WORLD FERTILITY SURVEY

TECHNICAL BULLETINS

Maximum Likelihood Estimation of the Parameters of Coale's Model Nuptiality Schedule from Survey Data

GERMAN RODRIGUEZ
AND
JAMES TRUSSELL

WORLD FERTILITY SURVEY

TECHNICAL
BULLETINS

Maximum Likelihood Estimation of the Parameters of Coale's Model Nuptiality Schedule from Survey Data

GERMAN RODRIGUEZ
AND
JAMES TRUSSELL

CONTENTS

ACKNOWLEDGEMENTS

INTRODUCTION

In this technical bulletin we develop procedures for fitting Coale's model nuptiality schedule to World Fertility Survey data, using the method of maximum likelihood. There are several reasons why one may be interested in fitting a model to WFS nuptiality data.

Firstly, the model may be used as a tool for smoothing the data or as an aid in assessing the quality of data. For example, fitting the model to distributions of marital status by age such as those obtained from WFS household surveys leads to smooth estimates of the proportion ever married by single years of age and helps identify ages where reporting is deficient.

Secondly, the model permits a succinct description of the marriage process in terms of three simple parameters, namely, the proportion of women in a cohort who will eventually marry and the mean and standard deviation of age at marriage for those who marry. If the model fits the data then these three parameters effectively capture all the information in the observed marriage schedules. In other words, the model permits parsimonious description without loss of information.

Thirdly, the model permits extrapolation from the incomplete experience reported at a cross-sectional survey by cohorts of women who are still undergoing the marriage process. This is perhaps the most important application in the context of distributions of age at marriage such as those obtained from WFS individual surveys, which are truncated or censored at the interview. Fitting the model to these data permits estimation of the proportions who will eventually marry as well as the mean and standard deviation of age at marriage, even for cohorts where only half the women who will ever marry have done so by the date of the survey.

Fourthly, the model itself is of interest to students of nuptiality, as it describes a complex process in terms of relatively simple mechanisms which have a behavioural basis or interpretation. The development of estimation procedures for WFS data permits validation of the model on a much more extensive data base than has heretofore been possible.

The procedures herein developed have been designed to estimate the parameters of the model, including mean age at marriage, making full use of the information available whilst properly taking into account the truncated or censored nature of the data. As such they represent a more refined analytic tool than the *ad hoc* procedures used to handle truncation in the estimation of mean age at marriage in WFS first country reports.

Finally, an important feature of the maximum likelihood approach adopted here is that it leads not only to estimates of the parameters of the model, but also to large sample estimates of the standard errors of the estimates, and large sample tests of the goodness of fit of the model.

This bulletin is organized in eight sections following this introduction.

In Section 1 we describe Coale's model nuptiality schedule, introduce its standard density and cumulative distribution functions, propose a reparameterization of the model in terms of its mean and standard deviation, and relate the model to a gamma distribution.

In Section 2 we consider estimating the parameters of the model for a synthetic cohort using data on marital status by age, of the type collected in the WFS household schedule. The basic features of the maximum likelihood procedures are described and illustrated, including estimation, standard errors, goodness of fit and robustness.

In Section 3 we discuss estimation of two of the parameters of the model for real cohorts, using data on age at marriage from a sample of ever-married women, of the type collected in the WFS individual interview. In addition to extending the estimation and goodness of fit procedures to this type of situation we introduce a test for homogeneity of cohorts.

In Section 4 we consider estimating all three parameters of the model for a real cohort by combining individual data on age at marriage with household data on marital status by age. We

propose two alternative procedures termed two-stage estimation and full information estimation.

In Section 5 we describe procedures appropriate for cases where data on marital status and age at marriage (of those ever-married) are available for the same sample of women, as it is the case in WFS surveys where all women in the reproductive ages, irrespective of marital status, are eligible for the individual interview.

In Section 6 we turn our attention to estimation using ungrouped or continuous data from ever-married or all-women sample, and discuss both parametric and non-parametric estimation of the nuptiality schedule from a truncated or censored sample.

In Section 7 we show that the model nuptiality schedule can also adequately replicate observed first birth schedules. This application may be used as either a diagnostic device for smoothing data or as a means of inferring the schedule of entry into cohabitation.

In Section 8 we refer briefly to the numerical procedures used to calculate the estimates and make some remarks concerning the evaluation of the cumulative distribution function. Reference is made to a computer package specially suited to handle the different types of data available from the WFS.

Throughout the paper the recommended procedures are illustrated using data from the Colombian National Fertility Survey of 1976, conducted as part of the WFS. The data are used not only to illustrate the maximum likelihood procedures, but also to compare methods of estimation, assess the robustness of the estimates, and compare results using grouped and ungrouped data.

1. COALE'S MODEL NUPTIALITY SCHEDULE

1.1 The Age Pattern of Marriage

Coale (1971) has presented empirical evidence to the effect that the distribution of age at first marriage in a female cohort takes the same basic form in a wide variety of populations, differing only in the location and scale of age at marriage and the proportion of the cohort eventually marrying.

Figure 1.1, reproduced from Coale (1971), illustrates vividly the existence of this common pattern. Panel A shows proportions ever-married by age for five different populations, and depicts clearly differences in location, scale and proportion ultimately marrying. Panel B shows the same data adjusted to give a proportion eventually married equal to one, and plotted with age standardised for location and scale, and reveals a remarkable uniformity in the age pattern of marriage.

The same type of uniformity is noted in observed schedules of first marriage frequencies, as illustrated in Figure 1.2, also reproduced from Coale (1971). Panel A shows first marriage frequencies for two cohorts and two cross-sections, differing in location and scale. Panel B shows the same data adjusted for location and scale, and reveals a common structure.

To represent this underlying structure, a "standard" schedule was constructed by making minor adjustments to the schedule of first marriage frequencies recorded in Sweden from 1865 to 1869. The standard frequencies, as well as the corresponding proportions ever-married by age, were tabulated by Coale (1971) in intervals of one-tenth of a year.

The question naturally arose as to whether this underlying pattern could be represented by a mathematical function. Trial and error lead Coale (1971) to find a closed-form expression for the risk of first marriage. Later, however, Coale and McNeil (1972) found an analytic expression for the frequency of first marriages that fits the Swedish standard — and hence many observed nuptiality schedules — remarkably well. The mathematical model will be introduced below.

1.2 Analytic Formulation of the Model

At this point we must introduce some notation. Let $f(a)$ represent the frequency of first marriages at exact age a, so that a proportion $f(a)da$ of a cohort marries between exact ages a and $a+da$.

Our development of the model proceeds in three stages. The function $f(a)$ may be related to the distribution of age at first marriage by writing

$$f(a) = c\, g(a), \tag{1.1}$$

where c is the proportion of the cohort eventually marrying, and $g(a)$ is the probability density function of age at first marriage among those who marry, so that a proportion $g(a)da$ of those who eventually marry do so between exact ages a and $a+da$.

The function $g(a)$ may in turn be related to a standard schedule of age at first marriage, by writing

$$g(a) = \frac{1}{k}\, g_s\left(\frac{a-a_0}{k}\right), \tag{1.2}$$

where a_0 is a location parameter which may be interpreted as the age at which a consequential number of marriages first occur, k is a scale parameter which may be interpreted as the rate at which marriage occurs (relative to the standard), and $g_s(z)$ is the standard schedule derived from Swedish data by Coale (1971).

Finally, the function $g_s(z)$ was found by Coale and McNeil (1972) to be very well approximated by the following probability density function:

FIGURE 1.1: Proportions ever-married by age, selected countries.
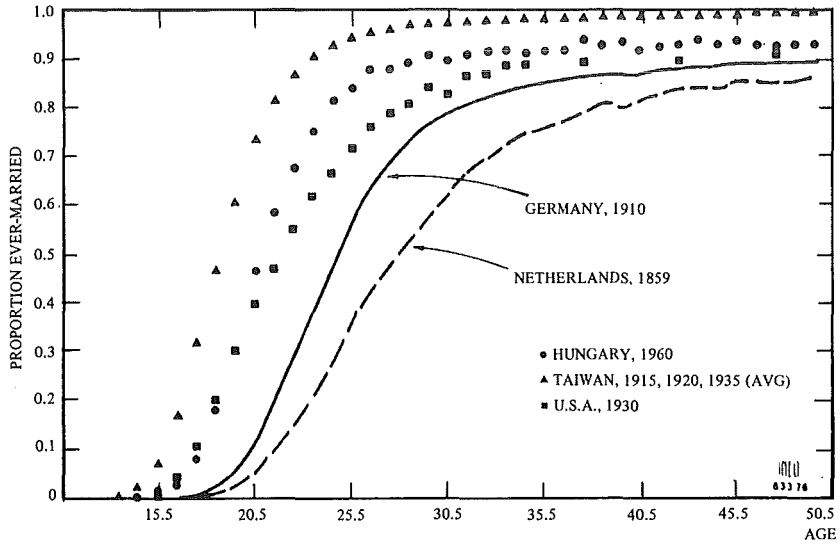
Panel A — Proportions ever-married, selected populations.
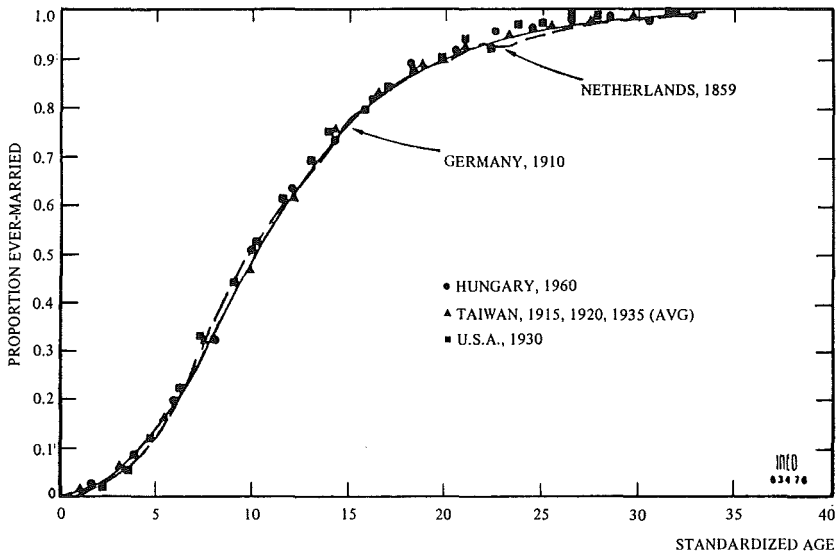
Panel B — Proportions ever-married, adjusted scale and origin selected populations.

