The Population Census as a Time Machine

Griffith Feeney

Demography - Statistics - Information Technology, Letter No. 4, 15 January 2014

> *The demographic history of a population is inscribed in its age distribution.*
>
> —Nathan Keyfitz

A population census is a "snapshot" of a population at a point in time. How can it tell us about the past? The answer is the demographer's secret weapon: Age.

"Time-plotting" is a method of deriving long term historical trends from population census data by using age as a proxy for historical time. It also provides a method for assessing the quality of census data. Let's look at three examples that show its power.

1. Children ever born: United States

Demographer Norman B. Ryder observed long ago that completed fertility for a birth cohort may be regarded as an estimate of the period total fertility rate at the time the cohort reaches its mean age at childbearing.

This suggests that mean numbers of children ever born to women over age 50 years may be used to study historical fertility trends. The United States censuses of 1910 and 1940–1980 provide a useful test of this idea.

Table 1 shows some conveniently available data. Because we expect most women to complete childbearing by age 45, the table includes women aged 45-49 years.

Consider first the 1970 census numbers. The average age of 45-49 year old women is 47.5 years. Taking the mean age at childbearing to be 30 years, the women in this cohort reached their mean age at childbearing, on the average, $47.5 - 30 = 17.5$ years prior to the census reference time. The census reference time is the beginning of 1 April. In decimal terms this is $(31 + 28 + 31)/365 = 0.247$.

The women in this cohort therefore reached their mean age at childbearing at time $1970.247 - 17.5 = 1952.747$. The 2.63 children per woman for these women is plotted at this time. The three-places-after-the-decimal precision is of course overkill for plotting; its value is allowing us to recover dates from times so expressed.

Because moving up 5 years in age corresponds to moving back 5 years in time, the means for older age groups are plotted at 1947.8, 1942.8, and so on back to 1922.8. The result is shown in Figure 1.

Applying the same procedure to the other censuses gives Figure 2. The plots from the different censuses will agree if the estimates are accurate. Deteriorating quality of reporting of children ever born as women get older will be revealed by discrepancies between the plots.

The consistency of the plots is very good indeed. There is almost no indication of deterioration of reporting of children ever born with increasing age. The greatest difference in the cohort comparisons is for women aged 70-74 years in 1960 with women aged 80-84 years in 1970, and this difference is only 2%.

Table 1 Mean children ever born to native white women: United States censuses of 1910 and 1940-1980

	Census year					
Age	1910	1940	1950	1960	1970	1980
45-49	4.14	2.72	2.25	2.20	2.63	2.99
50-54	4.37	2.84	2.45	2.14	2.40	2.86
55-59	4.68	2.99	2.69	2.24	2.21	2.65
60-64	4.70	3.05	-	2.46	2.17	2.42
65-69	4.85	3.22	-	2.68	2.28	-
70-74	4.88	3.38	-	2.90	2.47	-
75-79	-	-	-	3.05	2.68	-
80-84	-	-	-	3.09	2.84	-

Sources *1980 Census of Population, Volume 1, Characteristics of the Population*, Chapter D, Detailed Population Characteristics, Part 1, United States Summary, Section a: United States Tables 253-310. Table 270, page 1-104, top left. *1970 Census of Population, Subject Reports, Women by Number of Children Ever Born*, Table 4, page 14. it U.S. Census of Population: 1960, Subject Reports, Women by Number of Children Ever Born, Table 4, page 10. *1950 United States Census of Population, Special Report*, P-E, No. 56, Fertility, Table 1, page 5C-17. For 1910 as well as 1940, *16th Census of the United States: 1940, Population, Differential Fertility, 1940 and 1910, Fertility for States and Large Cities*, Table 3, page 13, for 1940, Table 4, page 15-16, for 1910.

The exception is the gap between the last point of the 1910 series (45-49 year old women) and the first point of the 1940 series (70-74 year old women). The implied decline in fertility, 1.5 children per woman per decade, is implausibly rapid. The first value is unlikely to be too high, so perhaps the second is too low.

But the agreement between the earliest two points for the 1960 census (age groups 75-79 and 80-84) and the 1940 points for the same cohorts (age groups 55-69 and 60-64) and the agreement in trend for the 1910s and 1920s cautions against concluding that these older women understated number of children ever born without further analysis.

