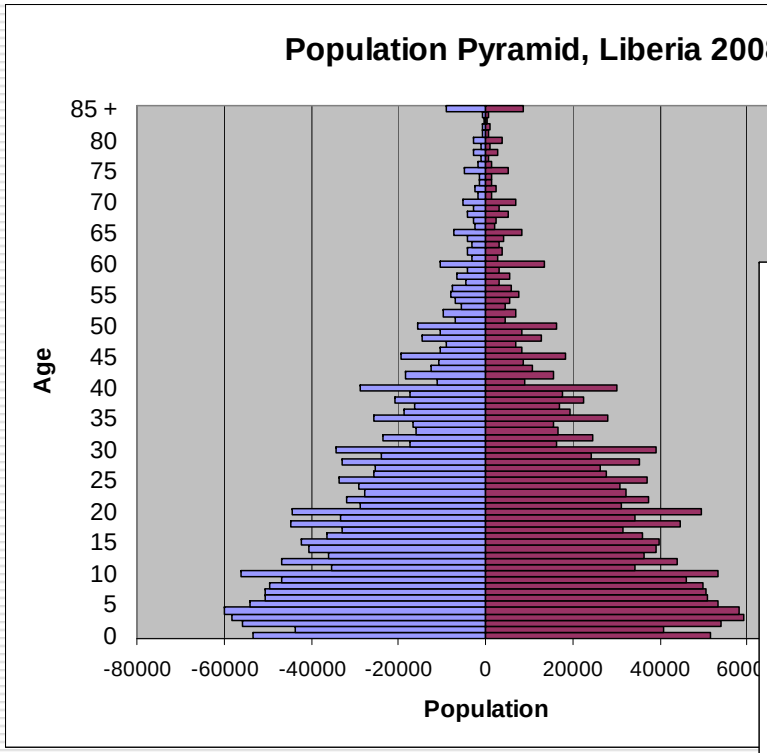




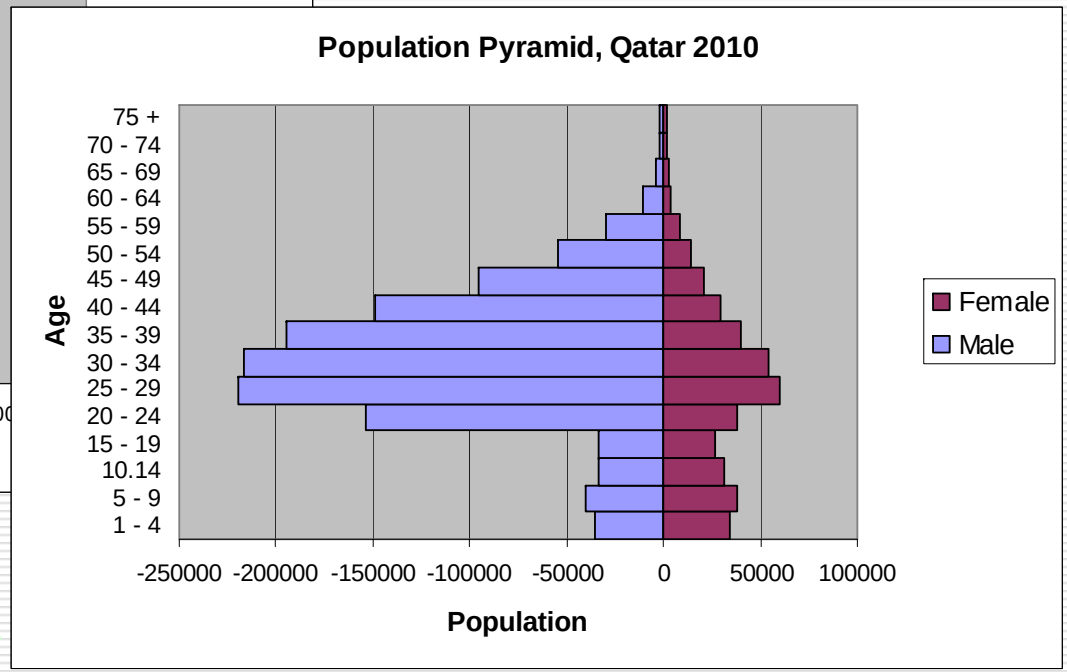
- Under enumeration of young children (< age 2)
- Age misreporting errors (heaping) among adults
- High fertility level
- Smaller population in 20-24 age group – extraordinary events in 1950-55?
- Smaller males relative to females in 20 – 44 - labor out-migration?