Source: Coale (1971).

9

FIGURE 1.2: First marriage rates by age for selected countries.
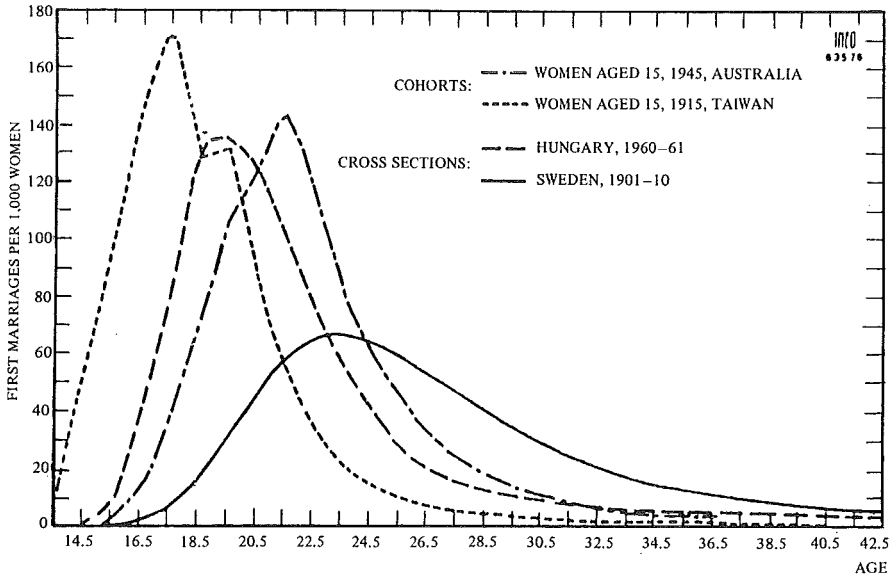
Panel A — First-marriage frequency (first marriages per thousand women) by single years of age, selected populations.
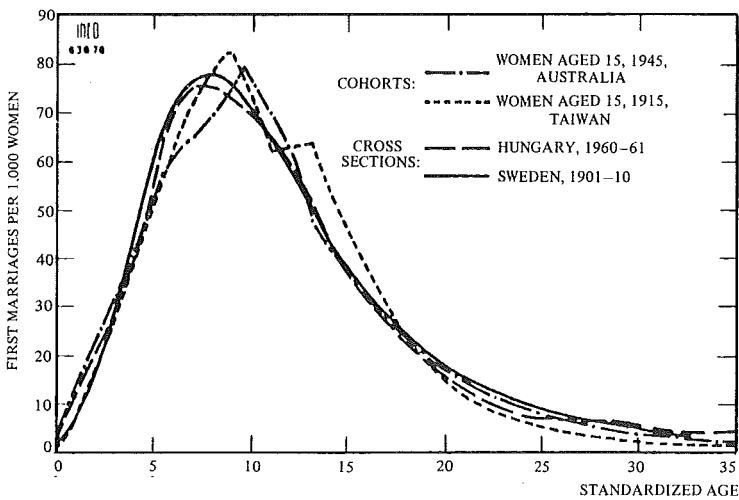
Panel B — First-marriage frequency, adjusted scale and origin, selected populations.

Source: Coale (1971).

$$g_S(z) = 0.1946 \, \text{Exp} \left\{ -0.174(z\text{-}6.06)\text{-Exp} \left[-0.288(z\text{-}6.06) \right] \right\} \qquad (1.3)$$

The function g(a) may now be written in full by substituting $(a\text{-}a_0)/k$ for z and dividing through by k at (1.3). Multiplication of the result by c gives an analytic expression for f(a). Thus, we have expressed f(a) in terms of a standard schedule (1.3) by using three parameters: ao, k and c.

A lucid account of developments leading to this analytic form for the standard schedule of first marriages may be found in Coale (1977), as well as the original and more technical paper by Coale and McNeil (1972).

The statistically inclined reader may be interested to know that the density at (1.3) represents the convolution of an infinite number of mean-corrected exponential random variables. This density, however, is in turn very closely approximated by the convolution of a normally distributed random variable and three exponential delays. Coale and McNeil (1972) have interpreted these components in Western cultures as representing the age of entry into the marriage market and the delays involved in finding a suitable partner, getting engaged, and getting married.

The conditional density given at (1.3) has mean and variance as follows:

$$E(z) = 11.36, \text{ and Var}(z) = 43.34. \qquad (1.4)$$

Changing variables from z to a=ao+kz gives, for any values of a_0 and k, the mean and variance of age at marriage (for those who marry) as

$$E(A) = a_0 + 11.36 \, k,$$
$$\text{and} \qquad\qquad\qquad\qquad (1.5)$$
$$\text{Var(A)} = 43.34 \, k^2.$$

It now remains only to define the proportion ever-married by exact age x among all women in a cohort as:

$$F(x) = \int_{-\infty}^{x} f(a)da. \qquad (1.6)$$

This function may be written as:

$$F(x) = c \, G(x), \qquad (1.7)$$

where G(x) is the cumulative distribution function of age at marriage for those who eventually marry,

$$G(x) = \int_{-\infty}^{x} g(a)da. \qquad (1.8)$$

This function may, in turn, be expressed in terms of the standard schedule by writing:

$$G(x) = G_S \left(\frac{x\text{-}a_0}{k} \right), \qquad (1.9)$$

where $G_S(z)$ is the standard cumulative distribution function of age at first marriage obtained integrating (1.3), that is:

$$G_S(z) = \int_{-\infty}^{z} g_S(t)dt. \qquad (1.10)$$

The question of evaluating this integral will be considered in Section 1.4 below.

1.3   A Standard with Mean 0 and Variance 1

The choice of a_0 and k as the location and scale parameters of the model is certainly valid, but somewhat arbitrary. One objection that may be raised is that these parameters are not easily interpretable, and thus do not provide a convenient basis for comparisons across cohorts or populations.

The location parameter a_0 is not the minimum age at marriage, but rather the age at which a "consequential" number of marriages first occurs. More precisely, the model implies that about one per cent of the women who will eventually marry have done so by age a_0, so that a_0 is close to the first percentile of the distribution.

The scale parameter k is literally the number of years in the standard schedule into which one year of marriage in the actual population may be packed, and therefore represents the rate of marriage relative to the Swedish standard. For example, in the standard about five per cent of the women who will eventually marry have done so by the end of the first age of marriage. If in an actual population k=2 it would mean that it takes two years for the same five per cent to marry, implying that the pace of marriage is slower than in the Swedish population of 1865-1869.

On the other hand, we have found that the statistic of greatest interest in fitting the model is usually the mean age at marriage, so that in actual practice one would translate a_0 and k into a mean and, say, a standard deviation, using (1.5). It thus seems more natural and convenient to reparameterize the model in terms of the mean and standard deviation rather than a_0 and k.

A new standard with mean 0 and variance 1 (analogous to the standard normal distribution), may be obtained from the existing standard (1.3) using (1.5) to find the values of ao and k that give the desired mean and variance. The required values are:

$$a_0 = -11.36/6.583 = -1.726,$$
$$\text{and} \tag{1.11}$$
$$k = 1/6.583 = 0.152.$$

Substituting $(a-a_0)/k$ for z and dividing through by k at (1.3), using these values of a_0 and k gives, as the new standard density function:

$$g_0(z) = 1.2813 \, \text{Exp}\{-1.145(z+0.805)-\text{Exp}[-1.896(z+0.805)]\} \tag{1.12}$$

The probability density function of age at first marriage, g(a), may be related to this new standard by writing

$$g(a) = \frac{1}{\sigma} \, g_0 \left(\frac{a-\mu}{\sigma} \right), \tag{1.13}$$

where μ is the mean age at marriage and σ is the standard deviation of age at marriage, among those who marry.

Similarly, the cumulative distribution function of age at first marriage G(x) may be written as

$$G(x) = G_0 \left(\frac{x-\mu}{\sigma} \right), \tag{1.14}$$

where G_0 is the new standard cumulative distribution function

$$G_0(z) = \int_{-\infty}^{z} g_0(t)dt. \tag{1.15}$$

We now consider the question of evaluating this integral.

1.4 Relationship of the Model to a Gamma Distribution

Unfortunately no closed form expression exists for the integrals given at (1.10) and (1.15) representing standard cumulative distributions of age at first marriage. These distributions, however, can be related quite easily to an incomplete gamma function, a result which greatly simplifies calculations, as simple algorithms exist for the calculation of the latter.

The density function used by Coale and McNeil (1972), may be written in general form as:

$$g(a) = \frac{\lambda}{\Gamma(a/\lambda)} \, \text{Exp}\{-a(a-\theta)-\text{Exp}[-\lambda(a-\theta)]\}, \tag{1.16}$$

12

where Γ denotes the gamma function and a, λ, θ are three parameters. The mean of this distribution is $\mu = \theta - \frac{1}{\lambda} \psi(a/\lambda)$, where $\psi = \Gamma' /\Gamma$ is the digamma function.