In any case, this imperfection does not detract from the value of the result. Birth registration in the United States became complete only in the early 1930s. Figure 2 tells a story of fertility decline over the 60 years between 1870 and 1930 that is not readily available from any other source.

There is no apparent logic to the choice of the age groups for the tabulations from the different censuses shown in Table 1. It is a pity that data for older women is not available for all the censuses. Evidently more attention to the details of tabulation planning would have been appropriate.

Too many people doubt the quality of reporting of children ever born data for older women. Doubting data quality is always appropriate, but it is never an excuse for not *looking* at the data—or to publish the data that makes looking possible. Looking at data extends knowledge, when the quality is poor as well as when it is good. Not looking perpetuates ignorance.

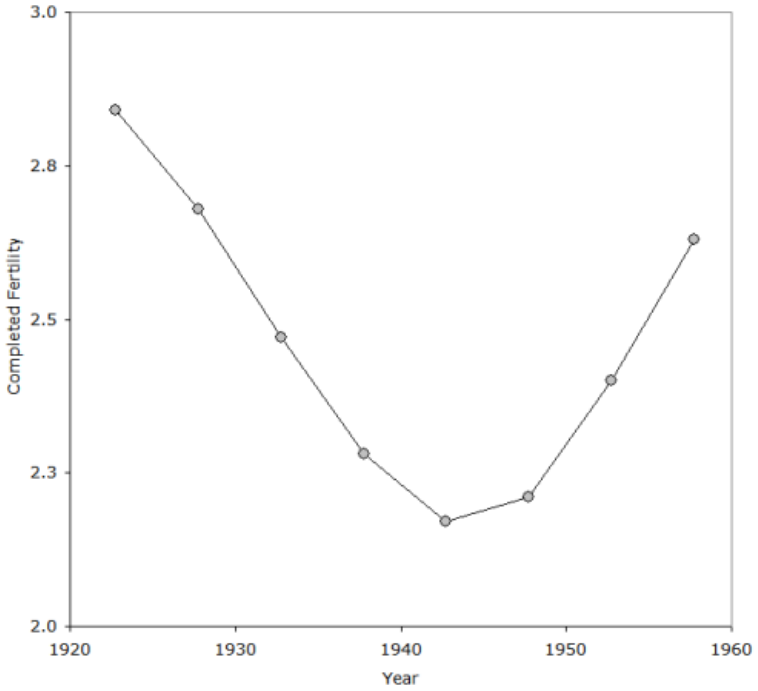

Figure 1 Time Plot of Mean Children Ever Born: 1970 US Census

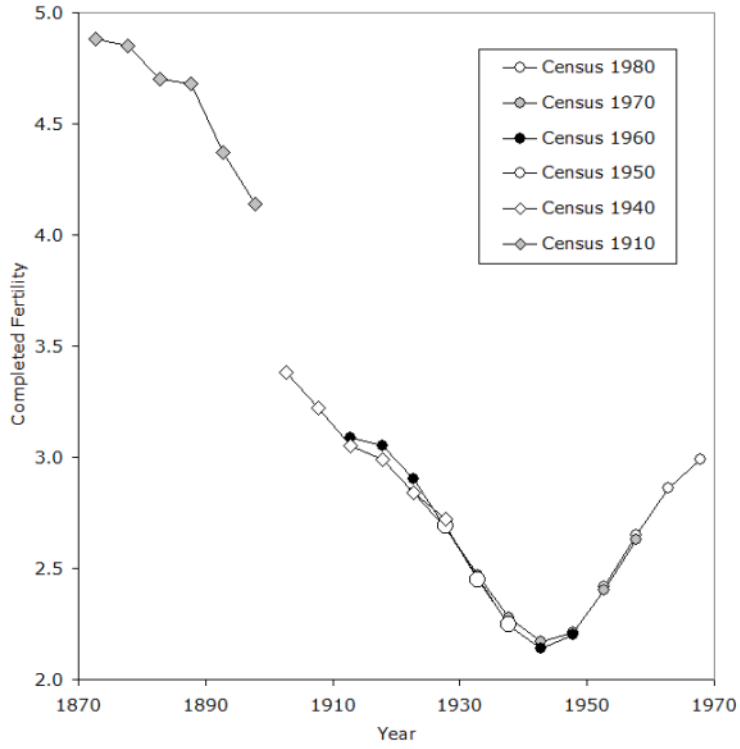

Figure 2 Time Plot of Mean Children Ever Born: United States Censuses of 1910, 1940, 1950, 1960, 1970 and 1980