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

Population pyramid (4) - detecting errors

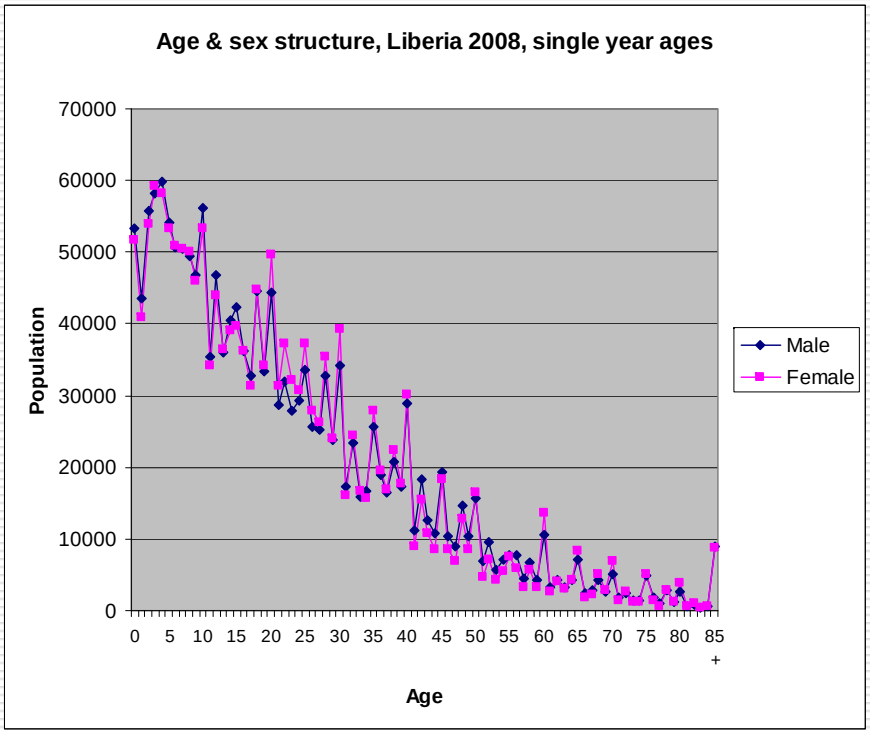


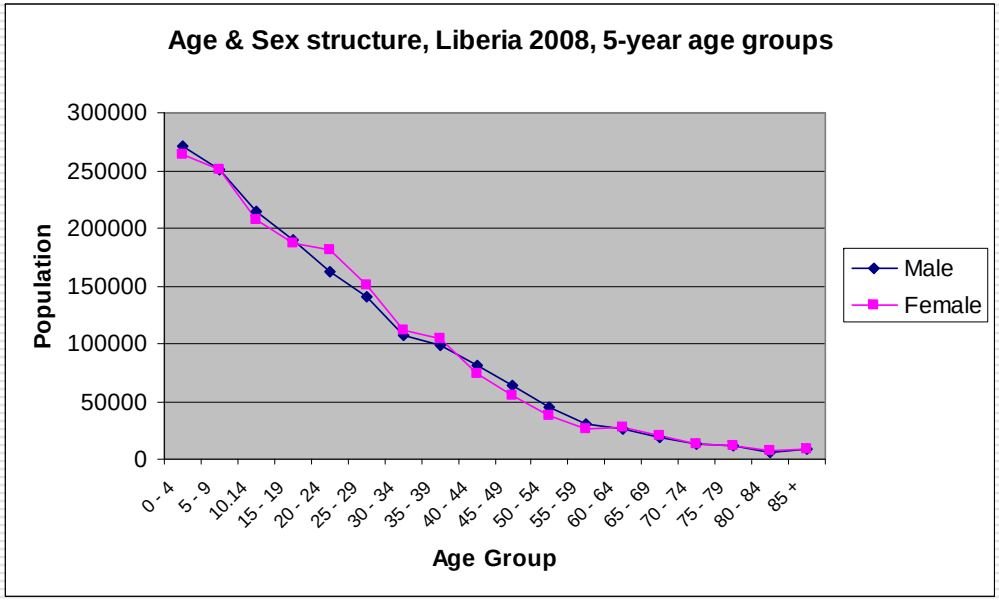
Age heaping



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)

- For this analysis we organize the data by birth cohort
- New cohorts will be added and older cohorts will be lost as we progress to later censuses
- Exclude open-ended age category

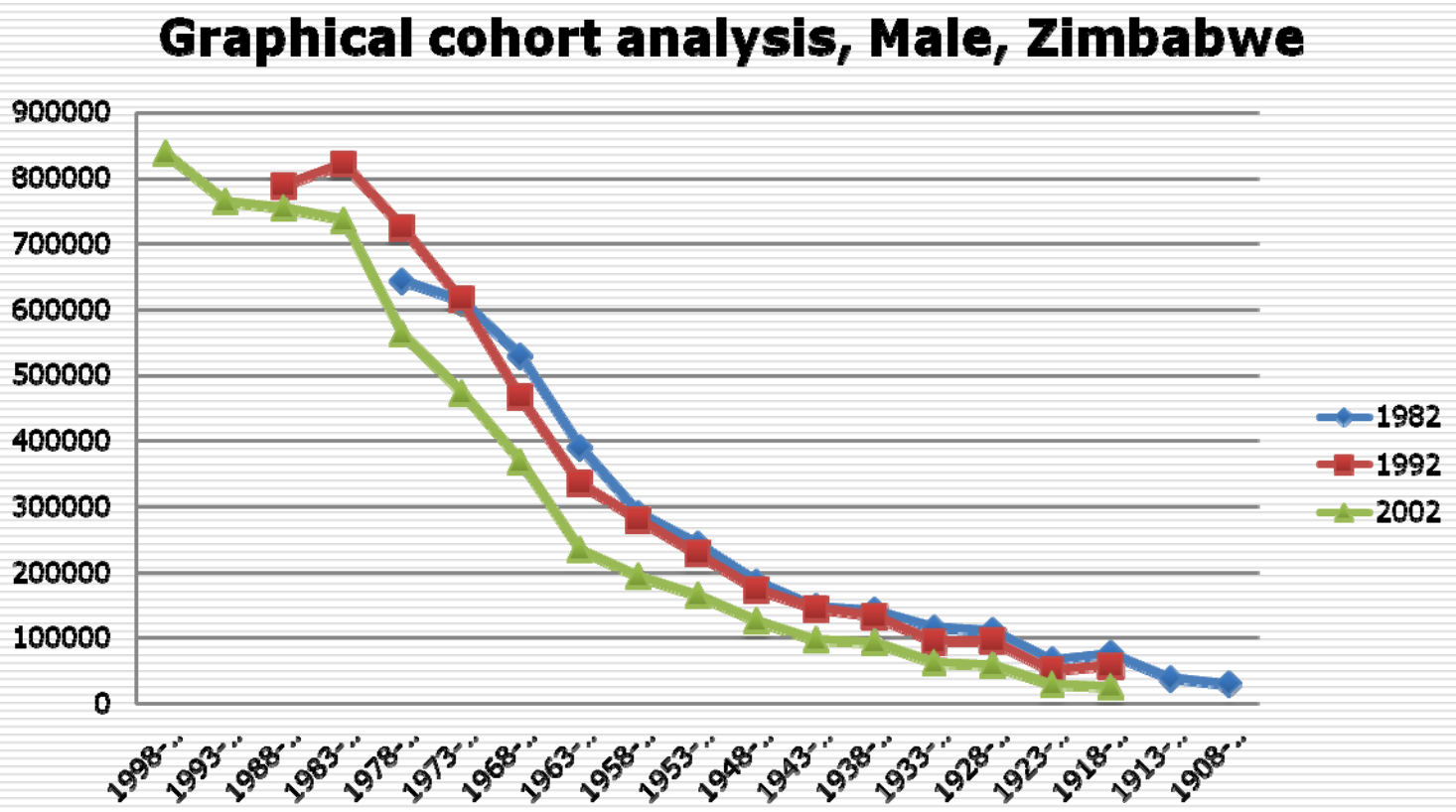
Source: Tabulated using data from *United Nations Demographic Yearbook*

Microsoft Excel — PopPyramids.xls

	A	B	C	D	E	F	G
1	Zimbabwe Graphical Cohort Analysis						
2							
3							
4		1982 Male	1982 Female	1992 Male	1992 Female	2002 Male	2002 Female
5	1998-2002					838062	838007
6	1993-1997					764453	769247
7	1988-1992			788963	795728	754587	757657
8	1983-1987			821319	832469	736686	766890
9	1978-1982	643330	665350	724905	731846	564034	658873
10	1973-1977	612760	619300	615728	632510	473984	513793
11	1968-1972	529750	518740	466837	523060	369836	360291
12	1963-1967	390160	412610	335713	376495	235692	268797
13	1958-1962	290380	364200	280066	326299	194702	239727
14	1953-1957	243420	281060	229360	259555	165437	191168
15	1948-1952	185400	206760	174266	189509	128029	173229
16	1943-1947	147920	170170	145437	143441	98417	112498
17	1938-1942	142050	139530	133261	147339	94447	99420
18	1933-1937	116490	110390	94713	86729	64301	67851
19	1928-1932	111780	90880	95510	84213	60311	62464
20	1923-1927	67400	60800	51202	50902	29997	34473
21	1918-1922	76850	65260	58279	62479	26764	32281
22	1913-1917	38810	38860				
23	1908-1912	29810	30500				

