REVISED COPY SENT TO MATH POP STUDIES 7-15-88

Recovering Event Histories By Cubic Spline Interpolation

Theodore Bergstrom
and
David Lam

Department of Economics
University of Michigan
Ann Arbor, Michigan 48109

July 1988

(Financial support for this research was provided by the National Institute of Child Health and Human Development, Grant No. 1-R01-HD19624)

ABSTRACT

If event history data are recorded in discrete intervals of time, errors are introduced when the data are converted from the unit in which they were recorded, such as date, to another unit such as age or duration. The problem is illustrated by the inconsistent age at marriage schedules published by two recent U.S. censuses. This paper develops a general method for treating problems of this type using cubic spline interpolation. The method is used to adjust U.S. age at marriage schedules, explaining a substantial part of the discrepancy in the 1960 and 1970 censuses.

KEY WORDS: cubic spline interpolation; event histories; age at marriage; U.S. census;

Introduction

Event histories in census and survey data must usually be recorded in discrete intervals of time. This simple fact can cause much confusion and vexation to users of such data. Consider data on individual histories of some event such as marriage, first birth, or labor force entry for some set of respondents surveyed on a particular date. Data for individuals could be recorded using the age of the respondent at the event, the date at which the event occurred, or the duration of time between the event and some other event such as birth or the survey itself. Given the exact survey date and the respondent's exact date of birth, any one of these three variables (age, date, or duration) determines the other two through simple identities. The choice of the variable in which to record the event may therefore appear to be arbitrary, or may be influenced by *a priori* beliefs that one variable is more accurately reported than the others.[1] The problem is that when any one of the three potential indicators is recorded using some discrete unit of time, such as months, quarters, or years, it becomes impossible to exactly recover the values of the other two indicators, even in units of the same length as those used for the variable recorded. As demonstrated below, the possible values of the two unrecorded variables which are consistent with the value of the one recorded variable always lie in an interval twice as long as the length of the time unit used to record the original variable, and must inevitably lead to ambiguity regarding the correct value for some observations.

This problem is usually dealt with in published data sources by *ad hoc* methods. Sometimes, as in the application to marriage histories we discuss below, the problem is exacerbated by the fact that different data sets relating to the same or similar events use different methods. In this paper, we propose a general method for treating discrete observations of event histories by using the available discrete data to fit a continuous cubic spline function. These continuous distributions can easily be integrated over desired intervals when one wants to reconstitute discrete information in any desired form. This method also allows

[1] Evidence that birth dates tend to be more accurately reported than ages, for example, motivated the U.S. census to change in 1960 from recording respondents' ages to recording respondents' dates of birth.

one to recover an estimate of the underlying continuous distribution of timing of historical events from data sources that have been compiled using different *ad hoc* methods, thus restoring comparability.

The problem of inaccuracies resulting from the use of discrete time in the recording of event histories is a common one in demography. The demographic handbook of Shryock and Siegel (1975) provides several examples and offers some simple solutions. Spline interpolation is also well known to statistical demographers. The technique is briefly discussed by Shryock and Siegel (1975) and is outlined in detail by McNeil et al. (1977). No previous work, to our knowledge, has exploited the application of spline interpolation to the problem of reconstructing event histories recorded in some time unit other than the unit desired by the researcher. As will be demonstrated below, our technique makes it possible to use cubic spline interpolation even when the observed data are known to be inaccurate. By explicitly modeling the systematic sources of error in the data we are able to extend standard spline interpolation through correctly measured data points to the case in which the observed data are known to be incorrect.

Age at Marriage in the U.S. Censuses

The general problem is illustrated by an example that arose in some of our recent research. We wanted to study the distribution of age at first marriage of men and women by year of birth. The U.S. censuses of 1960 and 1970 both publish tables for males and females reporting the joint distribution of age at first marriage and age at census.[2] Since the information collected in both censuses is retrospective, each provides information on the marriage experience of cohorts born as early as 70 years before the census. Our plan was to use information from the earlier census to determine the first marriage rates of cohorts that had few survivors by the time of the later census and to use the later census for marriage rates that depend on weddings that took place after 1960. For some cohorts and ages we would have two independent estimates of marriage rates, one from each census.

Before we could confidently use the data from either census, and certainly before we could graft together results taken from the two sources, we had to investigate the consistency of the two different census reports. When both censuses report estimates of a first marriage rate for the same cohort and age, the two estimates should be almost the same.[3] As it turns out, this is conspicuously not the case. Table 1 shows the marriage rates implied by

[2] The tables appear in the *Age at Marriage* subject reports for both years. See Table 2, U.S. Census Bureau, 1966: 26–37, and Table 2, U.S. Census Bureau, 1973: 29–76.

[3] Differential mortality by age at marriage, due perhaps to the correlation between income and age at marriage, will introduce some discrepancies, but some simple calculations indicate that this is unlikely to be a major factor in the cases analyzed here. Other possible sources of discrepancy between the two censuses include immigration between the censuses, differential underenumeration by age, and reporting and recall bias leading to systematic changes in a cohort's reported marriage history from one census to the next. Sampling error may introduce some discrepancies (the published tables on age at marriage are based on 5 percent samples from the census), but are unlikely to create the kind of systematic discrepancies we consider here.

the 1960 and 1970 census for ages 16 through 34 for white females born in 1914 and 1924.[4] Columns 2 and 5 show the first marriage probabilities implied by the 1960 census for the 1924 and 1914 birth cohorts, while columns 3 and 6 show the probabilities implied by the 1970 census for the same cohorts.[5]

Columns 4 and 7 indicate the discrepancy between the two censuses by the ratio of the 1970 probabilities to the 1960 probabilities. Column 8 shows the average of these ratios for the entire series of cohorts with overlapping data, 1910–1924. The table shows that for women born between 1910 and 1924, the first marriage rates in the late teens as estimated from the 1970 census are on the order of 5 to 15 per cent smaller than the rates estimated for the same cohorts based on the 1960 census. At the same time, the estimates of first marriage rates for women in their middle and late twenties based on the 1970 census are 3 to 15 percent higher than the rates estimated from the 1960 census.

To see the importance of these differences, imagine that a researcher was interested in the relative frequency of teen-age marriages in the 1960's as compared to the 1950's. If he looked at the 1970 census data for women born in, say, 1950, he would find that about 35% of these women married between their 15th and 20th birthdays. Suppose that he found some earlier research based on the 1960 census that reported the marriage rates for this age-group. He would find that about 49% of the women born in 1940 married between their 15th and 20th birthday. But the researcher would have been seriously misled. While it is true that there was a very large decrease in the rate of teen-age marriages over this period, there is reason to suspect that almost half of the apparent change is a statistical illusion due to a change in the census bureau's conventions for reporting data. This becomes apparent when we observe that the 1970 census also reports on the marital history of the 1940 cohort. According to the 1970 census report, about 43% of women born in 1940 married between their 15th and 20th birthday rather than 49% as reported for the same cohort by the 1960 census.

A clue to the source of these discrepancies can be found in a difference in the way that the two censuses converted reported data into published tables. In both the 1960 and 1970 censuses, people were asked the quarter and year they were born and the quarter and year in which they were first married. From these data the Census Bureau constructed tables

[4] The census report of age at census is actually based on respondents' date of birth, reported in quarters. Age is calculated as the difference between the quarter of the census and the quarter of birth, with fractions dropped (see U.S. Census Bureau, 1963: x–xii). It should be pointed out, that the year of birth cohorts so defined are actually for years running from April 1 to March 31 rather than for regular calendar years. To avoid inessential complications, we state some of our later examples as if the year of birth were measured in calendar years.

[5] For example, the .077 probability of marrying at age 17 in Column 2 is derived by dividing 89,475 white women reported as age 35 with age at first marriage 17 in the 1960 table on age at marriage (U.S. Census Bureau, 1966, Table 2, pp. 32–33) by 1,156,770 total white women age 35 reported in U.S. Census Bureau (1963), Table 157, p. 1–358. The .069 probability of marrying at age 17 in column 3 is derived by dividing 79,558 white women age 45 with age at marriage 17 in the 1970 census age at marriage table by 1,146,697 total white women age 45 in the same table (U.S. Census Bureau, 1973, Table 2, p. 59).