If we set $a=0.174$, $\lambda=0.288$ and $\theta=6.06$, with the constant $\lambda/\Gamma(a/\lambda)$ resulting 0.1946 and the mean $\mu=11.36$, we obtain the Swedish standard given at (1.3). Alternatively, setting $a=1.145$, $\lambda=1.896$ and $\theta=0.805$, with the resulting constant $\lambda/\Gamma(a/\lambda)$ equal to 1.2813, we obtain the new standard with mean 0 and variance 1 given at (1.12).

More generally, setting $a=1.145/\sigma$, $\lambda=1.896/\sigma$ and $\theta=\mu-0.805\sigma$, we obtain a distribution with mean μ and variance σ^2. In all these formulations the ratio a/λ is constant at 0.604 so that the model has only two parameters. (The question of whether the model may be generalised by allowing a/λ to be arbitrary may well deserve further research.)

The cumulative distribution function corresponding to (1.16) is given by the integral

$$G(x) = \int_{-\infty}^{x} g(a)da = \int_{-\infty}^{x} \frac{\lambda}{\Gamma(a/\lambda)} \mathrm{Exp}\{-a(a-\theta)-\mathrm{Exp}[-\lambda(a-\theta)]\}da. \tag{1.17}$$

Consider the change of variables

$$y = e^{-\lambda(a-\theta)}, \text{ so that } a = \theta - \frac{1}{\lambda} \log y. \tag{1.18}$$

Then

$$G(x) = \frac{1}{\Gamma(a/\lambda)} \int_{e^{-\lambda(x-\theta)}}^{\infty} y^{\frac{a}{\lambda}-1} e^{-y} dy, \tag{1.19}$$

which, recalling the definition of the gamma function, may be written as (Coale and McNeil, 1972, p.748)

$$G(x) = 1 - \frac{1}{\Gamma(a/\lambda)} \int_{0}^{e^{-\lambda(x-\theta)}} y^{\frac{a}{\lambda}-1} e^{-y} dy, \tag{1.20}$$

or more simply, as

$$G(x) = 1 - I[e^{-\lambda(x-\theta)}; \frac{a}{\lambda} -1], \tag{1.21}$$

where $I(w,p)$ denotes the incomplete gamma function

$$I(w,p) = \frac{1}{\Gamma(p+1)} \int_{0}^{w} y^{p}e^{-y}dy. \tag{1.22}$$

Thus, for any values of the parameters a, λ and θ (or μ and σ), the cumulative distribution function $G(a)$ may be evaluated in terms of an incomplete gamma function with parameter $\frac{a}{\lambda}=0.604$. In particular, the result may be used to evaluate the new standard cumulative distribution function as

$$G_0(z) = 1 - I [e^{-1.896(z+0.805)};-0.396]. \tag{1.23}$$

Approximations to the incomplete gamma function will be discussed in Section 8.2.

This formulation shows, incidentally, that age at marriage a (with parameters a, λ, θ or μ and σ) is distributed as $\theta = \frac{1}{\lambda} \log y$ where y has a standard gamma distribution with parameter $\frac{a}{\lambda}=0.604$, that is, age at marriage is distributed as a linear function of the logarithm of a standard gamma random variable.

A table of values of the new standard cumulative distribution function $G_0(z)$ is given in Appendix Table 5.

2. ESTIMATION FROM HOUSEHOLD DATA

2.1 The Data-Notation

The household schedule used in the WFS collects data on age at the interview and current marital status for all females between the ages of 15 and 49 or a similar age range. These data are usually tabulated by single years of age.

Table 2.1 shows such a set of data from the household interview of the Colombian National Fertility Survey of 1976, with a total of 12905 usual female residents between the ages of 15 and 49, of whom 7361 had been or were married legally or consensually.

We now consider fitting a model nuptiality schedule to this type of data, treating the different ages as representing a *synthetic cohort.*

The resulting parameter estimates will, of course, not apply to the experience of a real cohort unless nuptiality has been unchanging in the past. Our experience indicates, however, that the resulting fitted model may be used to smooth the data even in cases of changing nuptiality.

Let us introduce the following notation with reference to Table 2.1:

x = age at interview in completed years, ranging from x_0 to x_1, in our example 15 to 49 (Column 1)

m_x = number of ever-married women age x completed years at the interview (Column 2)

s_x = number of single (never married) women age x completed years at the interview (Column 3)

$n_x = m_x + s_x$ = total number of women age x completed years at the interview (Column 4).

In fitting the model we assume that age x completed years represents $x + \frac{1}{2}$ exact years.

2.2 Maximum Likelihood Estimation

We shall treat the number m_x married by age x completed years as having a binominal distribution with parameters n_x and Π_x — where Π_x denotes the probability of being ever-married by age x completed years — independently for each age.

The likelihood of the data is then a product binominal distribution. The logarithm of the likelihood function, except for a constant representing the binominal coefficients, is

$$\log L = \sum_{x=x_0}^{x_1} \{m_x \log \Pi_x + s_x \log (1-\Pi_x)\}. \tag{2.1}$$

The unrestricted maximum likelihood estimators (m.l.e.'s) of the parameters Π_x, obtained by maximising (2.1), are simply the proportions ever-married in the sample,

$$P_x = \frac{m_x.}{n_x} \tag{2.2}$$

These values are shown in Table 2.1 (Column 5) and present some obvious irregularities. Particularly noticeable are the low values at ages 35, 40 and 45, suggesting that either ever-married women are less likely to heap their ages, or that women who heap ages under-report marriage. One objective in fitting a model may be to smooth these proportions.

Under Coale's model nuptiality schedule the probability of being married by age x, assuming that women age x completed years are on the average $x + \frac{1}{2}$ exact years, is

$$\Pi_x = F(x+\frac{1}{2}), \tag{2.3}$$

14

TABLE 2.1: Observed and fitted proportions ever-married by age.
Colombia household survey (1976).

	Number of Women			Proportion Ever-married		
Age (1)	Ever Married (2)	Never Married (3)	Total (4)	Observed (5)	Fitted (6)	Difference (7)
x	m_x	s_x	n_x	P_x	$\hat{\Pi}_x$	$P_x - \hat{\Pi}_x$
15	16.	656.	672.	.024	.026	—.002
16	48.	662.	710.	.068	.063	.004
17	71.	584.	655.	.108	.121	—.013
18	120.	559.	679.	.177	.195	—.019
19	176.	398.	574.	.307	.278	.029
20	255.	379.	634.	.402	.361	.042
21	198.	270.	468.	.423	.439	—.016
22	267.	259.	526.	.508	.509	—.002
23	287.	197.	484.	.593	.571	.022
24	275.	185.	460.	.598	.623	—.025
25	340.	197.	537.	.633	.666	—.033
26	292.	124.	416.	.702	.702	—.001
27	274.	101.	375.	.731	.732	—.001
28	303.	103.	406.	.746	.756	—.010
29	242.	69.	311.	.778	.776	.002
30	332.	96.	428.	.776	.792	—.016
31	145.	34.	179.	.810	.804	.006
32	261.	54.	315.	.829	.815	.014
33	215.	29.	244.	.881	.823	.058
34	201.	32.	233.	.863	.830	.033
35	344.	84.	428. ·	.804	.835	—.032
36	255.	40.	295.	.864	.840	.025
37	211.	33.	244.	.865	.843	.022
38	262,	60.	322.	.814	.846	—.032
39	177.	31.	208.	.851	.848	.003
40	306.	75.	381.	.803	.850	—.047
41	119.	15.	134.	.888 ·	.852	.036
42	209.	28.	237.	.882	.853	.029
43	148.	17.	165.	.897	.854	.043
44	152.	18.	170.	.894	.855	.040
45	240.	56.	296.	.811	.855	—.044
46	148.	25.	173.	.855	.856	—.000
47	163.	21.	184.	.886	.856	.030
48	190.	35.	225.	.844	.856	—.012
49	119.	18.	137.	.869	.857	.012
TOTAL	7361.	5544.	12905.			

FIGURE 2.1: Observed and fitted proportions ever-married; household survey data.
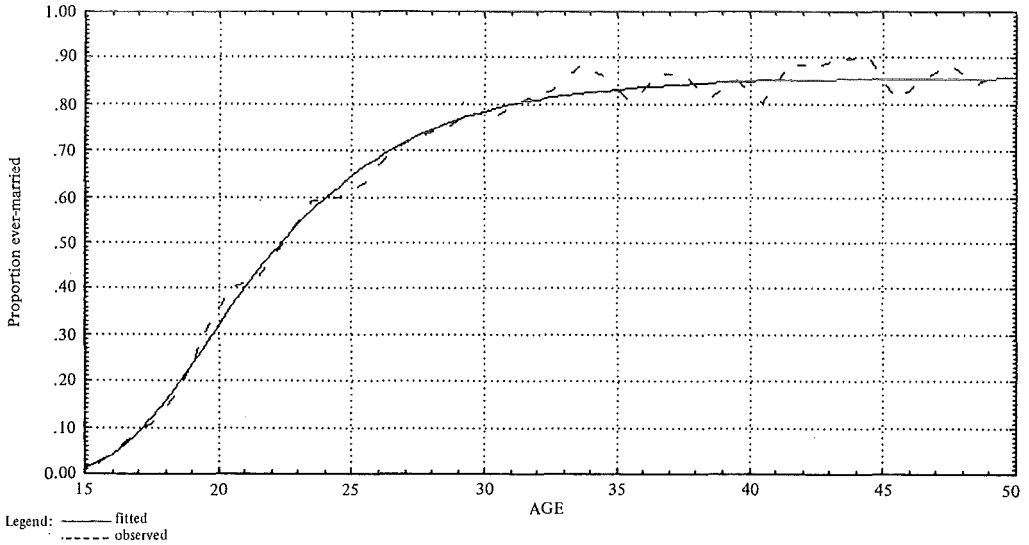
where F is the cumulative frequency function defined in Section 1.2 and depends on three parameters: μ, σ and c.