Italy 2010 | Italy for screenshots | Zimbabwe 1982 single year | Zim 1982 5year | Liberia 2008

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

$$ {}_5AR_x = \frac{2 * {}_5P_x}{{}_5P_{x-n} + {}_5P_{x+n}} $$

${}_5AR_x$ = The age ratio for the age group x to x+4

${}_5P_x$ =The enumerated population in the age category x to x+4

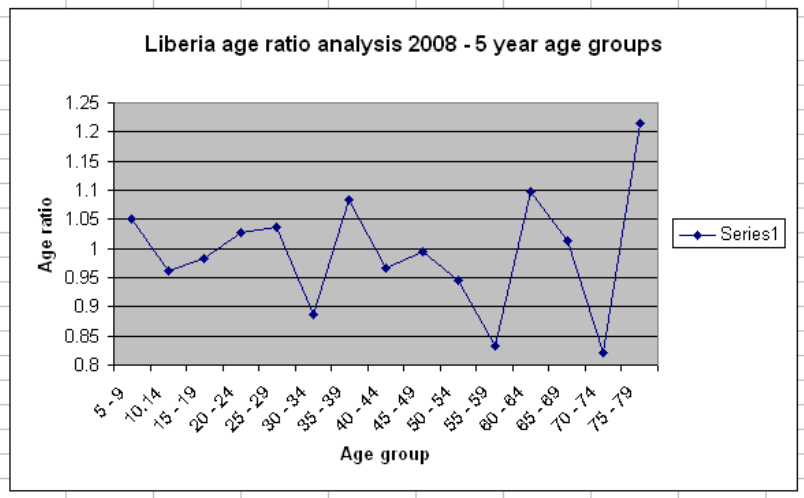
${}_5P_{x-5}$ = The enumerated population in the adjacent lower age category

${}_5P_{x+5}$ = The enumerated population in the adjacent higher age category

Age ratios (3) - example

PopPyramids.xls

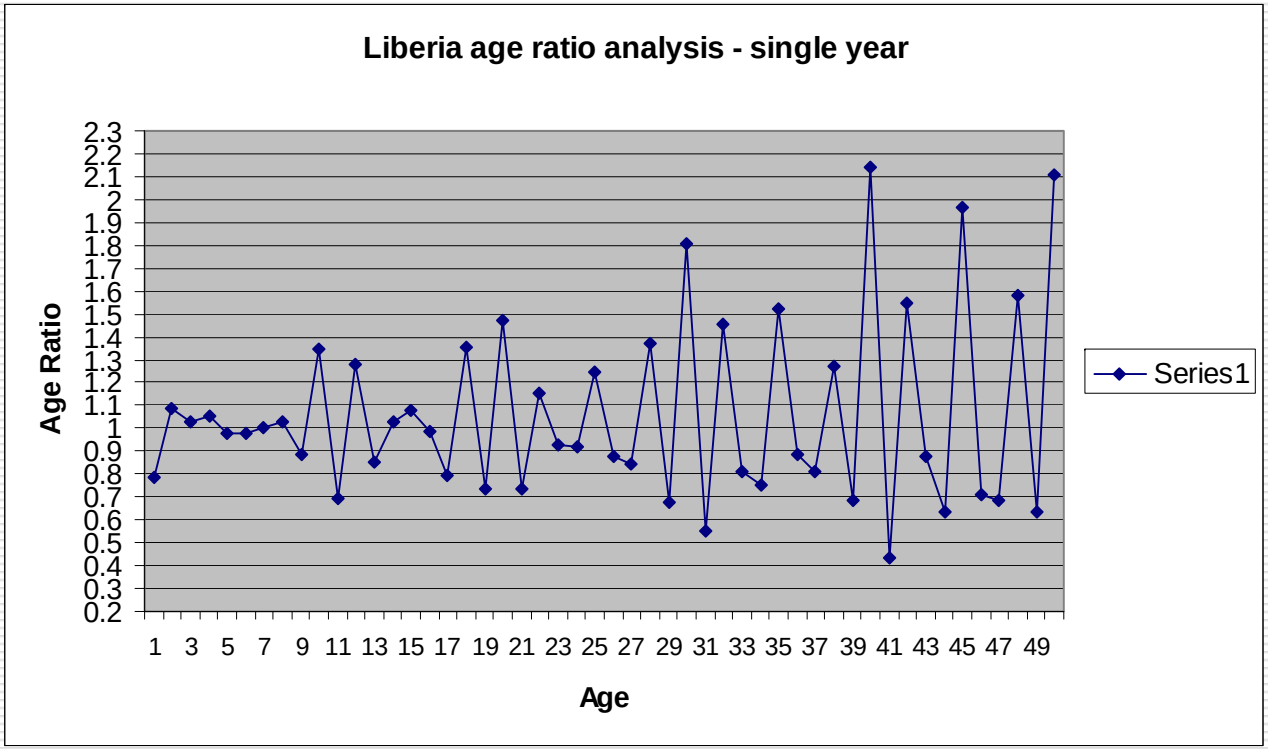
	A	B	C
1	**Liberia**		
2	**B07 Population by age, sex and urban/rural residence**		
3	**Series 200: 21 Mar; Census - de facto - complete tabulation; Final figure, complete**		
4	21 Mar 2008		
5		Total Both Sexes	
6	Age		
7	0 - 4	534475	
8	5 - 9	501931	1.04991
9	10.14	421666	0.960924
10	15 - 19	375695	0.982728
11	20 - 24	342930	1.027424
12	25 - 29	291858	1.037603
13	30 - 34	219632	0.886696
14	35 - 39	203536	1.084458
15	40 - 44	155737	0.966281
16	45 - 49	118807	0.995546
17	50 - 54	82940	0.946442
18	55 - 59	56460	0.831701
19	60 - 64	52830	1.097572
20	65 - 69	39807	1.01321
21	70 - 74	25746	0.820982
22	75 - 79	22913	1.213837
23	80 - 84	12007	