Table 1
Age-Specific First Marriage Rates Implied by 1960 and 1970 U.S. Censuses White Female Cohorts, 1914, 1924, and Average 1910–1924

	1924 Birth Cohort			1914 Birth Cohort			Average
							1970/1960
Age at	1960	1970	1970/1960	1960	1970	1970/1960	Ratio
Marriage	Census	Census	Ratio	Census	Census	Ratio	1910-1924
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
16	0.047	0.040	0.851	0.038	0.035	0.921	0.861
17	0.077	0.069	0.896	0.055	0.051	0.927	0.891
18	0.100	0.093	0.930	0.076	0.069	0.908	0.943
19	0.094	0.090	0.957	0.088	0.084	0.955	0.962
20	0.101	0.097	0.960	0.090	0.089	0.989	0.967
21	0.126	0.123	0.976	0.089	0.087	0.978	0.971
22	0.103	0.107	1.039	0.082	0.080	0.976	1.004
23	0.071	0.075	1.056	0.067	0.066	0.985	1.010
24	0.048	0.051	1.063	0.058	0.056	0.966	1.017
25	0.034	0.039	1.147	0.053	0.053	1.000	1.032
26	0.025	0.026	1.040	0.046	0.045	0.978	1.019
27	0.019	0.019	1.000	0.034	0.034	1.000	1.034
28	0.014	0.015	1.071	0.023	0.022	0.957	1.037
29	0.012	0.011	0.917	0.015	0.016	1.067	1.036
30	0.009	0.009	1.000	0.015	0.014	0.933	1.044
31	0.007	0.008	1.143	0.016	0.017	1.063	1.071
32	0.006	0.006	1.000	0.014	0.014	1.000	1.048
33	0.005	0.005	1.000	0.011	0.011	1.000	1.035
34	0.004	0.004	1.000	0.008	0.008	1.000	1.044

Note: The 1924 birth cohort refers to persons aged 35 and 45 at the time of the 1960 and 1970 censuses respectively, i.e. those born April 1, 1924 to March 31, 1925. The 1914 birth cohort refers to persons aged 45 and 55 in the 1960 and 1970 censuses respectively, i.e. those born April 1, 1914 to March 31, 1915.

Source: Numerators for 1960 taken from U.S. Bureau of the Census (1966), Table 2; Denominators (total number in birth cohort at time of census) for 1960 taken from U.S. Bureau of the Census (1963), Table 157. Numerators and denominators for 1970 taken from U.S. Bureau of the Census (1973), Table 2.

reporting "age at marriage" by "age at census." If, for example, a woman reported that she was born in the second quarter of 1930 and first married in the fourth quarter of 1958, then either census would correctly identify her as having been born in 1930 and as having first married when she was 28 years old. But what about someone who was born in the second quarter of 1930 and married in the second quarter of 1958? If her birthday was in May, 1930 and she married in April, 1958 then she was 27 when she married, but if she was born in April, 1930 and married in May, 1958 then she was 28. Whether she was 27 or 28 years old at the time of her wedding depends on which comes earlier in the quarter, her birthday or her wedding day. The census bureau did not have this information and had to allocate persons born in the same quarter of the year to one of the two adjacent age cells according to some more or less arbitrary procedure.

The census bureau made this allocation in a different way in 1970 than it did in 1960. The 1960 census report dealt with the problem of people whose wedding days and birthdays occurred in the same quarter by assuming that half of them were married before their birthdays and half after.[6] The 1970 census report, on the other hand, constructs its tables of the age distribution of marriage based on the assumption that *all* of the people whose wedding date and birthday are in the same quarter were married after their birthdays. The published data from the two censuses, therefore, provide two sets of tables, calculated by different methods, purporting to show the joint distribution of "age at census" and "age of marriage." Looking at the same cohort in the two censuses, we have two alternative estimates of the proportion of the cohort that married at each age. Even abstracting from other potential sources of bias in these estimates, we will show below that neither of the methods used to construct the tables can generate the correct schedule of age at marriage for cohorts. But we will show that it is possible essentially to undo the *ad hoc* methods used in the construction of the census tables and to use these tables to find interpolated continuous-time estimates of the age distribution of first marriage for each cohort. The method that we use has applications to many similar problems, some of which will be described at the conclusion of the paper.

The General Problem of Event Histories in Discrete Time

Suppose that your have data on individual histories of some event i, such as marriage, first birth, or labor force entry, for the population alive at some census or survey date. Denote the exact date of event i as D_i^*, the exact age at which that event occurred as A_i^*, the exact length of time between event i and event j as T_{ij}^*, and the exact length of time between event i and the survey date as T_{ic}^*.[7] Individual data on event i could be recorded using either age, date, or duration. Given the survey date D_c^* and the respondent's exact

[6] *More precisely*, persons who were married in the first and fourth quarter were assumed to have married before their birthdays, while persons married in the second and third quarters were assumed to have married after their birthdays (See U.S. Census Bureau, 1966:x-xi).

[7] By exact date, age, and time, we mean that these variables are measured in continuous time.

date of birth D_b^*, any one of these three variables determines the other two through simple identities. When any one of the three potential indicators is recorded using some discrete unit of time, however, it becomes impossible to exactly recover the values of the other two indicators, even in units of the same length as those used for the variable recorded.

To model the problem formally, suppose we make an arbitrary choice of unit of time, such as months, quarters, or years. We can express the exact date, age, and time elapsed for some event i as the sum of the number of completed units of time plus the fractional remainder. We adopt the notation

$$\begin{aligned} D_i^* &= D_i + d_i \\ A_i^* &= A_i + a_i \\ T_{ij}^* &= T_{ij} + t_{ij}, \end{aligned}$$

where D_i^*, A_i^*, and T_{ij}^* represent the exact date, age, and duration measured in continuous time, D_i, A_i, and T_{ij} are integers representing the number of completed units of time, and $0 \leq (d_i,a_i,t_{ij}) < 1$ represent the fractional remainders. The survey records either D_i, A_i, or T_{ic}, the truncated value of one of the three possible indicators. Suppose the researcher's problem is that the survey records the truncated date of the event, D_i, and the truncated date of birth, D_b, and the researcher is interested in the age of the respondent at the event, A_i. This is exactly analogous to the issue facing the researchers constructing the age at marriage tables from the U.S. Census based on census questions on the date of marriage and date of births.

Consider a person who experienced event i at exact age $A_i^* = A_i + a_i$. Given date of birth $D_b^* = D_b + d_b$, the exact date of event i is $D_i^* = D_b + d_b + A_i + a_i$. The truncated date of the event will be

$$D_i = \begin{cases} D_b + A_i & \text{if} \quad a_i + d_b < 1 \\ D_b + A_i + 1 & \text{if} \quad a_i + d_b > 1 \end{cases} \tag{1}$$

Given the data, a possible estimate of the truncated age at the event is $\hat{A}_i = D_i - D_b$. This is the estimate used without adjustment for the construction of the published age at marriage tables from the 1970 U.S. census. We see immediately from (1) that

$$\hat{A}_i = \begin{cases} A_i & \text{if} \quad a_i + d_b < 1 \\ A_i + 1 & \text{if} \quad a_i + d_b > 1 \end{cases} \tag{2}$$

where A_i is the true truncated age at the event.

If we calculate $\hat{A}_i = D_i - D_b$, what exact ages at the event A_i^* would be consistent with our observed $\hat{A}_i$? We adopt an indicator function $\delta(a_i + d_b)$ such that

$$\delta = \begin{cases} 0 & \text{if} \quad a_i + d_b < 1 \\ 1 & \text{if} \quad a_i + d_b > 1 \end{cases}$$

If $\delta = 1$, then the estimator $\hat{A}_i$ will assign an incorrect integer age at marriage. From (2), the relationship between the exact age at event i and the estimated age at event i is $A_i^* = \hat{A}_i + a_i - \delta$, where the values of both δ and a_i are unknown to the researcher. Although δ and a_i are not independent, any value of a_i is in principle consistent with δ equal to either zero or one. The fact that certain values of a_i are more likely to produce a particular value of δ will be exploited below, but for now we note that we cannot rule out any combinations of the two *a priori*. Since $\delta \in [0, 1]$ and $a_i \in [0, 1)$, we see that $a_i - \delta \in [-1, 1)$, and therefore

$$A_i^* \in [\hat{A}_i - 1, \hat{A}_i + 1). \tag{3}$$

From (3) we see the fundamental problem created by changing the unit of analysis of an event. No matter how accurate the responses or how complete the data, we must inevitably introduce error into our data if we change from the date of an event to the age of the respondent when the event occurred. The range of the possible ages consistent with a given recorded date will always be twice as long as the length of the interval in which the date was recorded. More importantly, a portion of the ages assigned by the estimator $\hat{A}_i$ must be incorrect. As an example, a woman with recorded year of marriage 1957 and recorded year of birth 1940 could in principle have married at any age between her 16th birthday and the day before her 18th birthday, given the possible values for the dates of the two events. Assigning all such women an age at marriage of 17 will inevitably assign some women too high an age at marriage.