2. Age-specific growth rates: Thailand

Age-specific growth rates are simple and useful but underutilized statistics. Growth rates for standard 5 year age groups may be regarded as estimates of growth rates in past numbers of births and plotted over time.

Given age distributions from censuses with reference times t_1 and t_2, persons in the 0-4 age group at the first and second censuses were born, respectively, during the periods $[t_1 - 5, t_1]$ and $[t_2 - 5, t_2]$.

The age-specific growth rate for this age group may be regarded as an estimate of the growth rate of numbers of births at the time mid-way between the mid-points of these two periods, which is $(t_1 + t_2)/2 - 2.5$. The growth rate for this age group is plotted at this time.

The reference times for the 1970 and 1980 censuses of Thailand are the beginning of 1 April, corresponding to times 1970.247 and 1980.247. The growth rate for the 0-4 age group is accordingly plotted at time $1975.247 - 2.5 = 1974.747$.

As for the time-plots of children ever born, moving up 5 years in age corresponds to moving back 5 years in time, so the growth rates for age groups 5-9, 10-14, ... are plotted at times 1969.747, 1964.747, ...

The resulting time-plot is shown in Figure 3. It represents nearly a century of historical experience, and a century for which alternative data sources are limited or non-existent.

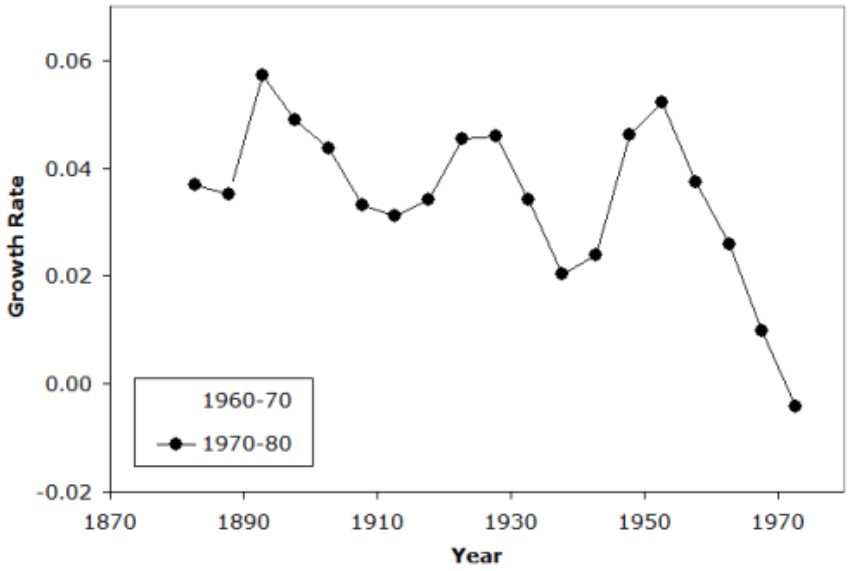

Figure 3 Time Plot of Age-Specific Growth Rates: Thailand Censuses of 1970 and 1980

The magnitude of the swings in age-specific growth rates is substantial, roughly 2-6% per annum through 1960, and below zero for the last point. Because most of the period is pre-demographic transition, one might have supposed it reasonable to assume a stable population.

If the population were stable, however, the age-specific growth rates would be constant over age groups and the time-plot would be a straight line. Figure 3 shows that the population

of Thailand from 1890 forward was very far from stable. What then were the causes of the instability?

As with the children ever born data, two or more plots provide a check on data quality. Figure 4 elaborates Figure 3 by adding the time-plot of age-specific growth rates for 1960–1970.