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

Age ratios (4) - example

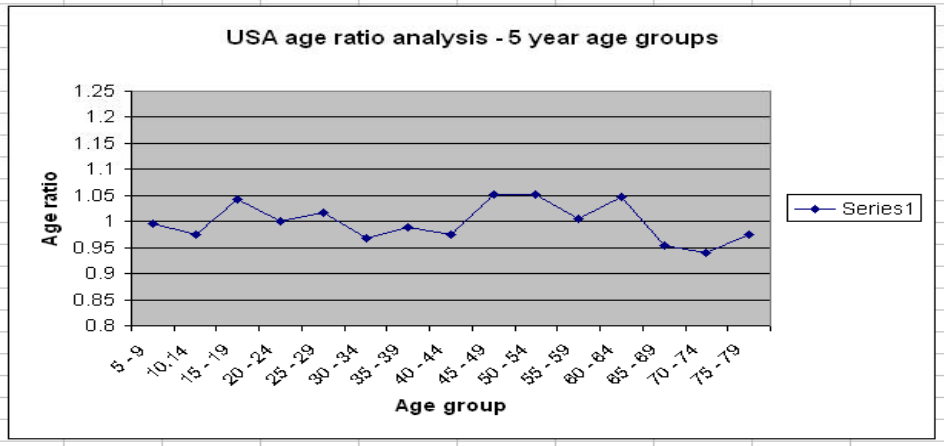
Age	Total Pop	Age Ratio
5	107400	0.97841386
6	101476	0.97463431
7	100834	1.0032036
8	99548	1.02888268
9	92673	0.88733244
10	109332	1.3471583
11	69642	0.6957069
12	90873	1.27942388
13	72411	0.85048831
14	79408	1.0278754
15	82098	1.08140362
16	72428	0.99046838
17	64152	0.79251367
18	89467	1.35862781
19	67550	0.73660106
20	93943	1.47351167



Age ratios (5) - example

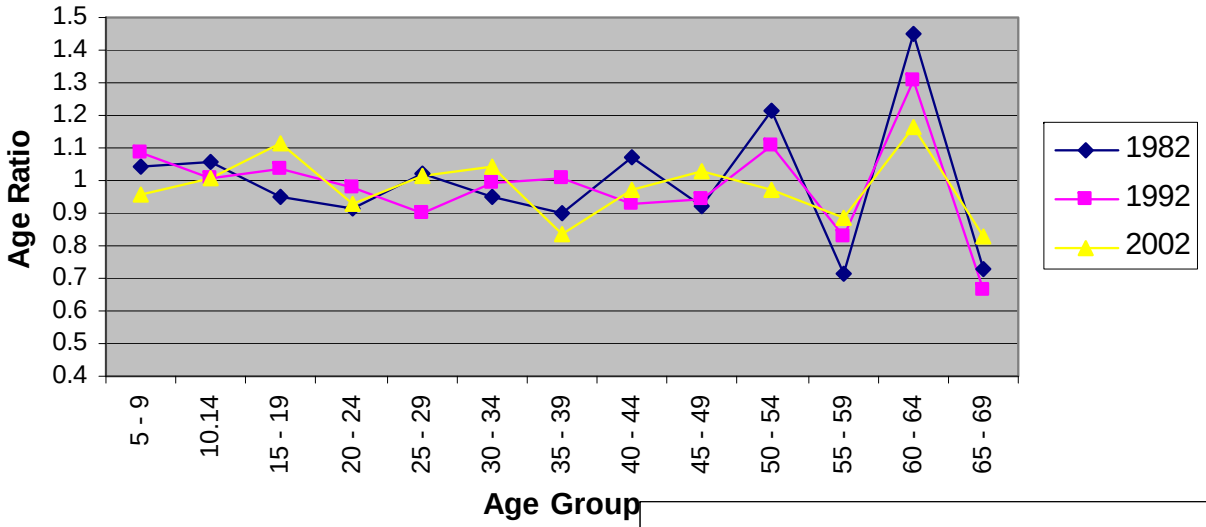
Microsoft Excel — PopPyramids.xls

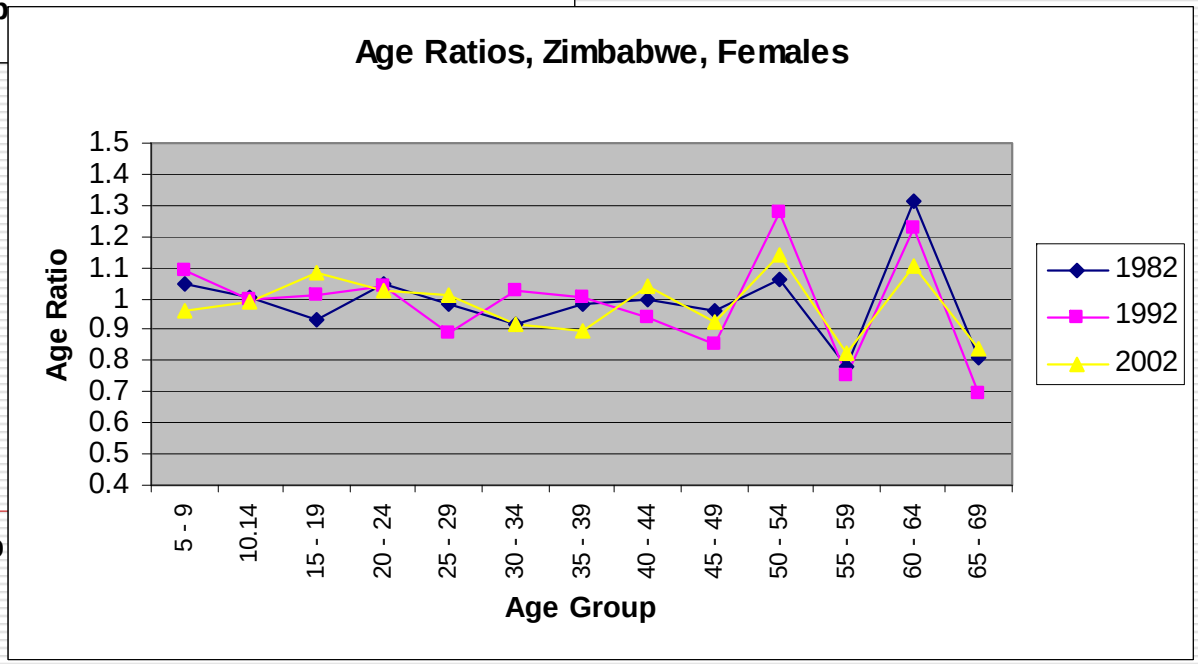
	A	B	C
1	**United States of America**		
2	**B07 Population by age, sex and urban/rural residence**		
3	**Series 335: 1 Apr; Census - de jure - complete tabulation; Final figure, complete**		
4	1 Apr 2010		
5		Total	
6	Age	Both Sexes	
7	0 - 4	20201362	
8	5 - 9	20348657	0.995566331
9	10.14	20677194	0.975592441
10	15 - 19	22040343	1.043004157
11	20 - 24	21585999	1.000690878
12	25 - 29	21101849	1.015779302
13	30 - 34	19962099	0.967121028
14	35 - 39	20179642	0.987913293
15	40 - 44	20890964	0.974204929
16	45 - 49	22708591	1.051589257
17	50 - 54	22298125	1.052458717
18	55 - 59	19664805	1.005459677
19	60 - 64	16817924	1.047843512
20	65 - 69	12435263	0.953036489
21	70 - 74	9278166	0.939415659
22	75 - 79	7317795	0.974309944
23	80 - 84	5743327	1.050135972
24	85 - 89	3620459	1.006844703
25	90 - 94	1448366	0.725688259
26	95 - 99	371244	0.494421767
27	100 +	53364	