Hope for a solution to this problem lies in the fact that all the combinations of dates which cause an individual to be assigned to one of the two categories in (2) are not equally likely. The probability that $\hat{A}_i$ erroneously assigns age $A_i + 1$ is the probability that $\delta = 1$, i.e. the probability that $a_i + d_b > 1$. If we assume that persons with integer date of birth D_b and integer event date D_i were born uniformly over the period $[D_b + 0, D_b + 1)$, then conditional on some a_i, the probability that $a_i + d_b > 1$ is the probability that a_i is greater than $1 - d_b$, where d_b is a random variable with uniform distribution over $(0, 1)$. Given d_b is uniform,

$$\begin{aligned} P[\delta = 0] =& P[a_i + d_b < 1] = 1 - a_i \\ P[\delta = 1] =& P[a_i + d_b > 1] = a_i. \end{aligned} \tag{4}$$

From (4), we see that conditional on exact event age $A_i + a_i$, integer event date D_i, and integer date of birth D_b, the probability that $\hat{A}_i$ incorrectly assigns age $A_i + 1$ goes to one as a_i goes to one, and goes to zero as a_i goes to zero. Persons who experience the event exactly on their birthday will always be assigned the correct integer age. The probability of being assigned an integer age that is one unit too high increases linearly as the length of time between the date of the last birthday and the date of the event increases.

It is clear that any discrete age group A_i will be assigned to some persons whose actual age group at the event was A_i but also to some persons whose actual age group at the event was $A_i - 1$. Once aware of this problem, the researcher's task is to minimize the number of persons assigned incorrect ages. Denote the unconditional probability that event i occurred at age $A_i + a_i$ as $f(A_i + a_i)$. Given a particular estimate $\hat{A}_i = \tilde{A}_i$, the conditional probability that the actual age at the event was some given value in the range of possibilities $(\tilde{A}_i - 1, \tilde{A}_i + 1)$ is the unconditional probability that the event occured at that age times the probability that the individual was assigned $\tilde{A}_i$ given the age at the event:

$$\begin{aligned} P[A_i^* = \tilde{A}_i + a_i | \tilde{A}_i] &= f(\tilde{A}_i + a_i) \cdot P[\delta = 0 | a_i] \\ P[A_i^* = \tilde{A}_i + a_i - 1 | \tilde{A}_i] &= f(\tilde{A}_i + a_i - 1) \cdot P[\delta = 1 | a_i]. \end{aligned} \tag{5}$$

Using the conditional probabilites of incorrect assignments in (4), the probability that the event occurred over some range $(\tilde{A}_i - 1, \tilde{A}_i - 1 + a_i)$ is given by

$$P[A_i^* \in (\tilde{A}_i, \tilde{A}_i + a_i) | \tilde{A}_i] = \int_0^{a_i} f(\tilde{A}_i + a)(1 - a)\, da. \tag{6}$$

and analogously,

$$P[A_i^* \in (\tilde{A}_i - 1, \tilde{A}_i - 1 + a_i) | \tilde{A}_i] = \int_0^{a_i} f(\tilde{A}_i - 1 + a) a\, da, \tag{7}$$

If we were to assume that the event was equally likely between ages $A_i - 1$ and $A_i + 1$, this result implies that we can calculate the exact proportion of all of those with estimated ages at the event $\hat{A}_i = \tilde{A}_i$ who experienced the event over any portion of the range $(\tilde{A}_i - 1, \tilde{A}_i + 1)$. If the event occurs with probability π uniformly over the range, then

$$\frac{P[A_i^* \in (\tilde{A}_i - 1, \tilde{A}_i - 1 + a_i) | \tilde{A}_i]}{P[\hat{A}_i = \tilde{A}_i]} = \frac{\pi \int_0^{a_i} a\, da}{\pi \left[\int_0^1 a\, da + \int_0^1 (1 - a)\, da \right]} = \frac{a_i^2}{2}. \tag{8}$$

So, for example, setting $a_i = 1$, the proportion of those for whom $\hat{A}_i = \tilde{A}_i$ who experienced the event between $\tilde{A}_i - 1$ and $\tilde{A}_i$ (i.e. the proportion incorrectly assigned by $\hat{A}_i$) is equal to 1/2. Under this assumption of uniform rates, then, if we want to reassign individuals who are assigned incorrect ages at marriage by the estimator $\hat{A}_i$, we should simply allocate half of those assigned $\hat{A}_i$ to $\hat{A}_i$ and half of them to $\hat{A}_i - 1$. This is, in a sense, the intuitive solution to the problem. It appears in some form in a number of handbooks on demographic techniques (e.g. Shryock and Siegel, 1971), and roughly corresponds to the method used by the Census Bureau in constructing the published age at marriage tables for 1960. The tables for 1970 were constructed using no correction for misallocation at all.

This simple solution will be unsatisfactory if the probability of the event is not uniform over the interval $(\tilde{A}_i - 1, \tilde{A}_i + 1)$. It may lead to significant biases, for example, if the probability of the event rises or falls rapidly with age (or, analogously, with duration) over some interval. We thus propose a powerful and quite general solution to the problem which allows the probability of the event to vary with age, and furthermore, solves for the exact age profile simultaneously. The technique can be demonstrated by considering the problem of the inconsistent age at marriage schedules based on the 1960 and 1970 censuses.

A Diagrammatic Explanation of Two Different Census Methods

In order to recover the actual age-at-marriage schedules from the published census tables, our goal is to "undo" the different accounting procedures used in the 1960 and 1970 tables. To see how to do this, we first look in detail at how the two methods apply to a specific case, women born in a particular year, who marry before their 18th birthday.

Figure 1 presents a hypothetical frequency distribution of age at first marriage which reflects the fact that the marriage rate increases rapidly between ages 16 and 18. The census bureau knows only the quarter in which each woman was born and the quarter in which she married. Women who marry before age 17 3/4 must necessarily have married in an earlier quarter than the quarter in which their 18th birthdays occurred. In Figure 1, these women are represented in the area S. The method used by either census will correctly record such women as having married before their 18th birthday. Similarly, women who marry after age 18 1/4 will have married in a later quarter than that in which their 18th birthday occurred and will be recorded by either method as not having married before age 18.

The difficulty arises when we consider women who married between ages 17 3/4 and 18 1/4. Consider the following four different histories of birth and marriage for women married between these ages.

Woman A was born in April, 1940 and married at age 17 years and 10 months. This means that she must have married in February, 1958. Since February and April are in different quarters, either census method will correctly ascertain that she married before her 18th birthday.

Woman B was born in June, 1940 and married at age 17 years and 10 months. She must have been married in April, 1958. Since April and June are in the same quarter, there is insufficient information available to determine whether she was married before or after her 18th birthday.

Woman C was born in April 1940 and married at the age of 18 years and 2 months. She must have been married in June, 1958. Since April and June are in the same quarter, there is again insufficient information to determine whether she married before or after her 18th birthday.

Woman D was born in June 1940 and married at the age of 18 years and 2 months.

She must have been married in August, 1958. She would therefore have her wedding day in a different quarter from her birthday and would accordingly be correctly determined by either census method to have married after her 18th birthday.

Insert Figure 1 about here.

How many of the women who marry at age 17 years and 10 months will, like Woman A, have their birthdays in a different quarter from their weddings? A bit of reflection should convince you that it will be those who are born in the last month of any quarter. If the birth rate were constant during each quarter, then this proportion would be 1/3. More generally, with a constant birth rate during each quarter, the probability will be $1 - t/3$ that a woman who marries $t < 3$ months before her 18th birthday will have her wedding day and her birthday in different quarters. The probability is $t/3$ that for her these dates fall in the same quarter. The area A in Figure 1 represents the proportion of women in the cohort who, like Woman A, are married at an age between 17 3/4 and 18 and whose wedding days fall in a different quarter from their 18th birthdays. The area B represents the proportion of women, who, like woman B, are married at an age between 17 3/4 and 18, and whose wedding days are in the same quarter as their birthdays.