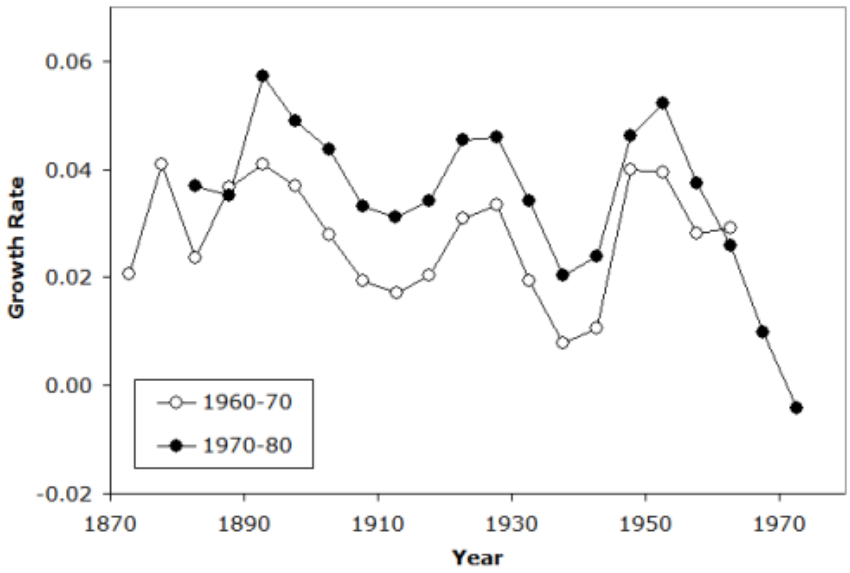

Figure 4 Time Plots of Age-Specific Growth Rates: Thailand Censuses of 1960-1970 and 1970-1980

Sources *Thailand Population Census: 1960, Whole Kingdom*, Tables 2 and 3, Central Statistical Office, National Economic Development Board. *1970 Population & Housing Census, Whole Kingdom*, Tables 3 and 4, pages 9-12, National Statistical Office, Office of the Prime Minister. *1980 Population & Housing Census, Whole Kingdom*, Tables 3 and 4, pages 14-22, National Statistical Office, Office of the Prime Minister.

The 1970–1980 rates are consistently higher than the 1960–1970 rates, and the difference between them is roughly constant over much of the time scale. If the age distributions from the three censuses are accurate, the plots will be close to identical where they overlap. What accounts for the separation?

The 1970–1980 rates are too high. The 1960–1970 rates are too low. The simplest way to account for this would be by higher omission in the 1970 census than in the earlier and later censuses.

This suggests an experiment: multiply the 1970 age distribution by a constant factor chosen to equalize the level of the two series and re-plot. How close to coincidence do we get? What factor achieves this?

Take as a measure of goodness-of-fit the sum of squared squares differences between points plotted at the same time, excluding points where the difference deviates from the general pattern.

The factor that minimizes this measure is 1.07. Multiply the 1970 census population numbers by this factor and recalculate the age-specific growth rates. Re-plotting the growth rates then gives Figure 5.

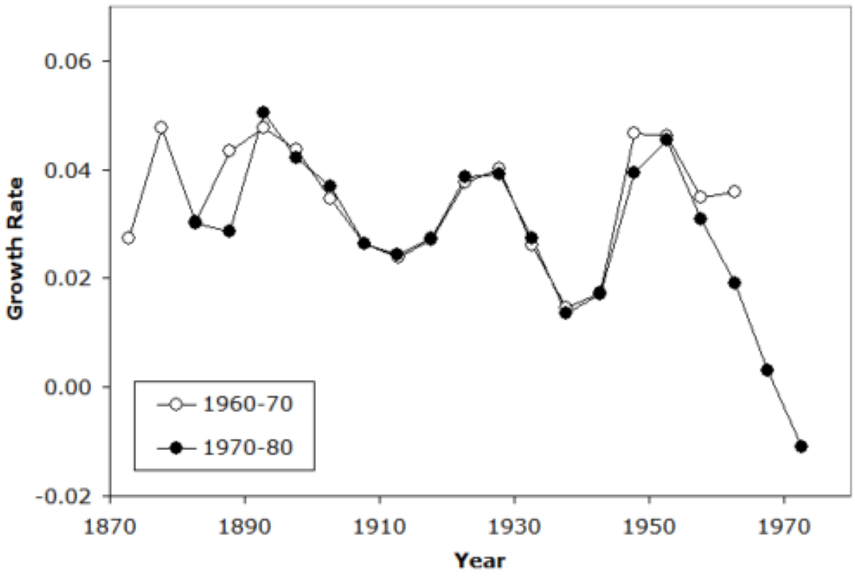

Figure 5 Time Plot of Age-Specific Growth Rates: Thailand Censuses of 1960-1970 and 1970-1980, 1970 census adjusted up by factor 1.07