Liberia 2008 / South Africa 1996 / Qatar 2010 / Turkey cohort / Liberia AR / **USA AR** / Thailand SR / China SR 1982 / China SR 1990 / China SR 2000

Ready


Age Ratios, Zimbabwe, Males
Age Ratio
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
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
Age Group
1982
1992
2002


Age Ratios, Zimbabwe, Females
Age Ratio
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
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
Age Group
1982
1992
2002

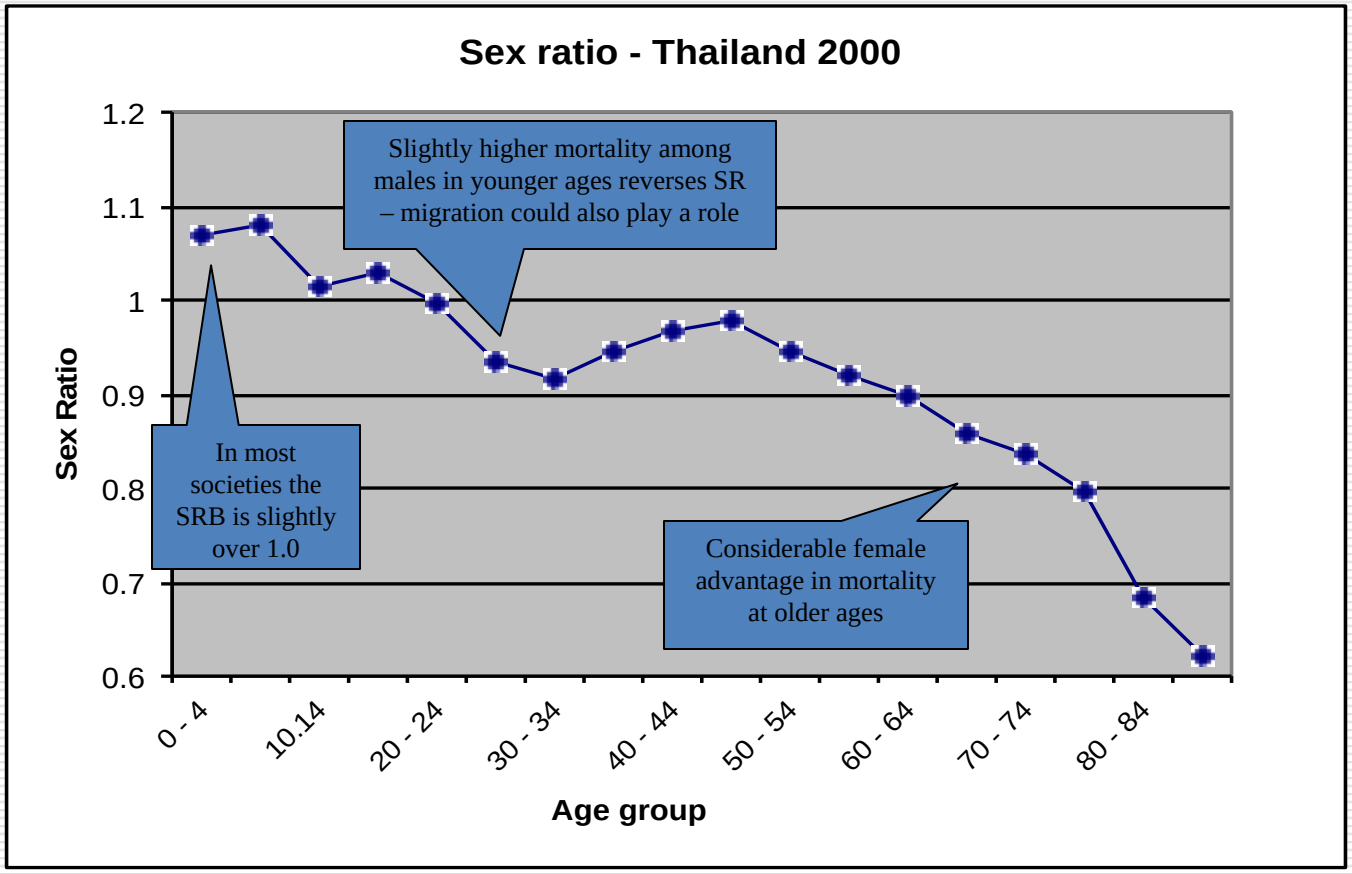
Sex ratios (1) - calculation

Sex Ratio = $_5M_x$ / $_5F_x$

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

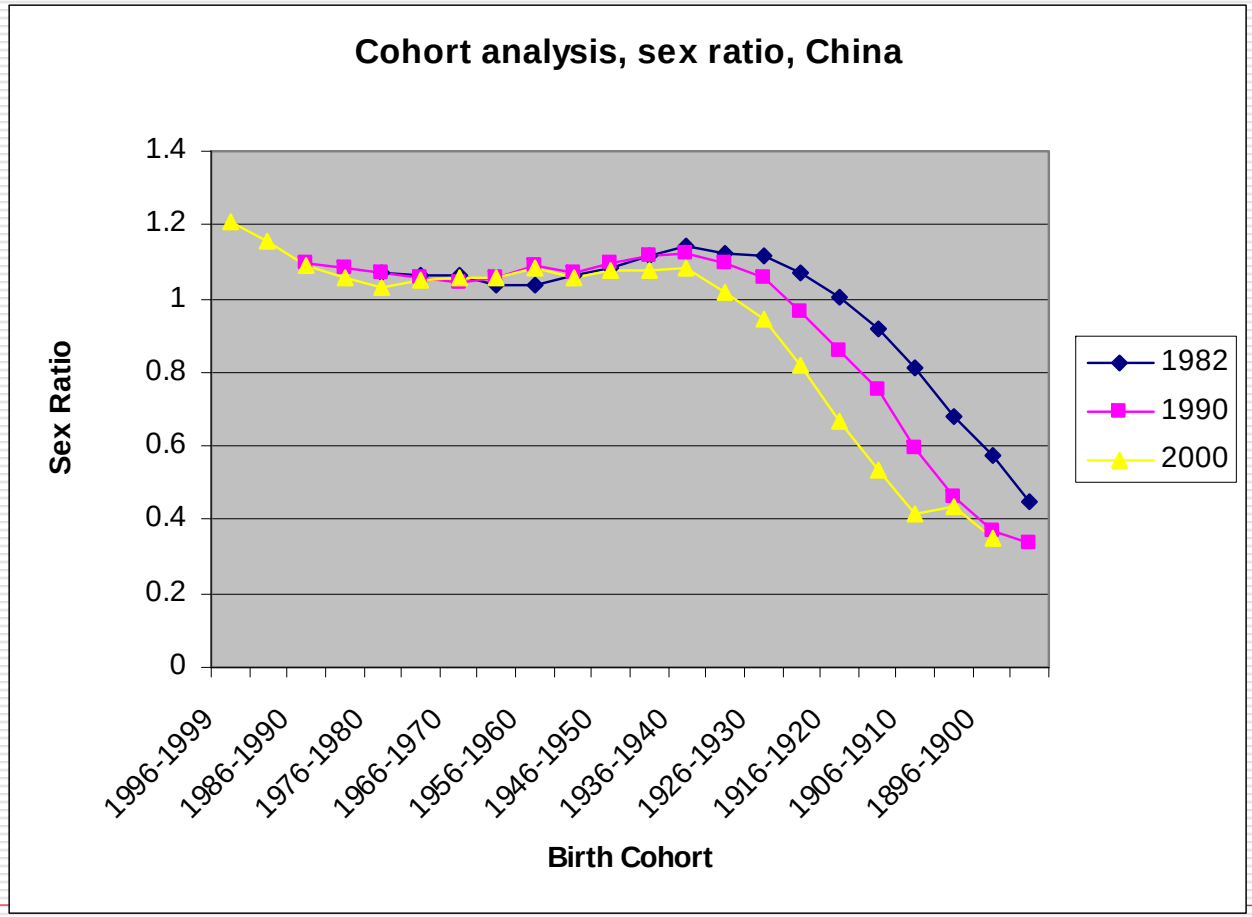
$_5F_x$ = Number of females enumerated in the same age group