By similar reasoning, we see that the proportion of women who marry t months after their 18th birthdays and whose birthdays and wedding days are in the same quarter will be $1 - t/3$. The area C in Figure 1 represents the proportion of women in the cohort who, like Woman C, fall into this group. Finally, the area D represents the proportion of women who, like Woman D, marry at an age between 18 and 18 1/4 but whose birthdays and wedding days are in different quarters.

The 1970 census method treats all women whose wedding days fall in the same quarter as their birthdays as having married after their birthdays. This means that the figures published in the 1970 census for women in cohort i, married before age 18 correspond to the sum of the areas in S and A in Figure 1. The "true" proportion, however, corresponds to the combined areas in S, A, and B. Thus the 1970 procedure incorrectly excludes the proportion corresponding to the area B from its estimate of persons married before age 18. If marriage probabilities were constant between ages 17 and 18, this area would represent 1/8 of the total marriages between 17 and 18, implying that the 1970 accounting procedure would incorrectly assign 1/8 of the women marrying at 17 to age 18. As can be seen in the figure, the area will be still larger if the probability of marriage rises between age 17 and 18, with the exact amount of the error depending on the slope of the age-at-marriage schedule.

From what we know about the way in which the 1970 method distorts the cumulative age distribution of marriage by age, it is quite easy to see the way in which it distorts the frequency distribution by age. In particular, as we have shown, about 1/8 of the women who marry at age 18 are counted as marrying at age 19 and about 1/8 of the women who marry at age 17 are counted as having married at age 18. But the number of women who marry at age 18 is much higher than the number who marry at 17, so that the net effect of the distortion is an underestimate of the number who marry at age 18. More generally, we can expect the 1970 method to understate the actual marriage rate at ages where the marriage rate is increasing with age, and to overstate the marriage rate at ages where the marriage rate declines with age.

We can also use Figure 1 to illustrate the effect of the 1960 procedure. The 1960 census counts half of the women whose 18th birthdays and wedding days fall in the same quarter as having married before their 18th birthday. This corresponds to the area in S plus the area in A plus half the combined area of B and C. The "true" proportion corresponds to the sum of the areas in A and B. The error in the 1960 procedure is, therefore, half of the difference between the area in C and the area in B. This number can be expected to be positive at ages where marriage rates are accelerating and negative at ages when they are decreasing with age. Therefore we would expect the 1960 census procedure to slightly overstate the cumulative proportions of the population married by age for young ages and to slightly understate this proportion for older ages. A systematic direction of bias in the distribution of marriages by single year of age is less clearcut for the 1960 procedure than for the 1970 procedure and is likely at any rate to be small.

It is interesting to notice that the qualitative pattern of discrepancies between the 1960 and 1970 estimates of the distribution of marriages by single years of age reported in Table 1 is exactly what would be predicted by our discussion. The marriage rates for either cohort as reported by the 1970 census are lower for younger women and higher for older women than the 1960 rates.

Correcting the 1960 and 1970 Data

Our method is simply this. We assume that the true cumulative density distribution is piecewise cubic and twice differentiable. For each of the two census years 1960 and 1970, we can use the reported age distributions together with information about how these figures were constructed to determine the parameters of the unique piecewise cubic c.d.f. that is consistent with the reported data. Therefore we find that no permanent damage has been done by the census bureau's *ad hoc* methods for either year. Even though the census bureau used different methods in the two years and even though both methods can give misleading results, we are able to use both kinds of data to construct consistent and conceptually appropriate estimates of the underlying continuous distribution of age at first marriage by date of birth. The mathematical details of this approach are reported in Appendix A.2.

This procedure is known in numerical analysis as "cubic spline interpolation." (Johnson and Reiss, 1982.) It would be possible to pursue our general strategy while making different assumptions about the functional form of the true distribution. For example, an alternative to interpolation would be to estimate a continuous distribution by a low order polynomial that does not have to pass through all of the observations. This might be appropriate if there were reason to think that there was substantial measurement error in the observed frequencies of marriage. In the case of census data collected from very large samples, it seems reasonable that sampling errors can be neglected. This will lead us to an interpolation procedure rather than least squares estimation, although we will take account of the systematic errors built into the data.

Of course there are infinitely many different continuous curves which could be drawn through a finite set of points. In order to select a unique curve, we must impose some criteria of simplicity that limits the number of free parameters of the curve to the number of observed points. One possible approach would be to fit a polynomial of sufficiently high degree through the observed points. This has the serious disadvantage that fitted polynomials of high degree can display very large oscillations between adjacent points, as demonstrated in Johnson and Riess (1982:237-238).

Rather than use cubic splines one could interpolate the cumulative distribution with piecewise polynomials of either higher or lower degree. We are primarily interested in the properties of the frequency distribution of marriage. If the cumulative density function of age at first marriage is a cubic spline then the frequency distribution will be piecewise quadratic and will be everywhere continuously differentiable. Lower order polynomial approximations would lack this smoothness. Higher order polynomials through the data could be determined by a requirement of continuity of derivatives of higher degree. But adding more smoothness in this way would have very little effect on the values taken by the frequency distribution of age of first marriage.

The properties of cubic splines and procedures for fitting them through observed data points are outlined in Appendix A.1. Cubic spline interpolation procedures have convenient matrix representations, and are straightforward to apply. Our problem differs from the standard applications of cubic spline interpolation, however, since the data points we are given by the two censuses are known to be systematically incorrect. We want to recover an original distribution from a set of points which are all erroneous measures of actual points on the distribution. We are able to proceed because, following the logic of the previous section, we can express the observed incorrect data as relatively simple functions of the true underlying distribution. Appendix A.2 provides a detailed presentation of the relationship between the true and observed data under the assumption that the true data can be expressed as a cubic spline. As shown in Appendix A.2, once the erroneous observed data have been expressed as a function of the correct data, it is simply matter of inverting the relationship to recover the actual distribution. The system has a straightforward matrix representation, and is easily operationalized using matrix operations on a microcomputer.[8]

Results

The published age-at-marriage tables for 1960 and 1970 both provide data on age at marriage from age 16 to 34 for single year cohorts born between 1910 and 1924.[9] We have

[8] The *Gauss* matrix programming language turned out to be particularly suitable for this endeavor. A copy of the *Gauss* routine that we wrote for this purpose is available from the authors on request.

[9] The binding constraint is that the 1960 table only gives single year age at marriage up to age 34.

applied our cubic spline adjustment procedure to both sets of data for 1910 to 1924 white female birth cohorts. As explained in detail in Appendix A.2, the adjustments differ for the two sets of data because of the different accounting procedures used to generate the 1960 and 1970 published tables. Table 2 summarizes the results of our procedure. Columns 2 through 7 show our estimated age-specific marriage probabilities for the 1924 and 1914 cohorts and the ratios of the 1970 and 1960 estimates. The table can be directly compared with the published data reported in Table 1. Column 8 shows the average of the ratios of the 1970 and 1960 estimates for all of the 1910-1924 female cohorts.

Figure 2 compares the original published data and the cubic spline estimates graphically by plotting column 8 of Tables 1 and 2. The figure shows the average for all of the 1910-1924 cohorts of the discrepancy in the 1970 and 1960 marriage probabilities, where the discrepancy is given by the ratio of the 1970 estimate to the 1960 estimate. Our procedure greatly improves the consistency of the two series, by eliminating about half the observed discrepancy. As shown in Figure 2, the cubic spline adjustments make their most noticeable improvements at ages 16 and 17, ages where the rapid increase in the probability of marriage should cause the 1970 accounting procedure to most underestimate the proportions marrying at young ages. It would be nice to be able to explain the entire difference between the 1960 and 1970 census estimates. While it is easy to think of additional sources of difference between the series, we have not so far been able to convincingly explain the remaining difference.

Insert Figure 2 about here.

Figure 3 shows the results when the same procedures are applied to census data for white males for the same cohorts. The solid line in Figure 3 indicates that the discrepancies in the male age at marriage data for 1960 and 1970 have a similar pattern to the discrepancies in the female data shown in the solid line in Figure 2, with the male data showing somewhat larger inconsistencies, especially at the youngest ages. The dashed line in Figure 3 shows that our 1960 and 1970 estimates after applying our cubic spline interpolation procedure are generally more consistent than the original series, with the most noticeable improvements again occurring at the youngest ages.

Insert Figure 3 about here.