The remarkably good fit provides persuasive evidence that the 1970 census was a less complete enumeration than the 1960 census and the 1980 census.

The relation between the factor k by which the enumerated population must be multiplied to give the true population and the proportion p of the true population that is omitted is given by the formulas $p = (k1)/k$ and $k = 1/(1p)$. A factor of $k = 1.07$ thus implies an omission p of 6.5% of the 1970 census population relative to the 1960 and 1980 censuses.

3. Literacy: Malawi

Literacy is an important variable for developing countries. Becoming literate, like having a child, is a life-cycle event whose distribution by age is similar between countries and over time.

To time-plot literacy data we assume, initially at least, that the distribution of ages at which persons in a birth cohort become literate has a mean that has been approximately constant over time. The proportion of literate persons for a given age group is then plotted at the time at which these persons reached the mean age of attaining literacy.

Table 2 shows literacy data from the 2008 census of Malawi. The first two columns show population and literate population by 5 year age group to age 80-84. The third column shows the proportion literate, the fourth column the time at which literacy was, on the average, attained.

The time at which 10-14 year olds attained their literacy is calculated by subtracting an assumed mean age at attaining literacy of 10 years from the average age of persons in the

Table 2 Proportions literate by age, with approximate time of attaining literacy: Malawi census of 8-28 June 2008

Age	Population	Literate	Prop	Time
5-9	1,968,299	425,010	0.216	-
10-14	1,670,391	1,273,679	0.763	2005.933
15-19	1,276,692	1,096,575	0.859	2000.933
20-24	1,240,329	1,014,651	0.818	1995.933
25-29	1,102,976	878,815	0.797	1990.933
30-34	827,547	614,351	0.742	1985.933
35-39	623,330	428,281	0.687	1980.933
40-44	441,231	292,602	0.663	1975.933
45-49	343,190	217,381	0.633	1970.933
50-54	269,634	156,618	0.581	1965.933
55-59	258,214	140,295	0.543	1960.933
60-64	184,679	96,931	0.525	1955.933
65-69	153,829	72,498	0.471	1950.933
70-74	106,020	45,226	0.427	1945.933
75-79	106,769	38,850	0.364	1940.933
80-84	55,970	18,513	0.331	1935.933

Source Government of Malawi, National Statistical Office, Population and Housing Census 2008, *Main Report*. Zomba, Malawi: September 2009. Population by age group from Table 5, Page 34, literate population by age group from Table 17, Page 69.

group, 12.5, and subtracting this difference from the reference time of the 2008 census, taken as the beginning of the first day of the enumeration, 8 June.

As for children ever born and age-specific growth rates, moving up 5 years in age corresponds to moving back five years in time, so the remaining times are obtained by repeated subtraction. The resulting time-plot is shown in Figure 6.

Figure 6 shows a nearly linear rise in literacy from 33 percent *circa* 1936 to 86 percent *circa* 2001, the next to last point in the plot. This is a profound achievement of social development, especially given that the population more than doubled during this time period. It is a development success story.

Figure 6 speaks also to social history. There is no apparent break in the increasing production of literate persons. In particular, there is no evidence of a discontinuity before and after the end of colonialism in mid-1964.

This suggests that the forces that propelled rising literacy did not change. What then were these forces, and why did they *not* change? Or, if they did change—is this not more plausible?—why is the effect not visible in the trend?

It is perhaps worth noting, for those of us who live in literate society, how profound a change comes with the possibility of communicating "at a distance" with persons in different places and times. It tends to a profound change in personal outlook. It makes possible a different kind of society.

The usual presentation of literacy data is of proportions of adults literate at successive censuses. The solid horizontal lines in Figure 6 show proportions literate for persons 15 years old and over as of the 2008 and 1998 censuses.