The log-likelihood (2.1) under the model (2.3) becomes

$$\text{Log } L = \sum_{x=x_0}^{x_1} \left\{ m_x \log [F(x+\tfrac{1}{2})] + s_x \log [1-F(x+\tfrac{1}{2})] \right\}. \tag{2.4}$$

This function depends on the data $\{m_x, s_x\}$ and the parameters μ, σ and c through F, and may be optimized numerically as noted in Section 8.

Maximum likelihood estimators (m.l.e.'s) of the parameters obtained using this method for the Colombian data are

$$\hat{\mu} = 22.44, \quad \hat{\sigma} = 5.28, \quad \text{and} \quad \hat{c} = 0.858 \tag{2.5}$$

The fitted mean age at marriage $\hat{\mu}$ is analogous to Hajnal's (1956) singulate mean age at marriage and may be interpreted in a similar way.

The fitted proportions ever-married by age are

$$\hat{\Pi}_x = \hat{F}(x+\tfrac{1}{2}), \tag{2.6}$$

where F denotes the cumulative frequency function F evaluated at the m.l.e.'s $\hat{\mu}$, $\hat{\sigma}$ and $\hat{c}$.

Table 2.1 (Column 6) shows fitted proportions ever-married for our example. Figure 2.1 compares the observed and fitted proportions.

2.3 Goodness of Fit of the Model

One advantage of the method of maximum likelihood is that it leads to a large sample test of the goodness of fit of the model, which we now present.

Under the product binomial model (2.1), the unrestricted m.l.e.'s of the parameters Π_x are the sample proportions P_x defined at (2.2), while the restricted m.l.e.'s of the same parameters under the model (2.3) are the fitted proportions $\hat{\Pi}_x$ defined at (2.6), leading to the likelihood ratio criterion

16

$$\chi_l^2 = 2 \sum_{x=x_0}^{x_1} \{m_x \ \log(P_x/\hat{\Pi}_x)+s_x \ \log[(1-P_x)/(1-\hat{\Pi}_x)]\}, \tag{2.7}$$

which is distributed in large samples as a chi-squared statistic with degrees of freedom

$$\nu = x_1 - x_0 - 2, \tag{2.8}$$

which is the number of ages or parameters in the unrestricted model (x_1-x_0+1) minus the number of parameters in the restricted model.

An alternative test criterion is the more familiar Pearson chi-squared statistic, which in this case is given by

$$\chi_p^2 = \sum_{x=x_0}^{x_1} n_x \ \frac{(P_x-\hat{\Pi}_x)^2}{\hat{\Pi}_x(1-\hat{\Pi}_x)}, \tag{2.9}$$

and is also distributed in large samples as a chi-squared variate with ν degrees of freedom. For our example we obtain

$$\begin{aligned}
\chi_1^2 &= 53.0, \ \text{P-value} = .011 \\
\chi_p^2 &= 52.7, \ \text{P-value} = .012 \\
\nu &= 32, \tag{2.10}
\end{aligned}$$

indicating a significant lack of fit.

Differences between observed and fitted values are given in Table 2.1 (Column 7), and show lack of fit particularly at ages ending in 0 or 5 at the extremes of the range, a possible consequence of heaping.

(As an alternative to raw residuals $P_x-\hat{\Pi}_x$ one may calculate standardized residuals

$$\sqrt{n_x}(P_x-\hat{\Pi}_x)/[\hat{\Pi}_x(1-\hat{\Pi}_x)]^{\frac{1}{2}} \tag{2.11}$$

where a value greater than 2 indicates a significant departure from the model.)

These results confirm what was visually obvious from a plot of the data in Figure 2.1; the observed proportions ever-married at the older ages are so erratic that no model could be expected to replicate them.

2.4 Standard Errors of the Estimates

A further advantage of the method of maximum likelihood is that it provides large sample approximations to the standard errors of the estimates.

Briefly, if $\hat{\theta}$ is a m.l.e. of a vector parameter $\underset{\sim}{\theta}$ then, under certain regularity conditions, the large sample distribution of $\hat{\underset{\sim}{\theta}}$ is normal with mean $\underset{\sim}{\theta}$ and variance-covariance matrix $\underset{\sim}{I}^{-1}(\theta)$ given by the inverse of the information matrix

$$\underset{\sim}{I}(\theta) = E\left[\frac{\partial \log L}{\partial \underset{\sim}{\theta}} \ \frac{\partial \log L}{\partial \underset{\sim}{\theta}'}\right] = -E\left[\frac{\partial^2 \log L}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'}\right] \tag{2.12}$$

The optimization procedures used here (see Section 8) provide numerical estimates of the matrix of second derivatives of the likelihood function, which in large samples should be reasonably close to the negative of its expected value, the information matrix.

For our example we obtain

$$s\hat{e}.\hat{\mu} = .146, \ s.\hat{e}.\hat{\sigma} = .162 \ \text{and} \ s.\hat{e}.\hat{c}. = .006 \tag{2.13}$$

17

These estimates are approximate and should therefore be interpreted with caution. The shape of the log-likelihood function is such that numerical estimates of the second derivatives in a neighbourhood of the optimum, and hence estimated standard errors, are unstable. Our experience indicates, however, that the numerical results provide at least a rough indication of the precision of the estimates.

A related question of interest is whether the estimated standard errors and the chi-squared statistics introduced earlier — which assume simple random sampling — are appropriate in stratified-clustered samples of the type used in the WFS.

Experience from the WFS indicates that design effects for nuptiality variables such as proportion ever-married and mean age at marriage are usually not far from unity, see Verma, Scott and O'Muircheartaigh (1980). Moreover, in later sections we shall be fitting the model to cohorts defined usually by five-year age groups, which are cross-classes and hence not likely to be seriously clustered. Under these circumstances we feel that treating the data as binomial should give a fairly good approximation to standard errors and chi-squared statistics.

2.5 Robustness of the Estimates

So far we have estimated the parameters of the model using all ages in the range 15 to 49, but clearly the procedure may be applied to any subset thereof. In theory four data points are required to estimate three parameters while reserving one degree of freedom for lack of fit, but in practice we would not recommend using less than 15 ages or data points.

Table 2.2 (lines 2 to 6) shows estimates of the parameters, as well as standard errors and the goodness of fit criterion, obtained by selecting progressively younger subsets of the age range. Note that the estimates of the parameters remain fairly stable, even when only ages 15 to 29 are used. One would expect this result if there had been no change in nuptiality in the recent past and if the data were of high quality. Note also that deleting the older ages increases the standard errors, as less data are used, but also improves the quality of fit, as the less reliable data points are ignored.

One of the difficulties posed by the poor quality of data for the older ages is that it makes estimation of c, the proportion eventually marrying, rather unreliable. In our example we

TABLE 2.2: Estimates of parameters of the model fitted to grouped marriage data from the Colombia household survey (1976).

	Ages	Estimates			Standard Errors			Goodness of Fit		
	(1) $x_0 x_1$	(2) $\hat{\mu}$	(3) $\hat{\sigma}$	(4) $\hat{c}$	(5) $s.e.\hat{\mu}$	(6) $s.e.\hat{\sigma}$	(7) $s.e.\hat{c}$	(8) χ_1^2	(9) ν	(10) p-value
	15–49	22.44	5.28	.858	.146	.162	.006	53.0	32	.011
	15–44	22.49	5.33	.861	.160	.174	.007	46.5	27	.011
	15–39	22.44	5.28	.858	.167	.179	.009	32.4	22	.071
	15–34	22.61	5.44	.872	.230	.234	.015	23.8	17	.126
	15–20	22.14	5.02	.830	.290	.272	.023	14.2	12	.286
	15–24	21.79	4.74	.794	.539	.452	.057	11.1	7	.135
Fix c	15–49	23.17	6.07	.90	.115	.145	—	102.9	33	.000
Fix c	15–39	23.10	6.00	.90	.112	.140	—	53.2	23	.000
Fix c	15–29	22.95	5.76	.90	.040	.050	—	21.2	13	.069
Fix c	15–24	22.71	5.46	.90	.141	.170	—	13.7	8	.089

have obtained values of c rather too low. An alternative is to set c at a fixed value and optimize the log-likelihood function (2.4) letting only μ and σ vary.

Table 2.2 (lines 7 to 9) shows estimates of μ and σ, as well as standard errors and goodness of fit tests obtained by fixing c at 0.90, which we believe to be a more plausible figure. The resulting estimates of mean age at marriage are quite stable, even when only ages 15 to 24 are used. Hence we have a strong indication that nuptiality patterns have not changed much in the recent past.

3. ESTIMATION FROM INDIVIDUAL DATA ON EVER-MARRIED WOMEN

3.1 The Data-Notation

The individual interview used in the WFS is usually applied to a sample of ever-married women between the ages of 15 and 49 or a similar age range, and collects information on age at marriage and age at interview. These data are frequently tabulated in single completed years of age.

Table 3.1 presents such a set of data for the cohort aged 25 to 29 in the individual interview of the Colombian National Fertility Survey. (Table 2 in the Appendix shows similar data for the cohorts aged 15 to 49.)

An important feature of this type of data for a sample of ever-married women, where each cohort is represented only by those who have married as of the interview, is that the distribution of age at marriage is *truncated* by age at the interview. This feature is reflected in Table 3.1 by the fact that there are no data below the main diagonal of the table.