Extensions

The method developed here for recovering continuous event histories from census survey data has broad applications. For example, an interesting footnote to the inconsistencies in the age at marriage data provide by the 1960 and 1970 censuses is that the 1950 census used still a different procedure. The 1950 census asked respondents their actual ages (rather than year of birth) and the duration of their marriage in single years. An analysis similar to that developed above demonstrates that women of any single year of age and single year of marriage duration could have married at any time during a two year age period. An attractive feature of the technique we have developed is that it is just as easily applied to

Table 2
Age-Specific First Marriage Rates Implied by 1960 and 1970 U.S. Censuses Adjusted Using Cubic Spline Procedure White Female Cohorts, 1914, 1924, and Average 1910-1924

	1924 Cohort			1914 Cohort			Average
Age at Marriage	1960 Census	1970 Census	1970/1960 Ratio	1960 Census	1970 Census	1970/1960 Ratio	1970/1960 Ratio, 1910-1924
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
16	0.047	0.044	0.945	0.037	0.038	1.003	0.942
17	0.078	0.073	0.939	0.055	0.054	0.979	0.943
18	0.102	0.094	0.923	0.077	0.071	0.923	0.954
19	0.092	0.089	0.975	0.088	0.085	0.965	0.967
20	0.101	0.100	0.988	0.090	0.089	0.988	0.966
21	0.130	0.124	0.957	0.090	0.086	0.965	0.963
22	0.102	0.103	1.006	0.082	0.079	0.955	0.990
23	0.071	0.071	1.012	0.066	0.064	0.968	0.988
24	0.047	0.049	1.031	0.058	0.055	0.956	0.993
25	0.034	0.038	1.113	0.053	0.052	0.985	1.005
26	0.025	0.025	0.989	0.046	0.044	0.944	0.990
27	0.019	0.019	0.978	0.034	0.032	0.961	1.011
28	0.014	0.014	1.048	0.023	0.021	0.915	1.010
29	0.012	0.011	0.874	0.014	0.016	1.078	1.013
30	0.009	0.009	1.002	0.015	0.014	0.933	1.021
31	0.007	0.008	1.118	0.016	0.017	1.058	1.048
32	0.006	0.006	0.967	0.014	0.013	0.963	1.026
33	0.005	0.005	0.966	0.011	0.011	0.958	1.000
34	0.004	0.004	0.942	0.008	0.008	0.942	0.981

Note: The 1924 birth cohort refers to persons aged 35 and 45 at the time of the 1960 and 1970 censuses respectively, i.e. those born April 1, 1924 to March 31, 1925. The 1914 birth cohort refers to persons aged 45 and 55 in the 1960 and 1970 censuses respectively, i.e. those born April 1, 1914 to March 31, 1915.

Source: Based on data reported in Table 1, adjusted for misallocation using cubic spline procedure described in text.

this 1950 problem, involving ages and durations, as it was to the 1960 and 1970 problems, which were based on inference from dates.

Analogous problems are common in demographic data. The well known demographic handbook of Shryock and Siegel (1971), for example, provides two simple examples. The first involves interpreting survey data giving the number of children born in the previous twelve month period to women in single year age groups (1971:696). The issue is how to allocate the births according to age of mother at the time of birth. Once again note that births in the previous year to twenty year old women will include on one extreme a birth the day before the survey to a women just short of her twenty-first birthday and on the other extreme a birth a full year prior to the survey to a woman who had just turned nineteen. The Shryock and Siegel solution assumes uniform age-specific fertility rates, and is analogous to the method used in the 1960 age at marriage tables. Our method can be easily applied to the problem, making it possible to relax the unrealistic assumption that fertility rates are constant from (for example) ages nineteen to twenty-one. Our cubic spline procedure will generate a continuous fertility schedule with a functional form flexible enough to closely approximate any empirical age pattern of fertility. To see why this problem falls into the same general class of problems as our census data on age at marriage, think of the problem as resulting from the fact that instead of women being asked the age at which their previous birth occured, women are asked the number of years since their previous birth, with the data recorded in single years of duration.

The second example we draw from Shryock and Siegel is a standard problem in estimating mortality from two censuses (1971:696). We observe the number of 10-14 year olds in the 1950 census and the number of 20-24 year olds in the 1960 census, and assuming no migration want to estimate age-specific mortality rates for the ten year period. We can think of our data as being the size of the 10-14 year old cohort in 1950 and the number who died in the subsequent ten years. It is precisely analogous to the fertility problem above once we recognize that we have data on "duration until death" in ten year intervals instead of direct data on age at death. We note that those in the cohort who died during the period may have died at any age between 10 and just short of age 25.

Shryock and Siegel again present a simple allocation based on an assumption of uniform deaths between ages 10 and 25. This procedure incorporates the fact that deaths at age 15, for example, are more likely to be included in the deaths to this cohort than are deaths at age 10, since most of the cohort had already survived past age 10 at the time of the 1950 census. The procedure does not incorporate the actual shape of the underlying mortality schedule, however, instead imposing the assumption of constant mortality between 10 and 25. While the assumption of constant mortality rates across a fifteen year age group might be tolerable for this particular example, it will obviously be unreasonable at both younger and older ages. More importantly, it is an unnecessary assumption given our more general procedure. We can greatly improve on the Shryock and Siegel approach by expressing the observed deaths as a function of the true continuous mortality schedule, assumed to be a cubic spline, and the probability of observing deaths conditional on the age at which they occur.

While many other examples could be discussed, it is clear that our procedure has applications to virtually all types of event history data. An important feature of the procedure is that it only assumes that the underlying continuous distribution of events can be represented by a cubic spline function, a quite unrestrictive assumption which makes the procedure appropriate for almost any problem. This makes it considerably more powerful than methods which impose particular functional forms or model schedules, such as parametric age at marriage schedules or model life tables. While our procedure is computationally more complex than methods used by the 1960 and 1970 censuses and the simple graphical solutions suggested by Shryock and Siegel, advances in microcomputing allow straightforward operationalization of our procedure and make it difficult to justify continued use of the restrictive and often *ad hoc* procedures commonly found in treatments of event history data.

Appendix

A1. On Cubic Spline Interpolation

This section presents a detailed discussion of cubic spline interpolation through exactly observed data.[10] We will then be able to demonstrate how cubic spline techniques can be used to "recover" a smooth age distribution of marriage function from the data calculated by the *ad hoc* methods of either the 1960 or 1970 U.S. census.

The name "cubic spline" comes from the fact that this procedure closely approximates a technique that has long been used by draftsmen. A draftsman who wishes to plot a smooth curve through a set of $n+1$ observations, (x_i, y_i) for $i = 0, \ldots, n$, will place a set of weights on a thin elastic rod called a spline. The weights are placed in such a way that the rod passes over each of the observed points. The draftsman then traces the curve formed by the rod.

The cubic spline function that we seek will consist of n cubic functions, $F_i(.)$, where the domain of $F_i(.)$ is the closed interval from x_{i-1} to x_i. The parameters are chosen to satisfy the following conditions: The function $F_i(.)$ must pass through the two points, (x_{i-1}, y_{i-1}) and (x_i, y_i). That is, for $i = 1, \ldots, n$:

$$F_i(x_{i-1}) = y_{i-1}$$

and

$$F_i(x_i) = y_i \; . \qquad (A1)$$

The cubic pieces must fit together smoothly. That is for $i = 1, \ldots, n-1$:

$$F_i'(x_i) = F_{i+1}'(x_i) \; . \qquad (A2)$$

and

$$F_i''(x_i) = F_{i+1}''(x_i) \; . \qquad (A3)$$

These conditions constitute $4n - 2$ equations. Since each of the cubic pieces has 4 parameters, we have $4n - 2$ equations in $4n$ unknowns. In many applications, including the one we have in mind, it is reasonable to complete the determination of the system by assuming that the curve is linear at the end points, that is $F_1''(x_0) = F_n''(x_n) = 0$. A cubic spline with these end point assumptions is known as a "natural spline."