Sex ratios (2) - plotting

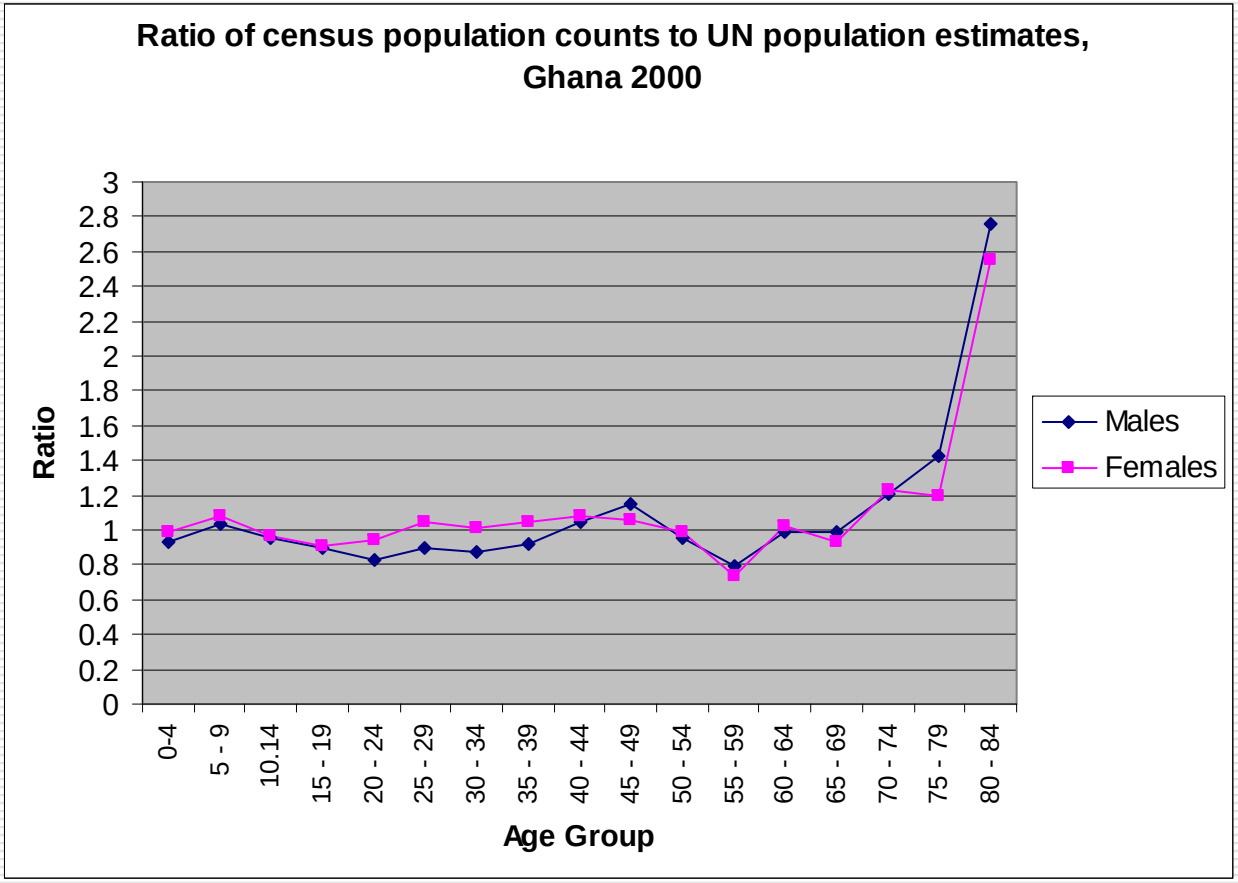


Source: Tabulated using data from *United Nations Demographic Yearbook*

Sex ratios (3) – cohort analysis

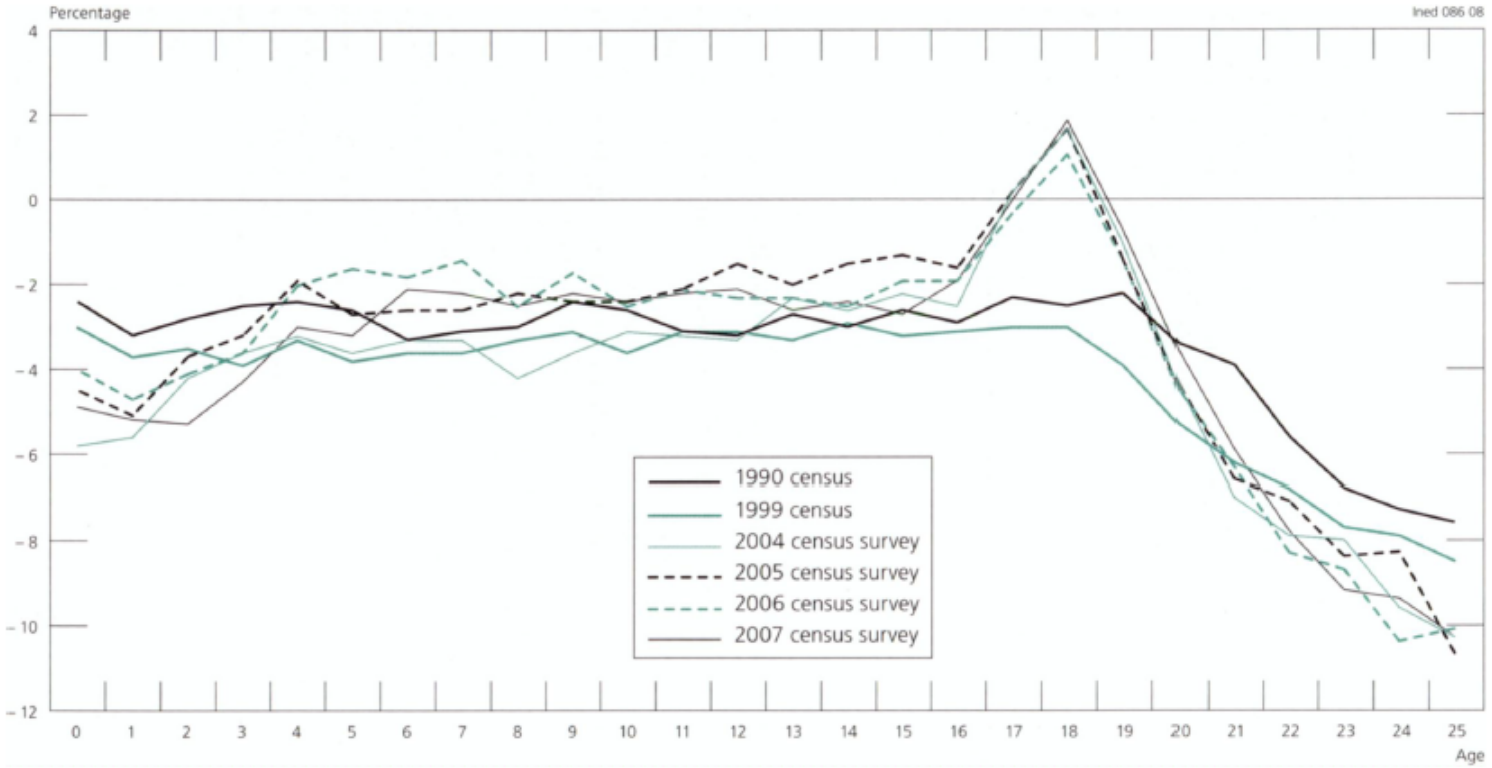


Age & sex ratio comparisons with external sources (1)



Age & sex ratio comparisons with external sources (2)

Figure 2. Comparison between enumerated population and total population inferred from vital records (births minus deaths), for persons aged 0-25 born in metropolitan France (relative differences)