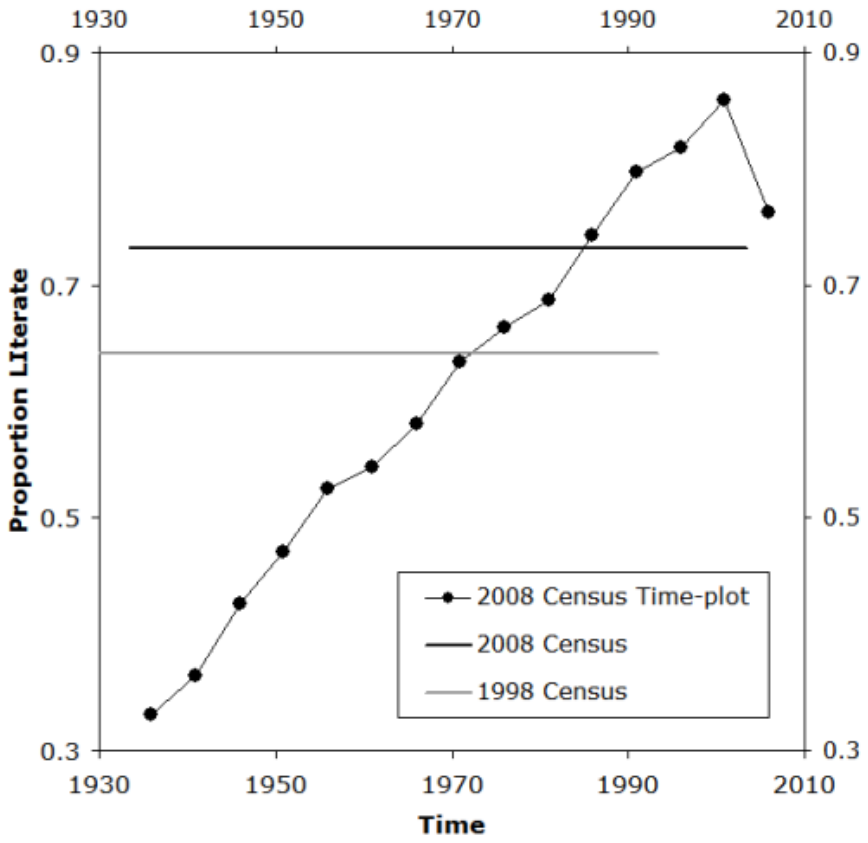

Figure 6 Time Plot of Proportions Literate by Age: Malawi 2008 Census

The rate of increase indicated by comparison of adult literacy proportions for the two censuses is the same as the rate indicated by the time-plot in this example. The change is less impressive than that shown in the time-plot, however, because it refers to change over a single decade. The time-plot shows change over six and a half decades.

Though it does not happen in this example, comparison of adult literacy from from successive censuses may show spurious changes in the rate of change. Adult literacy is a weighted average of literacy for constituent five year age groups, the weights being the proportions of adults in each age group. Changing age distribution may distort literacy levels in the same way that crude birth rates distort changes in fertility levels.

4. How to explain the downturn?

There are several possible explanations for the down turn at the end of the series in Figure 6. One is that the trend of rising literacy was for some reason reversed.

A second is that the mean age at attainment of literacy is higher than 10 years, so that the 10-14 age group includes persons who will become literate, but have not yet attained literacy.

A third explanation is that, even though the mean age at literacy is 10 years, there is a substantial variability about this mean. This also means that some 10-14 year old persons have not yet become literate, though they may be expected to become literate in the future.

It is important to decide which of these explanation is correct. We would be disappointed to see 65 years of success reversed by the last plotted point, and worse than disappointed to find that we drew this conclusion spuriously.

The simplest empirical test is a second time plot for a preceding census, assuming that there is one and that the necessary data is available. If the second plot also shows a downturn at the end of the series, we have solid if circumstantial evidence that the explanation is the inclusion of persons who are too young. Additional plots for earlier censuses may provide stronger evidence.

Further evidence may be provided by data on school enrolment and educational attainment. Details will vary with the particulars of the census questions, but it will usually be possible to obtain a reasonable proxy for the distribution of ages at which literacy is attained, e.g., a table of persons enrolled in school by single year of age and highest level of education entered.