In the class of twice continuously differentiable functions passing through these n points, the natural spline has the smallest possible curvature over the interval, where the curvature

[10] The discussion draws heavily on Johnson and Riess (1982).

of a function F over a closed interval $[a, b]$ is defined as the integral of the squared second derivative of F over the interval. It turns out that the physical spline placed by the draftsmen will shift so as to minimize the "strain energy" on itself which is proportional to the curvature over the interval from x_0 to x_n. Thus the equivalence to the draftsman's procedure. In addition to being the "simplest" smooth function fitting the data, the natural cubic spline has the advantage that there is a well developed theory of error bounds.[11]

To simplify notation, we present only the analysis for the case where $x_{i+1} - x_i = 1$ for all i. The results can be readily extended to cases of non uniform intervals. It is convenient to represent the cubic pieces of the spline, $F_i(.)$, running from x_{i-1} to x_i in the parametric form

$$F_i(x_{i-1} + t) = (m_{i-1}(1-t)^3 + m_i t^3)/6 + tc_i + (1-t)d_i \qquad (A4)$$

for t between 0 and 1, where the m's,the c's, and the d's are the parameters to be fit. The parametric form that we have specified guarantees automatically that conditions (A3) are satisfied. This can be verified by evaluating the expressions $F_i''(x_{i-1}+t)$ at $t = 0$ and $t = 1$. It is seen that $F_i''(x_i) = F_{i+1}''(x_i) = m_i$ for $i = 1, \ldots, n-1$. The natural spline imposes the additional assumption that $m_0 = m_n = 0$. This condition, together with the Equations (A1)-(A3) will be sufficient to determine all parameters.

Evaluating Equation 1 for $t = 0$, and $t = 1$, applying (A1), and rearranging terms, one finds that for $i = 1, \ldots, n$

$$d_i = y_{i-1} - m_{i-1}/6 \qquad (A5)$$

and

$$c_i = y_i - m_i/6. \qquad (A6)$$

Equations (A5) and (A6) can then be used to eliminate the parameters c_i and d_i from (A4), implying

$$F_i(x_{i-1} + t) = (m_{i-1}(1-t)^3 + m_i t^3)/6 + t(y_i - m_i/6) + (1-t)(y_{i-1} - m_{i-1}/6). \qquad (A7)$$

Differentiating (A7) with respect to t, we have

$$F_i'(x_{i-1} + t) = (-m_{i-1}(1-t)^2 + m_i t^2)/2 + \Delta\ y_i - (m_i - m_{i-1})/6 \qquad (A8)$$

[11] A detailed discussion of this and of other properties of the cubic spline can be found in Johnson and Riess (1982).

where $\Delta y_i = y_i - y_{i-1}$. Condition (A2) requires that $F_i'(x_{i-1}+1) = F_{i+1}'(x_i)$ for $i = 1, \ldots, n-1$. Calculating the appropriate derivatives and rearranging terms, this condition is equivalent to

$$m_{i-1} + 4m_i + m_{i+1} = 6(\Delta\ y_{i+1} - \Delta\ y_i) \qquad (A9)$$

for $i = 1, \ldots, n-1$. The system of linear equations in (A9) can be written as a matrix expression

$$Qm = 6(\Delta_2\ y + z) \qquad (A10)$$

where the matrix Q is an $(n-1)$ by $(n-1)$ matrix with diagonal entries equal to 4, with entries of 1 on its first superdiagonal and on its first subdiagonal and with zeros everywhere else, where m, y, and z are $n-1$ dimensional column vectors with transposes $(m_1, \ldots, m_{n-1})$, $(y_1, \ldots, y_{n-1})$, and $(y_0, 0, \ldots, 0, y_n)$, and where Δ_2 is the double-lag operator with -2 on its diagonal, with 1's on its first superdiagonal and first subdiagonal and with zeros everywhere else. The matrix Q is nonsingular, so it is possible to solve for m by the matrix equation

$$m = 6Q^{-1}(\Delta_2\ y + z)\,. \qquad (A11)$$

Having solved for m, one can immediately solve for the c_i's and d_i's from Equations (A5) and (A6). The cubic spline function is then fully determined.

A.2 Cubic Splines Consistent With the Census Reports

Our problem is a little more complicated than fitting cubic splines through actual observations because of the different forms in which the census data is reported. First consider the 1970 census. In the construction of the 1970 age at marriage tables, all persons who married in the same quarter of the year as that in which their birthdays occurred are implicitly assumed to have married after their birthdays. The cumulative density function obtained by adding the annual marriage figures reported in the 1970 census tables is, therefore, always an underestimate. The persons who married before any age i but are not counted as having done so by the 1970 census tables are those whose birthdays and wedding days fall in the same quarter with their birthdays coming earlier in the quarter than their wedding days.

We can correct for this underestimate if we assume that for any cohort, the birthrate is constant during each quarter (but might differ from quarter to quarter) and that the probability of marrying at a given age is independent of the time of the year in which one is born. Notice that all of the persons who are incorrectly excluded from the count of persons having married at age i will have married at ages in the interval from $i - 1/4$ to i. Of those persons who marry at age $i - 1 + t$ where $3/4 < t < 1$, only those whose birthdays fall sufficiently late in the quarter in which they are born will have their wedding days in the same quarter as their birthdays. For example, someone who marries at age 17 years and 10 months will have her birthday and her wedding day in the same quarter only if her birthday falls in the last month of the quarter in which she was born. Given our assumption of a uniform birth rate over the year, the probability that someone is born in the last month of a quarter is 1/3. Hence about 1/3 of the women who married at age 17 years and 10 months are excluded from the 1970 census count of persons married before age 18. More generally, of the persons who marry at age $i-1+t$ where $3/4 < t < 1$, the fraction $4t-3$ will be incorrectly counted as having married after their ith birthdays.

Referring back to Figure 1, let us denote the *number* of women in a given cohort who fall into the regions corresponding to S, A, B, C, and D for age i, as $n_S(i)$, $n_A(i)$, $n_B(i)$, $n_C(i)$, and $n_D(i)$. As we have argued, the cumulative age distribution of marriage reported by the 1970 census understates the number of marriages before age i by the amount $n_B(i)$. We can derive an exact expression for this area given the assumptions that the cumulative marriage schedule is a cubic spline and that births are uniform within any given quarter. Let us define e_i to be the number of persons who married before age i and are counted by the 1970 census as having married after age i. If the cumulative age distribution of marriage is $F(t)$, then

$$e_i = n_B(i) = \int_{3/4}^{1} F'(i-1+t)(4t-3)\,dt\,. \tag{A12}$$

Now if F is a cubic spline function, we can use the explicit form of $F'(.)$. In particular, for ages $i-1+t$ where $0 < t < 1$,

$$F_i(i-1+t) = (m_{i-1}(1-t)^3 + m_i t^3)/6 + tc_i + (1-t)d_i\,. \tag{A13}$$

Differentiating this expression and substituting into (A12) allows us to write

$$e_i = \int_{3/4}^{1} ((-m_{i-1}(1-t)^2 + m_i t^2)/2 + c_i - d_i)(4t-3)\,dt\,. \tag{A14}$$

Calculating this integral, we find that

$$e_i = \frac{81}{1536}m_i - \frac{1}{1536}m_{i-1} + \frac{1}{8}(c_i - d_i)\,. \tag{A15}$$

Finally, we can use Equations (A5) and (A6), to eliminate c_i and d_i from (A15). We then have for $i = 1, \ldots, n$,

$$e_i = \frac{49}{1536}m_i + \frac{31}{1536}m_{i-1} + \frac{1}{8}(y_i - y_{i-1})\,. \tag{A16}$$

The equations in (A16) imply the following matrix equation where e and y are the $n-1$ vectors whose transposes are $(e_1, \ldots, e_{n-1})$ and $(y_1, \ldots, y_{n-1})$,

$$e = Wm + (1/8)\Delta\ y \tag{A17}$$

where W is the $n-1$ by $n-1$ matrix with diagonal entries equal to 49/1535, with entries on its first subdiagonal equal to 31/1536 and with zeros elsewhere, and where Δ is the "lag operator", a matrix with 1's on the diagonal, -1's on the first subdiagonal, and zeros everywhere else.