From the point of view of estimation, truncation requires that we work with conditional probabilities of marriage — that is the probability of marrying at a certain age conditional on marrying by the current age of the cohort — rather than marriage frequencies. The use of such conditional probabilities underlies all developments in this section.

TABLE 3.1: Tabulation of age at marriage by age at interview for women aged 25–29 at the time of the survey, Colombia (1976).

Age at Marriage	Age at Interview x				
(1) a	(2) 25	(3) 26	(4) 27	(5) 28	(6) 29
11	0	1	1	1	1
12	2	4	0	8	2
13	4	4	4	6	3
14	8	5	8	8	4
15	14	10	7	13	8
16	14	12	9	16	12
17	8	10	15	13	7
18	15	13	11	16	12
19	17	19	9	10	16
20	13	18	9	12	9
21	12	8	12	15	11
22	1	11	12	6	10
23	10	8	4	7	5
24	8	6	11	4	3
25	(1)	7	6	3	4
26		(1)	1	4	5
27			(2)	2	4
28				(2)	5
29					(2)
Total ever-married	127	137	121	146	123
Ever-married by exact age x	126	136	119	144	121

In fitting Coale's model nuptiality schedule this circumstance implies that we will be able to estimate two of the parameters of the model, namely μ and σ, governing age at marriage, but not c, the proportion of the cohort ultimately marrying. The use of additional information to estimate c will, however, be considered in Section 4.

Let us introduce the following notation with reference to Table 3.1

x = age at interview in completed years, ranging from x_0 to x_1, (in our example 25 to 29)

a = age at marriage in completed years, ranging from a_0 to x (in our example 11 to x), for the cohort aged x

m_{ax} = number of women married at age a completed years and now aged x completed years

m_x = total number of ever-married women aged x completed years at the interview.

At this point we must note that truncation creates one further problem, namely the treatment of women marrying at their current age of m_{xx}. The difficulty is that the cohort aged x completed years at the interview has experienced a full year of exposure to marriage at each age a $<$ x completed years, but less than a year of exposure at age x itself.

One possibility is to assume that women aged x completed years at the interview are on the average $x+\frac{1}{2}$ exact years, treat women marrying at their current age as marrying between exact ages x and $x+\frac{1}{2}$, and work with probabilities of marriage conditional on marrying by exact age $x+\frac{1}{2}$.

A simple alternative, which avoids any bias introduced by the above assumption and simplifies some further developments, is to ignore women marrying at their current age. For the cohort aged x completed years at the interview we simply truncate the experience at exact age x and work with probabilities of marriage conditional on marrying by exact age x. For this purpose we redefine

$$m_x = \sum_{a=a_0}^{x-1} m_{ax} = \text{total number of women aged x completed years at the interview who had married by exact age x.}$$

In the following discussion we will adopt this simpler alternative. Although extensions to use all data will be obvious in most cases, the details are cumbersome and will not be given.

3.2 Maximum Likelihood Estimation

Let us consider fitting the model to a real cohort aged x_0 to x_1 completed years at the interview. This may be a single-year cohort such as women aged 25 or a group of cohorts such as women aged 25 to 29. In all cases, however, we work with the data in single-year form.

We shall treat the numbers $\{m_{ax}\}$ married at each age a $<$ x for the cohort aged x as having a multinomial distribution with parameters m_x and $\pi_{a|x}$, where

$\pi_{a|x}$ = Probability of marrying between exact ages a and a+1 conditional on marrying by exact age x,

Note that for each of the cohorts in the age group x_0 to x_1 we have introduced a different set of conditional probabilities.

The likelihood of the data for the cohorts x_0 to x_1 is then a product multinomial distribution. The logarithm of the likelihood is, except for a constant representing the multinomial coefficients,

$$\log L = \sum_{x=x_0}^{x_1} \sum_{a=a_0}^{x-1} m_{ax} \, \log \, (\pi_{a|x}). \tag{3.1}$$

The unrestricted maximum likelihood estimators of the conditional probabilities $\{\pi_{a|x}\}$, obtained maximising (3.1), are simply

21

$$P_{a|x} = \frac{m_{ax}}{m_x} , \tag{3.2}$$

the sample proportions of women married between exact ages a and a+1 among those married by exact age x.

Under Coale's model nuptiality schedule, the probability of marrying between exact ages a and a+1 conditional on marrying by exact age x is given by

$$\pi_{a|x} = \frac{G(a+1)\text{-}G(a)}{G(x)} , \tag{3.3}$$

where G denotes the cumulative distribution function of age at marriage with parameters μ and σ defined at (1.8).

Expression (3.3) is simply the ratio of the probability of marrying between exact ages a and a+1 conditional on ever-marrying, to the probability of marrying by exact age x conditional on ever-marrying.

Note that we have used the same cumulative distribution function G with parameters μ and σ for all single-year cohorts in the age-group x_0 to x_1; that is, we are fitting the same model schedule to all cohorts in the group.

The log-likelihood function (3.1) under the model (3.3) becomes

$$\log L = \sum_{x=x_0}^{x_1} \sum_{a=a_0}^{x-1} m_{ax} \left\{ \log[G(a+1)\text{-}G(a)] \text{-}\log[G(x)] \right\}. \tag{3.4}$$

The function (3.4) depends on the data $\{m_{ax}\}$ and on the parameters (μ,σ) through the cumulative distribution function G, and may be optimized numerically as noted in Section 8.

Estimates obtained using this procedure for the cohort aged 25 to 29 in the Colombian individual survey are

$$\hat{\mu} = 21.22 \text{ and } \hat{\sigma} = 5.98 \tag{3.5}$$

Note that although we have worked with conditional probabilities of marriage we have been able to estimate the mean and standard deviation of the complete distribution of age at marriage. This result is possible because both the truncated distribution (3.3) and the complete distribution G depend on the same parameters μ and σ.

It should be noted, however, that the estimates of the parameters μ and σ which fit the truncated experience of a cohort still going through the marriage process may not necessarily fit the complete experience of the same cohort once it finishes marrying, a subject which will be discussed in more detail in Section 5.5.

Approximate standard errors of the estimates, obtained from a numerical approximation to the information matrix, are

$$\hat{se}.\hat{\mu} = .362 \text{ and } \hat{se}.\hat{\sigma} = .303 \tag{3.6}$$

These estimates are relatively unstable, depending somewhat on the optimization procedure used, but they provide at least a rough indication of the precision of the estimates.

Estimates of the parameters and associated standard errors for six 5-year cohorts in the Colombian individual survey are given in Table 3.2 (Columns 1 to 5).

The results for the cohorts aged 20-24 to 35-39 indicate an increase in mean age at marriage of approximately one year over the past 10 to 15 years. For the youngest cohort the results are unreliable, as indicated by the large standard errors. For the cohorts 40-44 and 45-49 the relatively higher means may represent mis-statement of age at marriage due to recall errors.

TABLE 3.2: Estimates of the parameters of the model fitted to grouped marriage data from the Colombia individual survey (1976).

Cohort	Estimates		St. Errors		Goodness of Fit			Homogeneity		
(1) $x_0 - x_1$	(2) $\hat{\mu}$	(3) $\hat{\sigma}$	(4) $s.e.\hat{\mu}$	(5) $s.e.\hat{\sigma}$	(6) $\chi^2 1$	(7) ν	(8) p	(9) $\chi^2 1$	(10) ν	(11) p
20–24	21.51	5.94	.640	.479	59.6	48	.121	40.7	38	.351
25–29	21.22	5.98	.362	.303	79.1	73	.292	65.9	58	.222
30–34	20.62	5.00	.247	.212	120.9	98	.058	88.4	78	.197
35–39	20.43	5.38	.251	.217	141.0	127	.188	108.9	102	.302
40–44	21.21	5.74	.263	.226	122.1	145	.917	92.3	117	.939
45–49	21.69	6.12	.320	.266	163.4	172	.669	132.6	139	.638

TABLE 3.3: Proportions marrying at each age among women 25–29 married by age at interview, Colombia (1976).

Age at Marriage	Age at Interview x					Pooled	Fitted	Difference
	25	26	27	28	29			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
a			$p_{a|x}$			$\bar{\pi}_{a|x}$	$\hat{\pi}_{a|x}$	$\bar{\pi}_{a|x} - \hat{\pi}_{a|x}$
11	.000	.007	.008	.007	.008	.006	.006	.000
12	.016	.029	.000	.056	.017	.023	.016	.007
13	.032	.029	.034	.042	.025	.030	.034	−.003
14	.063	.037	.067	.056	.033	.048	.055	−.007
15	.111	.074	.059	.090	.066	.075	.074	.001
16	.111	.088	.076	.111	.099	.091	.088	.003
17	.063	.074	.126	.090	.058	.077	.095	−.018
18	.119	.096	.092	.111	.099	.097	.095	.002
19	.135	.140	.076	.069	.132	.103	.090	.012
20	.103	.132	.076	.083	.074	.088	.083	.006
21	.095	.059	.101	.104	.091	.084	.074	.010
22	.008	.081	.101	.042	.083	.058	.064	−.006
23	.079	.059	.034	.049	.041	.049	.055	−.006
24	.063	.044	.092	.028	.025	.046	.047	−.001
25		.051	.050	.021	.033	.037	.039	−.003
26			.008	.028	.041	.025	.033	−.008
27				.014	.033	.022	.028	−.006
28					.041	.041	.023	.018
Number of cases	126.	136.	119.	144.	121.			

Fitted proportions married by exact age x among women who will eventually marry:

Cohort	25	26	27	28	29
$\hat{G}(x)$	.789	.825	.855	.880	.900

3.3 Goodness of Fit of the Model

The unrestricted m.l.e.'s of the conditional probabilities $\{\pi_{a|x}\}$ under the product multinomial model (3.1) are the *observed* proportions of women married between exact ages a and a+1 among those married by exact age x, $\{p_{a|x}\}$ defined at (3.2).