Distributions of persons enrolled in school at different educational levels may however show unexpectedly long tails. Considering only persons in a birth cohort who eventually become literate, for example, the age by which 99 percent have attained literacy may be over 20 years, even if the mean age is 10 years.

If this is the case, avoiding a spurious downturn at the end of the time-plot with five year age groups means restricting attention to persons over age 25 years. This results in a time-plot that ends nearly two decades before the census, a disadvantage if one is interested in recent as well as long term historical trends.

It would be advantageous to have a way of avoiding the spurious downturn without foregoing information on recent trends. Given an approximate distribution of the ages at which literate persons in a birth cohort attain literacy, a factor might be derived by which to adjust proportions literate for younger age groups to compensate for incomplete attainment of literacy. The idea is similar to, and indeed derives from, William Brass's `P/F` ratio method for estimating current fertility.

5. Conclusion: Assessing the quality of census data

We began with the idea of deriving information on historical trends from population censuses, but we have seen that overlapping time-plots from successive censuses provide a method for assessing data quality.

The elaboration of frameworks for assessing the quality of official statistics has had at least one unanticipated, unintended and unwelcome consequence: an implicitly reduced emphasis on assessing *accuracy*. Timeliness, accessability, availability of meta-data and all the rest are indeed important, but how much do they matter if the statistics are seriously inaccurate?

Errors of 1-2 percent are expected and do not impair usability for most purposes. The primary emphasis is on therefore on assessing the likelihood of larger errors.

Classical statistics provides remarkably little help. Sampling theory is irrelevant because the census is a complete enumeration. With an n of millions, tests of significance are unhelpful because even the most trivial substantive differences will be statistically signficant.

Textbooks duly note "non-sampling" errors, but where are the methods for dealing with them? The residual *non-...* definition implies a variety that tends to rule out general

methods. What can be said about "non-linear" models other than that they are ... not linear?

We have seen how time-plotting gives persuasive evidence of the accuracy of average numbers of children ever born in United States censuses.

We have seen how time-plotting age-specific growth rates gives both evidence of differential completeness of enumeration in censuses of Thailand and a method for quantifying the omission.

Time-plotting literacy from multiple censuses, though not illustrated here, similarly provides a general method for assessing the quality of literacy data.

Time-plotting turns population censuses into time machines, but it also provides a general method of assessing errors in census statistics. What has been done for literacy may be extended to educational attainment. What has been done for children ever born may be extended to parity progression ratios, to marriage, perhaps even to divorce.

This is a lot of gain for very little pain; for time-plotting is, after all, a very simple method.

6. Resources

Time-plotting was introduced in my paper "The use of parity progression models in evaluating family planning programs", African Population Conference, Dakar 1988, Volume 3, International Union for the Scientific Study of Population. For a succinct general exposition see Time-plotting Life Cycle Events, 30 May 2009, available on my website demographer.com.

For further applications to children ever born data see The Analysis of Children Ever Born Data for Post-Reproductive Age Women, Notestein Seminar, Office of Population Research, Princeton University, 14 November 1995, and Period parity progression measures of fertility in Japan, NUPRI Research Paper No. 35, Nihon University Population Research Institute, Tokyo, December 1986.

For an application to marital status data see Progression to first marriage in Japan: 1870-1980. Griffith Feeney and Yasuhiko Saito. NUPRI Research Paper No. 24, Nihon University Population Research Institute, Tokyo, March 1985. Figure 1, page 3, time plots proportions of women ever marrying from the 12 population censuses taken between 1920 and 1980.

All of these publications are available on my website gfeeney.com.

A table showing the correspondence between dates and decimal fractions of a year given to three places after the decimal is given in Annex Table I.1 of Methods for Estimating Adult Mortality, published by the United Nations Population Division.

Spreadsheets containing data and sources for the examples given above may be downloaded using the links ceb.us.xls, asgr.thailand.xls and literacy.malawi.xls

The Keyfitz quote is the first line of his paper "On the Interpretation of Age Distributions", Journal of the American Statistical Association, Volume 62, Issue 319, 1967, available on jstore.