Denote the number of persons reported by the 1970 census to have married before age i as r_i, with r representing the $n-1$ vector whose transpose is $(r_1, \ldots, r_{n-1})$. The actual marriage rates, y, are related to the numbers reported by the census according to the simple vector equation

$$y = r + e\,. \tag{A18}$$

Substituting from (A17) into (A18), we have

$$y = r + Wm + (1/8)\Delta\ y\,. \tag{A19}$$

But we can replace m in Equation (A19) by using Equation (A11). This gives us

$$y = r + 6W(6Q^{-1}(\Delta_2\ y + z)) + (1/8)\Delta\ y\,. \tag{A20}$$

Collecting the terms involving y and rearranging, we have

$$(I - S)y = r + 6WQ^{-1}z \tag{A21}$$

where

$$S = W(6Q^{-1}\Delta_2 + (1/8)\Delta)\,. \tag{A22}$$

Finally, then we can solve for the "true" age distribution y by simple matrix inversion:

$$y = (I - S)^{-1}(r + W6Q^{-1}z)\,. \tag{A23}$$

We have already computed the entries in the two square matrices in (A23). The vector r is known, since it is just the vector of cumulative marriage rates reported by the 1970 census. The vector z has zeros everywhere except possibly in its first and last entries which are respectively, y_0 and y_n. For our problem, we can without serious distortion assume that $y_0 = 0$ and that y_n is observed correctly as r_n. Therefore, the actual work of solving for the "true" cumulative age distribution of marriage for any cohort is simply a matter of a few matrix operations.

We can calculate estimates of the age distribution of marriages based on the 1960 census tables by using similar methods. Referring back to Figure 1, the cumulative distribution at a given age implied by the 1960 tables erroneously omits half of the marriages in area B and erroneously includes half of the marriage in area C. Where e_i^* denotes the error in the 1960 census report, we have

$$e_i^* = (n_B(i) - n_C(i))/2 \qquad (A24)$$

We can write an integral expression for $n_C(i)$ and solve the integral explicitly in much the same way as we did in Equations (A12)–(A16) where we solved for $n_B(i)$. When this is done, we find that

$$n_C(i) = \int_0^{1/4} F_{i+1}(i+t)'(1-4t)\,dt = \frac{49}{1536}m_i + \frac{31}{1536}m_{i+1} - \frac{1}{8}(y_{i+1} - y_i) \qquad (A25)$$

Let e^* be the vector whose transpose is $(e_1^*, \dots, e_{n-1}^*)$ and let y^* be the transpose of $(y_1^*, \dots, y_n^*)$ where y_i^* is the true fraction of the population married by age i. Making appropriate substitutions, we find that y^* satisfies the matrix equation

$$e^* = W^*m - (1/16)\Delta_2\, y^* \qquad (A26)$$

where W^* is an $(n-1)$ by $(n-1)$ matrix with the elements on its first superdiagonal equal to $-31/3072$, elements on its first subdiagonal equal to $31/3072$ and zeros elsewhere, and where Δ_2, m, and y are defined as before. We can then proceed exactly as we did with the 1970 data to substitute for m and solve for y. We define

$$S^* = W^*(6Q^{-1}\Delta_2 - (1/16)\Delta_2)\,. \qquad (A27)$$

Let r^* be the age distribution reported by the 1960 census, and let z^* be a vector with zeros except for its last entry, which is r_n^*. Then in exact analogy to the derivation of (A23),

$$y^* = (I - S^*)^{-1}(r^* + W^*6Q^{-1}z^*)\,. \qquad (A28)$$

This completes our task of showing how to use cubic spline functions to estimate continuous cumulative distribution functions of age of marriage and to make the data reported from the two different census methods comparable.

A.3 The piecewise linear case

If we were willing to assume that the density function of age of marriage were a step function and hence the cumulative distribution function was piecewise linear, the computations we made in the last section would be greatly simplified. The cumulative density function of marriage rates as a function of age would then be just

$$F(i-1+t) = y_{i-1} + t(y_i - y_{i-1}) \tag{A29}$$

for all ages i and any t between zero and 1.

The number of persons incorrectly excluded from the 1970 report of people married by age i is still described by (A12) above. But in the piecewise linear case, the expression for F' is much simpler, and we can replace (A14) by the following

$$e_i = \int_{3/4}^{1} (y_i - y_{i-1})(4t-3)\,dt\,. \tag{A30}$$

When we perform this integration we find that

$$e_i = 1/8\,(y_i - y_{i-1})\,. \tag{A31}$$

The counterpart of Equation (A19) is simply

$$y = r + 1/8\Delta\ y \tag{A32}$$

and so the corrected solution for y, corresponding to (A24) is just

$$y = (I - 1/8\Delta)^{-1} r\,. \tag{A33}$$

A similar simplification can be made for computing the 1960 corrections.

In this case, we find that the necessary correction is

$$e_i^* = -\frac{1}{16}\Delta_2\ y^* \tag{A34}$$

and hence the corrected estimate for the cumulative age distribution function for marriages based on the 1960 census data is

$$y^* = (I + \frac{1}{16}\Delta_2)^{-1} r^*\,.$$

References

Johnson, L.W. and Riess, R.D. (1982) *Numerical Analysis.* Reading, Mass: Addison-Wesley.

McNeil, D.R., Trussell, T.J., and Turner, J.C. (1977) Spline interpolation of demographic data. *Demography* 14(2): 245–252.

Shryock, H.S., Siegel, J.S., and Associates (1971) *The Methods and Materials of Demography.* Washington, D.C.: U.S. Bureau of the Census, U.S. Government Printing Office.

U.S. Bureau of the Census (1963) *U.S. Census of Population, 1960, Detailed Characteristics: United States Summary,* Final Report PC(1)-1D. Washington, D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census (1966) *U.S. Census of Population, 1960, Subject Reports: Age at First Marriage,* Final Report PC(2)-4D. Washington, D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census (1973) *U.S. Census of Population, 1970, Subject Reports: Age at First Marriage,* Final Report PC(2)-4D. Washington, D.C.: U.S. Government Printing Office.

Figure 1.
Cumulative Proportion Married by Age and
Proportion Misallocated by 1960 and 1970 Censuses

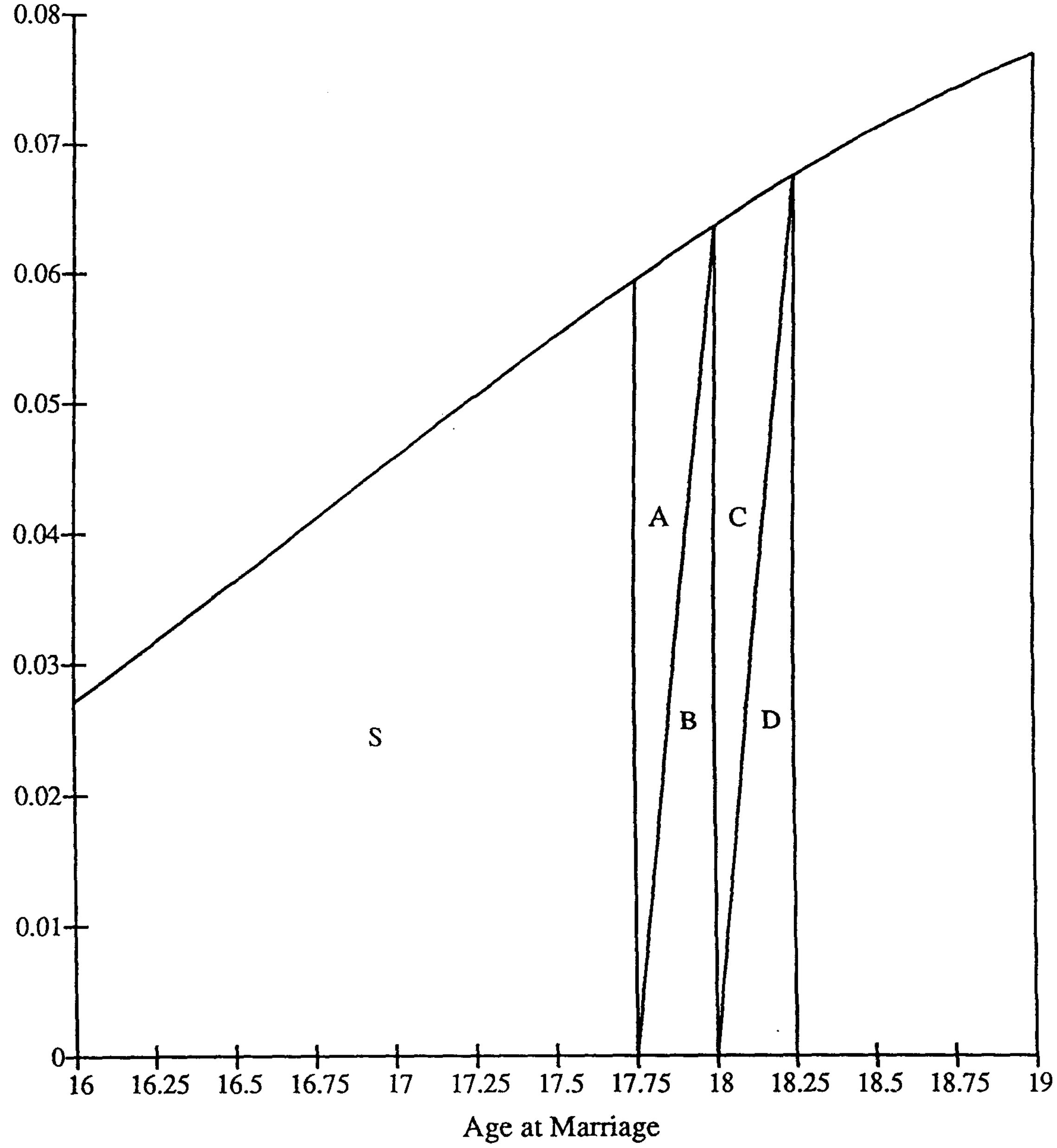


Figure 2.
Discrepancy in Estimated Proportion Marrying by Age.
Ratio of 1970 Estimate to 1960 Estimate.
Average for 1910-1924 White Female Cohorts.