The restricted m.l.e.'s of the same conditional probabilities, under the restrictions (3.3) imposed by the model nuptiality schedule, are given by

$$\hat{\pi}_{a|x} = \frac{\hat{G}(a+1)-\hat{G}(a)}{\hat{G}(x)} , \tag{3.7}$$

where $\hat{G}$ denotes the cumulative distribution function G evaluated at the m.l.e.'s $(\hat{\mu}, \hat{\sigma})$.

We shall refer to the $\{\hat{\pi}_{a|x}\}$ as the *fitted* proportions married between exact ages a and a+1 among those married by exact age x.

Table 3.3 shows observed proportions for the cohorts 25 to 29 (Columns 2-6), and fitted proportions corresponding to the cohort aged 29 (Column 8). Fitted proportions for the other cohorts may be calculated using the fitted values $\hat{G}(x)$ given at the bottom of the table, and the following relation, which follows from (3.7).

$$\hat{\pi}_{a|x-1} = \hat{\pi}_{a|x} \frac{\hat{G}(x)}{\hat{G}(x-1)} \tag{3.8}$$

The likelihood ratio and Pearson chi-squared statistics for testing the goodness of fit of the model are given by

$$\chi_1^2 = 2 \sum_{x=x_0}^{x_1} \sum_{a=a_0}^{x-1} m_{ax} \log (p_{a|x}/\hat{\pi}_{a|x}) \tag{3.9}$$

and

$$\chi_p^2 = \sum_{x=x_0}^{x_1} \sum_{a=a_0}^{x-1} m_x (p_{a|x}-\hat{\pi}_{a|x})^2/\hat{\pi}_{a|x}, \tag{3.10}$$

and are distributed in large samples as chi-squared statistics with degrees of freedom ν given by

$$\nu = \sum_{x=x_0}^{x_1} (x-1-a_0)-2 \tag{3.11}$$

which is the total number of independent cells, $x-1-a_0$ for each cohort aged x, minus the number of parameters estimated; note that the last cell in each cohort contains truncated data which were ignored.

(If there is an age a_1 such that no one in the cohorts x_0 to x_1 has married after that age (i.e. $m_{ax}=0$ for $a>a_1$) we ignore such cells in calculating the chi-squared statistics and correct the degrees of freedom accordingly. Other cells with zero entries ($a_0 \leqslant a \leqslant a_1$) are, however, included in the calculations.)

For the cohort 25 to 29 we have,

$$\begin{array}{lll} \chi_1^2 = 79.1 & , \text{p-value} = .292 \\ \chi_p^2 = 74.1 & , \text{p-value} = .441 & (3.12) \\ \nu = 73 \end{array}$$

indicating a fairly good fit to the data.

FIGURE 3.1: Adjusted observed and fitted proportions ever-married by each age among those who will ever marry in the cohort aged 25–29; individual survey data.
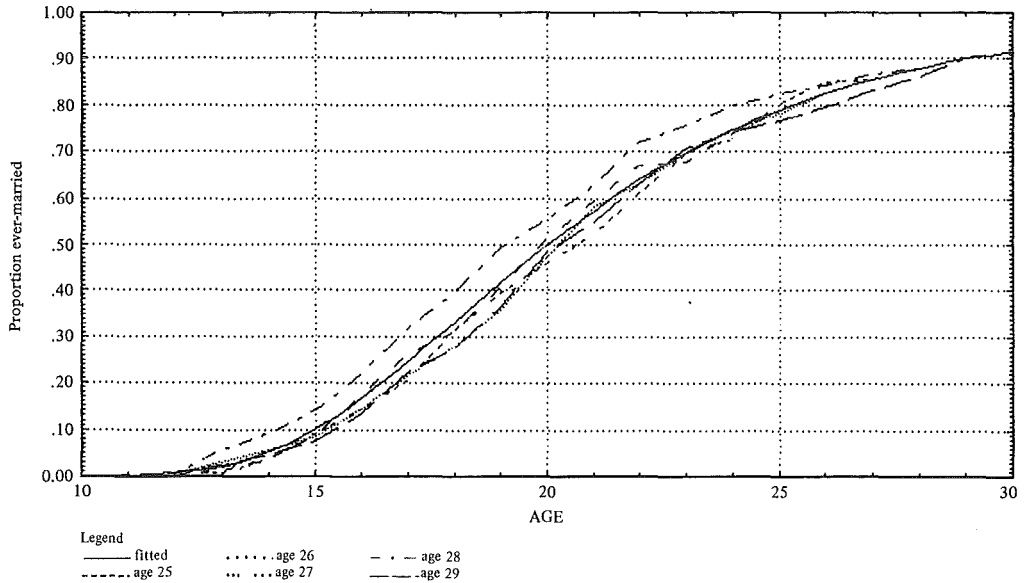
FIGURE 3.2: Adjusted observed and fitted proportions marrying at each age among those who will ever marry for the cohort aged 25–29; individual survey data.
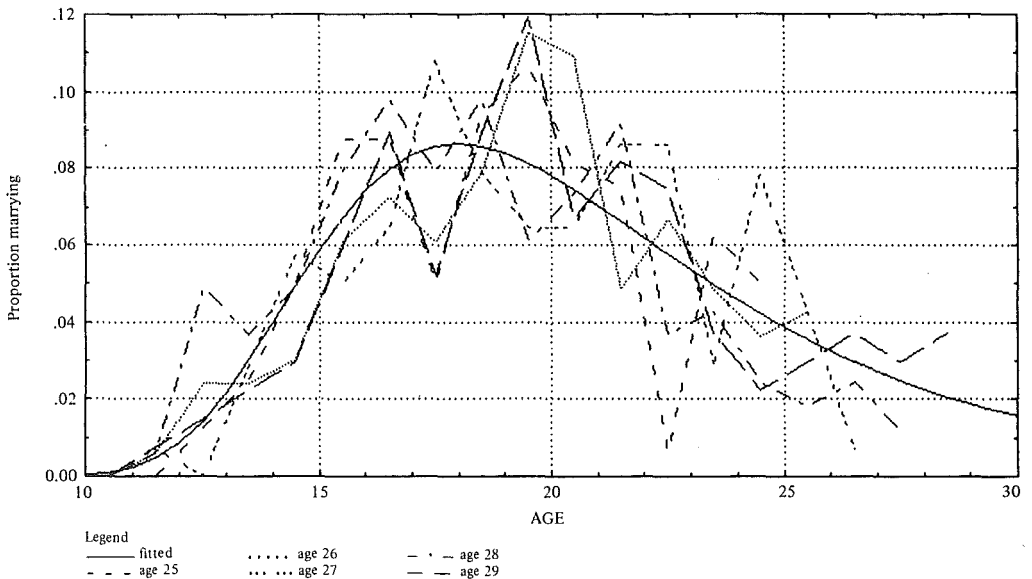
Values of the likelihood ratio chi-squared statistic, as well as its degrees of freedom and associated p-value, are shown in Table 3.2 (Columns 6-8) for the six 5-year cohorts in the individual interview of the Colombian survey. In general the model fits the data fairly well.

A visual impression of the goodness of fit of the model to each individual cohort x may be obtained by plotting the observed and fitted proportions marrying at each age among all women married by exact age x. Alternatively, one may accumulate these data and plot observed and fitted proportions married up to each age, among all women married by exact age x. In either case, a separate plot is required for each individual cohort as the conditioning age varies.

Another type of plot, which has certain advantages, may be obtained by calculating *adjusted* sample proportions married up to each age among all women who will eventually marry

$$P_{ax} = \hat{G}(x) \sum_{a=a_0}^{a-1} p_{a|x} , \quad a_0 \leqslant a \leqslant x \tag{3.13}$$

and plotting these together with $\hat{G}(a)$, the fitted cumulative distribution function. Note that an uncumulated adjusted sample proportion marrying at each age would be given by $G(x)p_{a|x}$; this uncumulated schedule is ordinarily more irregular than the cumulated version and thereby reveals distortions more readily.

Figures 3.1 and 3.2 show both types of plots for the cohorts aged 25 to 29 in the Colombian survey. One advantage of this type of plot is that all single-year cohorts in the age group x_0 to x_1 may be displayed on the same graph.

Note, however, that the adjusted values defined at (3.13) are a mixture of observed and fitted proportions, and in particular must necessarily agree with the fitted distribution at exact age x, as is visually evident in Figure 3.1.

3.4   Homogeneity of Cohorts

As noted earlier, the model may be fitted to a single-year cohort, such as women aged 25, or to a group of cohorts, such as women aged 25 to 29, by assuming that they have all followed the same nuptiality schedule.

In the latter case lack of fit of the model, as indicated by the tests introduced in the previous section, may be due to the fact that the different single-year cohorts in the group have not followed the same nuptiality schedule, or to genuine lack of fit of the model to their common schedule.

In order to distinguish these cases we now introduce a test for homogeneity of cohorts, by fitting a model where all single-year cohorts in the age group x_0 to x_1 are assumed to follow the same schedule which is otherwise unrestricted.

To do this we consider the product multinomial model (3.1) with parameters m_x and $\{\pi_{a|x}\}$. Recall that $\pi_{a|x}$ is the probability of marrying between exact ages a and a+1 conditional on marrying by exact age x, and that we introduced a different set of conditional probabilities for each cohort.