Note: The difference is the coverage difference between the total number of persons counted in the census or annual census survey, and the number of births in metropolitan France for the same cohort, after subtracting the deaths occurring up to the time of the census or annual census survey considered.
Source : INSEE, 1990 and 1999 censuses, EAR 2004-2007 and vital records.

Source: Guy Desplanques and Godfrey Rogers. 2008. Strengths and Uncertainties of the French Annual Census Surveys. *Population* 63(3).

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

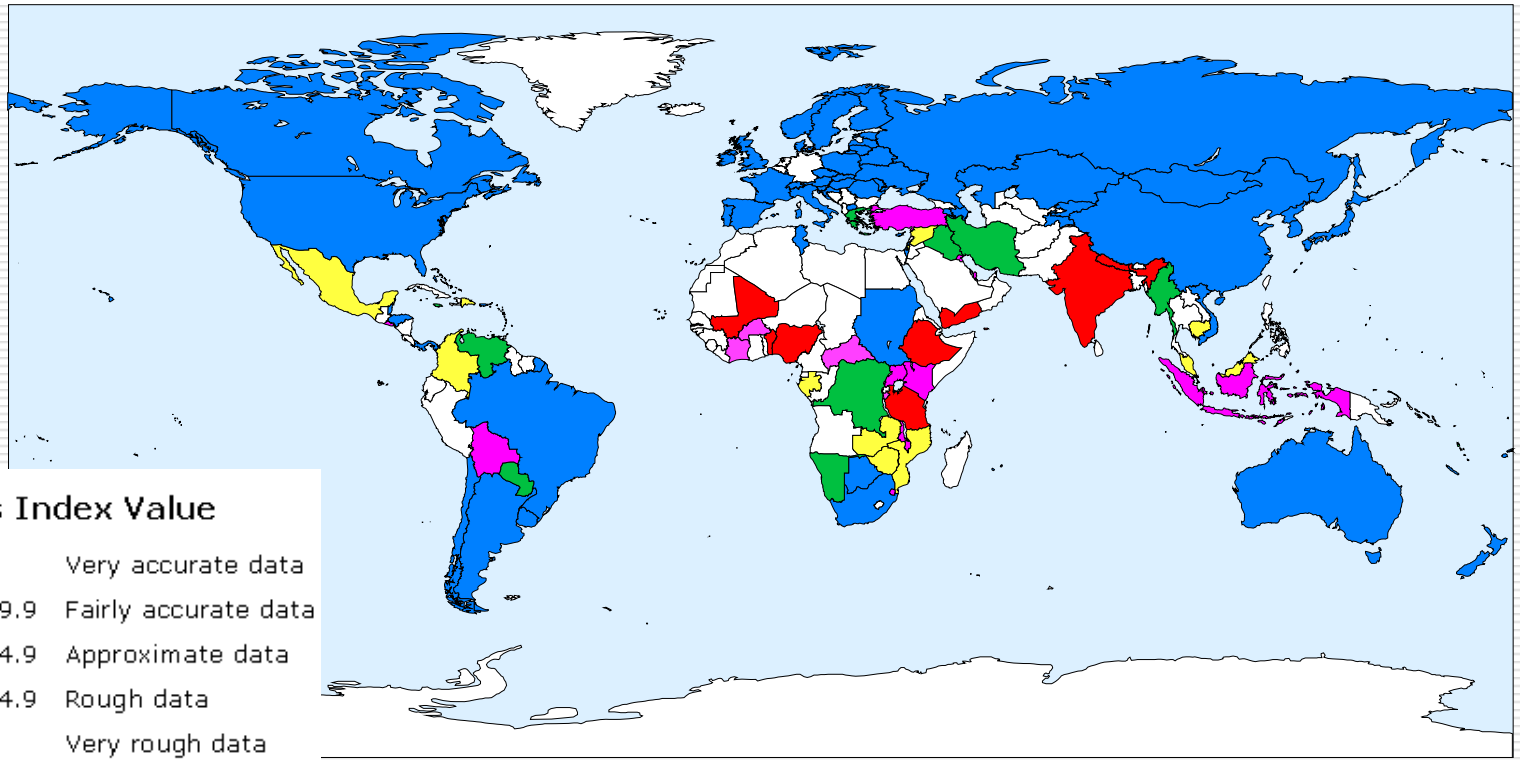
Whipple`s Index (3)