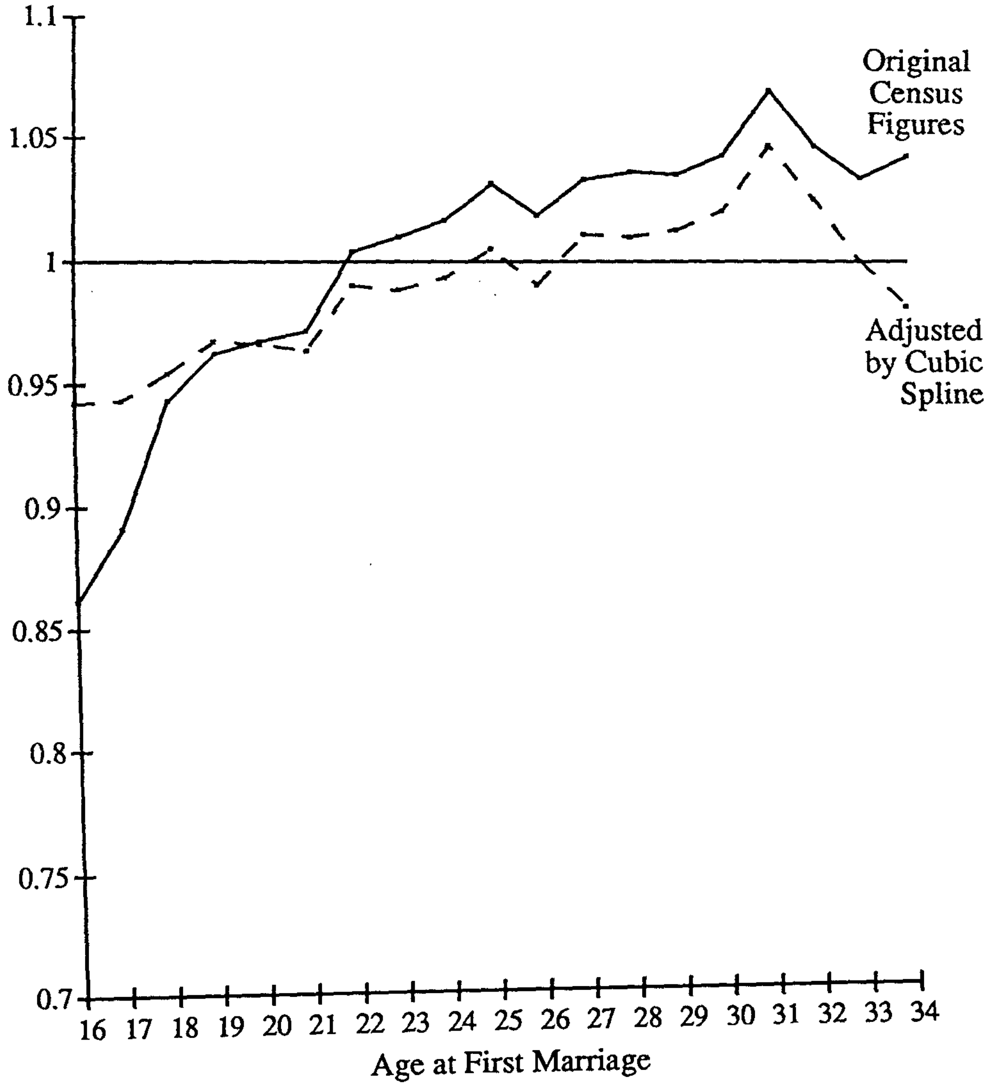
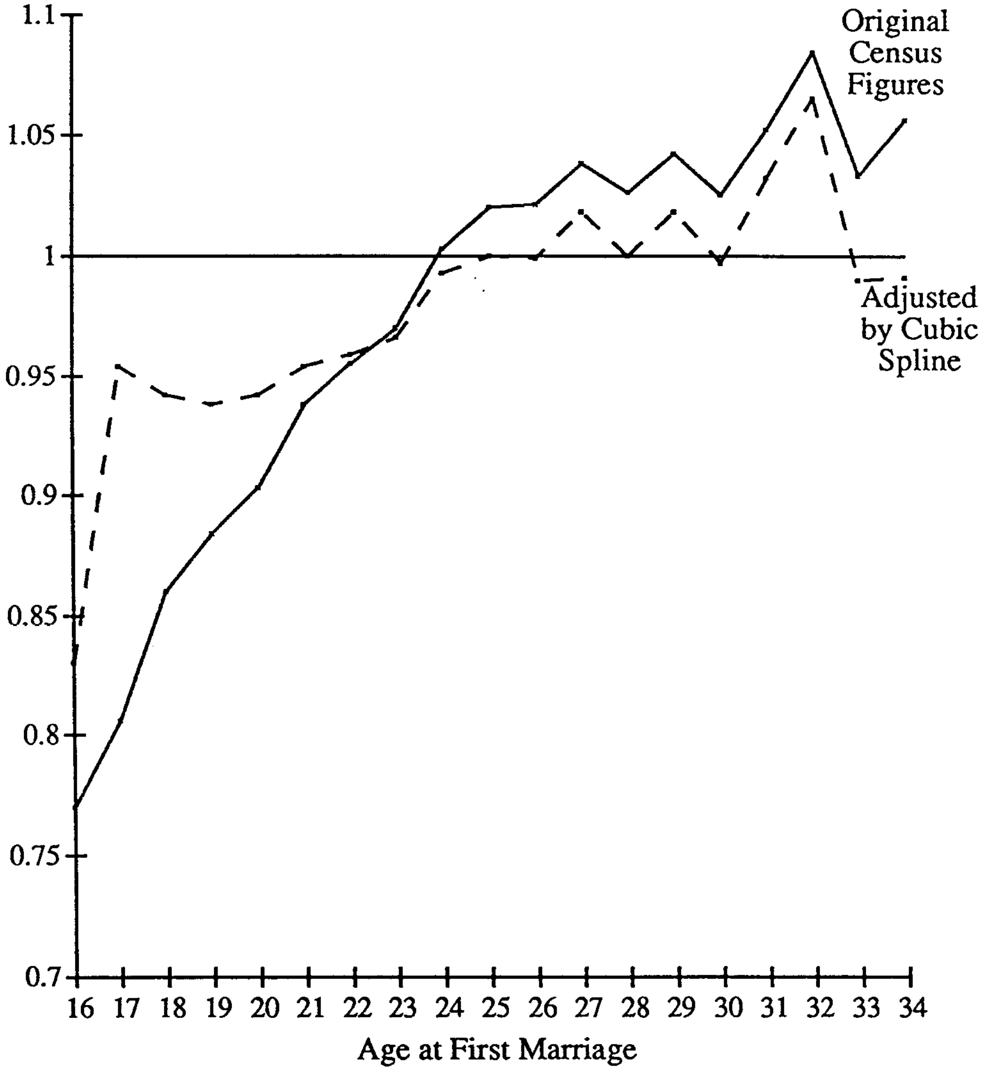


Figure 3.
Discrepancy in Estimated Proportion Marrying by Age.
Ratio of 1970 Estimate to 1960 Estimate.
Average for 1910-1924 White Male Cohorts.


Age at First Marriage
Original Census Figures
Adjusted by Cubic Spline