We now write all sets of conditional probabilities in terms of a common set $\{\pi_{a|x_1}\}$, which for convenience will be taken to refer to the older cohort.

$$\pi_{a|x} = \frac{\pi_{a|x_1}}{\sum_{a=a_0}^{x-1} \pi_{a|x_1}} , \quad x < x_1 \tag{3.14}$$

26

Thus we have written $\pi_{a|x}$ as the ratio of the probability of marrying between exact ages a and a+1 conditional on marrying by exact age $x_1 > x$, to the probability of marrying by exact age x conditional on marrying by exact age x_1.

The likelihood of the data under the set of restrictions (3.14) becomes

$$\log L = \sum_{x=x_0}^{x_1} \sum_{a=a_0}^{x-1} m_{ax}[\log(\pi_{a|x_1}) - \log \sum_{a=a_0}^{x-1} \pi_{a|x_1})]. \tag{3.15}$$

Asano (1965) has derived m.l.e.'s for multinomial distributions supplemented by incomplete sets of observations. A direct extension of his work to suit the truncated nature of our data shows that the estimates that maximise (3.15) may be calculated recursively as follows

$$\bar{\pi}_{a|x_1} = \begin{cases} \dfrac{m_{ax_1}}{m_{x_1}} & , a = x_1 - 1 \\[2em] \dfrac{m_{ax_1-1} + m_{ax_1}}{m_{x_1-1} + (m_{x_1} - m_{x_1-1x_1})}(1-\bar{\pi}_{x_1-1|x_1}) & , a = x_1 - 2 \\[2em] \dfrac{\sum\limits_{x=\max(x_0,a+1)}^{x_1} m_{ax}}{\sum\limits_{x=\max(x_0,a+1)}^{x_1} \sum\limits_{a=a_0}^{a} m_{ax}}(1-\sum\limits_{a=a+1}^{x_1-1}\bar{\pi}_{a|x_1}) & , a = a_0,...,x_1 - 3 \end{cases} \tag{3.16}$$

Thus, we first calculate $\bar{\pi}_{x_1-1|x_1}$, use this estimate to calculate $\bar{\pi}_{x_1-2|x_1}$, and carry on calculating successively $\bar{\pi}_{x_1-3|x_1}$ down to $\bar{\pi}_{a_0|x_1}$.

The restricted m.l.e.'s of the conditional probabilities $\{\bar{\pi}_{a|x}\}$ applying to the cohort aged x, under the set of restrictions (3.14), are given by

$$\bar{\pi}_{a|x} = \dfrac{\bar{\pi}_{a|x_1}}{\sum\limits_{a=a_0}^{x-1} \bar{\pi}_{a|x_1}}, \quad x < x_1 \tag{3.17}$$

(We use the notation π bar to distinguish these estimates from those obtained under Coale's model, which we denoted π hat.)

We shall refer to the $\bar{\pi}_{a|x}$ as the *pooled* estimates of the conditional probabilities of marrying between exact ages a and a+1 given marriage by exact age x. Pooled estimates pertaining to the cohort aged 29 in the Colombian survey are shown in Table 3.3 (Column 7). Pooled estimates for younger cohorts may be calculated using (3.17).

The unrestricted estimates of the same conditional probabilities are, of course, the sample proportions $\{p_{a|x}\}$ defined at (3.2).

The likelihood ratio and Pearson chi-squared statistics for testing the homogeneity of the cohorts aged x_0 to x_1 are given by

$$\chi_1^2 = 2\sum_{x=x_0}^{x_1} \sum_{a=a_0}^{x-1} \max \log (p_{a|x}/\overline{\pi}_{a|x}), \tag{3.18}$$

and

$$\chi_p^2 = \sum_{x=x_0}^{x_1} \sum_{a=a_0}^{x-1} m_x (p_{a|x}-\overline{\pi}_{a|x})^2/\overline{\pi}_{a|x}, \tag{3.19}$$

and are distributed in large samples as chi-squares with degrees of freedom ν given by

$$\nu = \sum_{x=x_0}^{x_1-1} (x-1-a_0) \tag{3.20}$$

which is the number of independent cells in the data, $(x-1-a_0)$ for each cohort aged x, minus the number of independent parameters estimated, (x_1-1-a_0).

(If there is an age $a_1 < x_1-1$ such that nobody in the cohorts aged x_0 to x_1 has married after age a_1, we substitute $x-1$ by min $(a_1,x-1)$ in (3.18)-(3.20), thus avoiding division by zero and correcting the number of degrees of freedom.)

For the cohorts aged 25 to 29 in the Colombian Survey we have

$$\chi_1^2 = 65.9 \qquad \text{p-value} = .222$$
$$\chi_p^2 = 60.1 \qquad \text{p-value} = .400$$
$$\nu = 58 \tag{3.21}$$

indicating that the cohorts may be considered to have followed the same nuptiality pattern (a hardly surprising result, since the test in the previous section had indicated that the same model schedule did fit these five cohorts well).

The likelihood ratio statistics for homogeneity of each of the six 5-year cohorts in the Colombian sample, as well as the corresponding degrees of freedom and associated p-values, are shown in Table 3.2 (Columns 9-11). All 5-year cohorts appear to be homogeneous, a fact consistent with the general impression that nuptiality has not been changing very much in Colombia.

In countries where age at marriage has been changing rapidly, however, one may find that 5-year cohorts are not homogeneous. In such cases a different model schedule should be fitted to each single year cohort in a heterogeneous group.

It is also possible that a χ^2 test will reveal that cohorts are not homogeneous even where it can be confidently assumed that nuptiality has not been changing; this situation is likely to arise when the quality of data is poor. In particular mis-statement of age can lead to the appearance of non-homogeneity and of a poor fit of the model to the data. In this case, the model can best be viewed as a diagnostic and smoothing device (Trussell, 1980).

Note that we have fitted two models to the same data, namely the model schedule defined by the restrictions (3.3) and the more general homogeneous schedule defined by the restrictions (3.14), and that these models are hierarchical, that is (3.3) is a subset of (3.14).

This nesting property permits us to compare the two models by simple subtraction of the chi-squared statistics and the corresponding degrees of freedom for each model. In the case of the likelihood ratio χ^2 the resulting statistic is the same that would be obtained directly from the observed and pooled proportions, namely

$$\chi_1^2 = 2\sum_{x=x_0}^{x_1} \sum_{a=a_0}^{x-1} \max \log (\overline{\pi}_{a|x}/\hat{\pi}_{a|x}). \tag{3.22}$$

FIGURE 3.3: Adjusted pooled and fitted proportion ever-married by each age among those who will ever-marry for the cohort aged 25−29; individual survey data.
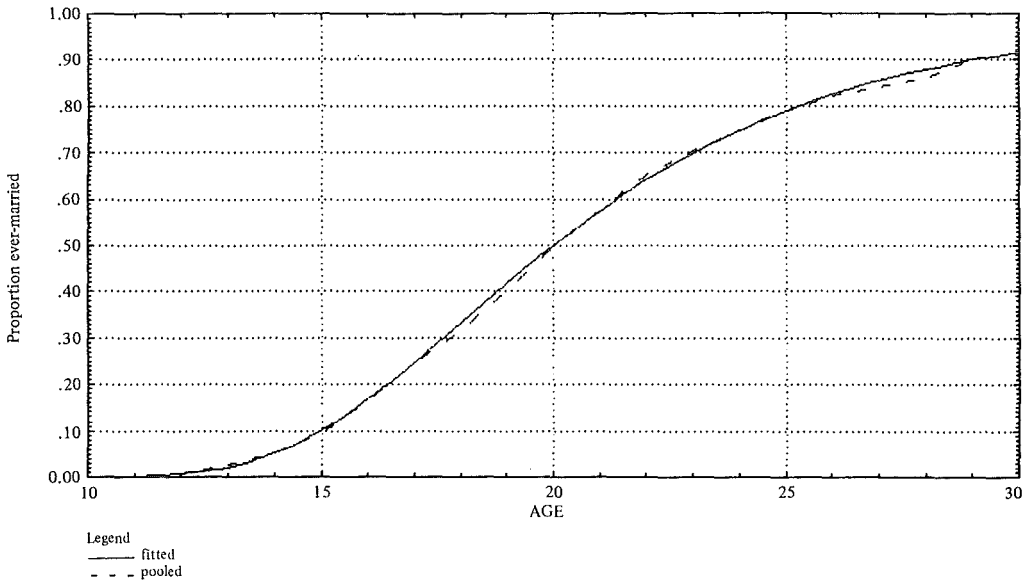
FIGURE 3.4: Adjusted pooled and fitted proportions marrying at each age among those who will ever-marry for the cohort aged 25−29; individual survey data.
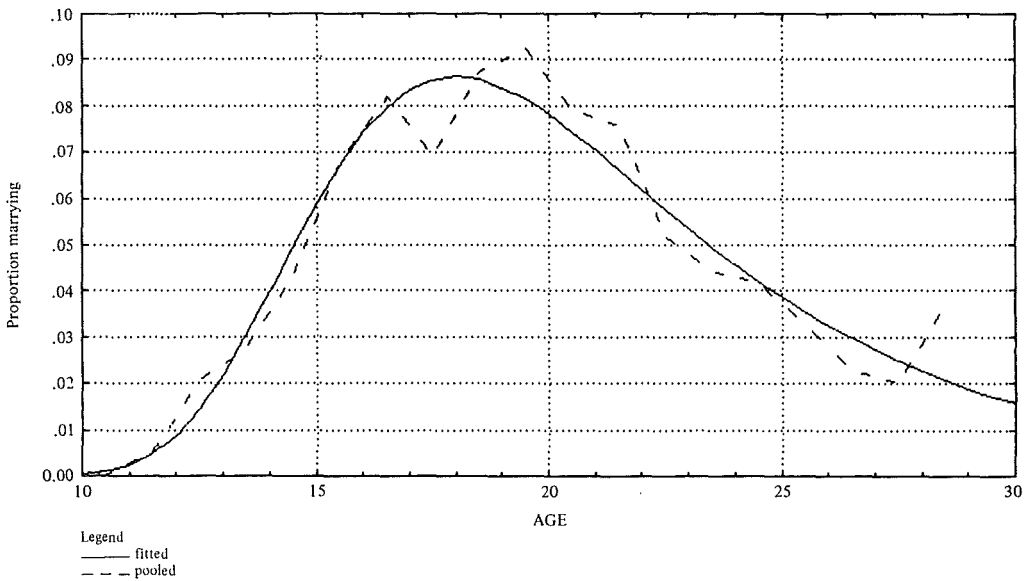
For the cohorts aged 25 to 29 in the Colombian survey we have

$$\chi_1^2 = 13.2 \qquad\qquad \text{p-value} = .588$$
$$\chi_p^2 = 14.1 \qquad\qquad \text{p-value} = .522$$
$$\nu = 15 \tag{3.23}$$

indicating that the model schedule agrees fairly well with the pooled estimates.