- 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

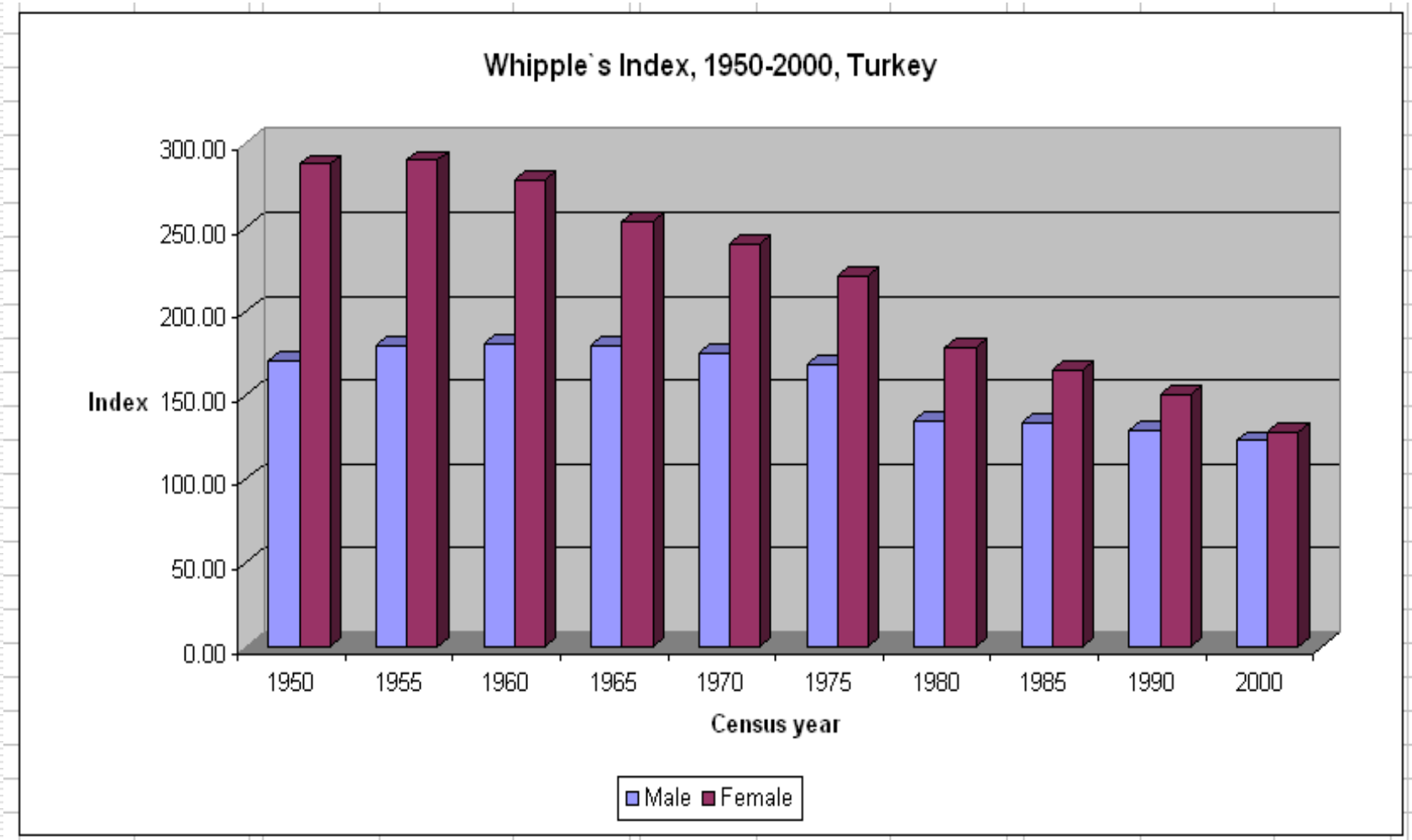
Whipple's index around the world

- 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

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

Myers' Blended Index: Example

	A	B	C	D	E	F	G	H	I	J	K	L
1					Myers' Blended Method - Liberia Census 2008 - Males							
2	Liberia 21 Mar											
3					(1)	(2)	(3)	(4)	(5)	(6)	(7)	
4					Sum of population ages ending in terminal digit		Weights (given)		Blended Population		Deviation from 10%	
5	Age	Male		Terminal digit	From 10+x	From 20+x	Column 1	Column 2	(1)x(3) + (2)x(4)	Percent =row / cell I16	abs((6) - 10)	
6	10	56118		0	197,432	141,314	1	9	1,469,258	14.45	4.45	
7	11	35473		1	105,215	69,742	2	8	768,366	7.56	2.44	
8	12	46839		2	137,505	90,666	3	7	1,047,177	10.30	0.30	
9	13	36026		3	103,187	67,161	4	6	815,714	8.02	1.98	
10	14	40403		4	110,561	70,158	5	5	903,595	8.89	1.11	
11	15	42379		5	140,744	98,365	6	4	1,237,924	12.18	2.18	
12	16	36293		6	103,303	67,010	7	3	924,151	9.09	0.91	
13	17	32750		7	91,697	58,947	8	2	851,470	8.37	1.63	
14	18	44658		8	126,683	82,025	9	1	1,222,172	12.02	2.02	
15	19	33327		9	92,754	59,427	10	0	927,540	9.12	0.88	
16	20	44382		TOTAL					10,167,367	100.00	17.89	
17	21	28639										
18	22	31848							Index of Age Preference = K16/2		8.95	
19	23	27822										
20	24	29260										
21	25	33589										
22	26	25717										
23	27	25187										
24	28	32784										
25	29	23729										
26	30	34119										
27	31	17253										
28	32	23432										

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 (in progress)

  http://demographicestimation.iussp.org/