A visual impression of the goodness of fit may be obtained by plotting the pooled and fitted proportions, or by calculating the *adjusted* pooled values

$$\bar{\Pi}_a = \hat{G}(x_1) \sum_{a=a_0}^{a-1} \bar{\pi}_{a|x_1} \tag{3.24}$$

and plotting these together with the fitted cumulative distribution function $\hat{G}(a)$, as discussed in Section 3.4. An example of such a plot is given in Figure 3.3, and the uncumulated version is displayed in Figure 3.4.

4. ESTIMATION FROM INDIVIDUAL AND HOUSEHOLD DATA

4.1 The Data

We now consider estimating all three parameters of the model schedule by combining household data on marital status and age at interview for all women, with individual data on age at marriage and age at interview for ever-married women.

The basic data are as given in table 2.1 for the household interview and Appendix Table 2 for the individual interview in the Colombian National Fertility Survey. In this case the individual interview was conducted in a sub-sample of the household survey.

The notation to be used has already been introduced in Sections 2.1 and 3.1. It will only be necessary to distinguish household and individual data by adding a prime to identify the latter. Thus, m_x represents the number of ever-married women aged x completed years in the household survey, while m'_x represents the number of ever-married women aged x completed years in the individual survey.

We now discuss two methods of combining these data, which we term "two-stage estimation" and "full information estimation". These methods differ in the extent to which they use household data.

4.2 Two-stage Estimation

Suppose that at stage 1 the parameters μ and σ have been estimated using individual data by the procedures described in Section 3.

For a real cohort aged x_0 to x_1 completed years at the interview, let $\hat{\mu}$ and $\hat{\sigma}$ denote the maximum likelihood estimators of the parameters, and let $\hat{G}$ denote the cumulative distribution function evaluated at the m.l.e.'s.

At stage 2 we consider the likelihood of the household data given at (2.1). Recall that we treat m_x as having a binomial distribution with parameters n_x and Π_x. Under the model

$$\Pi_x = F(x+\tfrac{1}{2}) = c\, G(x+\tfrac{1}{2}). \tag{4.1}$$

We now treat G as *known* by substituting $\hat{G}$. This reduces the log-likelihood function to

$$\log L = \sum_{x=x_0}^{x_1} \left\{ m_x \log[c\hat{G}(x+\tfrac{1}{2})] + (n_x-m_x)\log[1-c\hat{G}(x+\tfrac{1}{2})] \right\}. \tag{4.2}$$

Differentiating with respect to c we obtain

$$\frac{\partial \log L}{\partial c} = \sum_{x=x_0}^{x_1} \left\{ \frac{m_x}{c} - \frac{(n_x-m_x)\hat{G}(x+\tfrac{1}{2})}{1-c\hat{G}(x+\tfrac{1}{2})} \right\} \tag{4.3}$$

In the case of a single-year cohort ($x_0=x_1$), setting this derivative to zero leads to the maximum likelihood estimator

$$\hat{c} = \frac{m_x/n_x}{\hat{G}(x+\tfrac{1}{2})} = \frac{P_x}{\hat{G}(x+\tfrac{1}{2})} \tag{4.4}$$

Thus, the estimate of the proportion c who will ultimately marry for the cohort aged x is simple the ratio of P_x, the proportion ever-married at exact age $x+\tfrac{1}{2}$ among all women, estimated from household data; to $\hat{G}(x+\tfrac{1}{2})$, the proportion ever-married at exact age $x+\tfrac{1}{2}$ among women who will eventually marry, estimated from individual data.

TABLE 4.1: Estimates of c obtained by treating μ and σ as known.

age x (1)	$p_x = m_x/n_x$ (2)	$\hat{G}(x+\frac{1}{2})$ (3)	$\hat{c}_x$ (4)
25	.633	.808	.784
26	.702	.841	.835
27	.731	.868	.842
28	.746	.891	.838
29	.778	.909	.856

Table 4.1 shows details of the application of this procedure to each single-year cohort in the age-group 25 to 29 in the Colombian survey, including values of P_x obtained from the house-hold data in 2.1, values of G evaluated at the m.l.e.'s $\hat{\mu}=21.22$ and $\hat{\sigma}=5.98$ obtained from individual data in Section 3.2, and the corresponding ratios or estimates of c.

In the case of a group of cohorts $(x_0 < x_1)$, setting the derivative (4.3) of the log-likelihood function to zero does not lead to an analytic expression for the m.l.e. of c. It is possible, how-ever, to derive a recursive relationship for this estimator.

Let $\hat{c}_x$ denote the estimate of c obtained from a single-year cohort age x by applying (4.4). Since m_x has a binomial distribution with parameters n_x and Π_x given at (4.1), and $G(x+\frac{1}{2})$ is assumed known, we have

$$E(\hat{c}_x) = c, \qquad (4.5)$$

and

$$\text{Var}(\hat{c}_x) = \frac{c[1 - c\hat{G}(x+\frac{1}{2})]}{n_x \hat{G}(x+\frac{1}{2})}. \qquad (4.6)$$

Consider now combining the different estimates of c_x by calculating a weighted average

$$\hat{c} = \frac{\sum_{x=x_0}^{x_1} w_x c_x}{\sum w_x} \qquad (4.7)$$

with weights equal to the reciprocals of the variance of $\hat{c}_x$,

$$w_x = \frac{n_x \hat{G}(x+\frac{1}{2})}{c(1 - c\hat{G}(x+\frac{1}{2}))} \qquad (4.8)$$

where in practice c must be replaced by its estimate $\hat{c}$.

The resulting estimate $\hat{c}$ is a minimum variance unbiased estimator of c, and hence a maximum likelihood estimator of c under the binomial model.

Recalling the definition of $\hat{c}_x$ given at (4.4), the weighted average (4.7) with weights (4.8) becomes

$$\hat{c} = \frac{\sum_{x=x_0}^{x_1} \dfrac{m_x}{1 - \hat{c}\hat{G}x}}{\sum_{x=x_0}^{x_1} \dfrac{n_x \hat{G}x}{1 - \hat{c}\hat{G}x}} \qquad (4.9)$$

32

TABLE 4.2: Two-stage estimates of c obtained by treating μ and σ as known.

age x_0 to x_1 (1)	$\hat{c}$ (2)	$\hat{s.e.\hat{c}}$ (3)	(no. of iterations)* (4) (5)	
20–24	.785	.015	1	(2)
25–29	.830	.012	1	(2)
30–34	.854	.010	1	(2)
35–39	.845	.010	1	(2)
40–44	.866	.011	1	(2)
45–49	.851	.011	1	(2)

*Number of iterations needed until there was no change in the third decimal place, initial value = average $\hat{c}$ for the 5 individual ages. The number of iterations when the initial value was 1.0 is given in parentheses.

where $\hat{G}_x$ is shorthand for $\hat{G}(x+\frac{1}{2})$. The expression (4.9) has $\hat{c}$ on both sides of the equation, but may be used to obtain the estimate iteratively. Starting with a value of c=1 we have found that only 2 or 3 iterations using (4.9) are needed.

Note from (4.7) that since $w_x = 1/var(\hat{c}_x)$, the variance of the m.l.e. $\hat{c}$ is simply

$$Var(\hat{c}) = 1/\sum_{x=x_0}^{x_1} \frac{n_x \hat{G}_x}{c(1-c\hat{G}_x)} \tag{4.10}$$

and may be estimated by substituting $\hat{c}$ for c in (4.10).

For the cohorts aged 25 to 29 in the Colombian survey we have

$$\hat{c} = .830 \text{ and } \hat{s.e.\hat{c}} = .012 \tag{4.11}$$

In many practical applications we have found that a simple unweighted average of the $\hat{c}_x$ gives a reasonable estimate of c, and that a single iteration using (4.9) with the unweighted average as the starting value is sufficient to obtain the m.l.e. Results for the other cohorts are given in Table 4.2.

4.3 Full Information Estimation

In the case of a single-year cohort the household data contain no information about the shape of the nuptiality schedule, but only about its level. In this circumstance the procedure described in the previous section extracts all available information from the data.

In the case of a group of cohorts, however, the household data contain some information about the shape of the schedule which is not used by the two-stage procedure. We now consider an alternative method which uses all available information.

The basic idea is to fit the model schedule simultaneously to the household and individual data by combining the procedures described in Sections 2 and 3.

Thus, for each real cohort aged x ($x_0 \leqslant x \leqslant x_1$) we treat the household data $\{m_x\}$ as having a binomial distribution with parameters n_x and Π_x defined in Section 2.2, and the individual data as having a multinomial distribution with parameters m'_x and $\pi_{a|x}$ defined in Section 3.2, independently for each age.

The joint likelihood of the data is then a product binomial/multinomial distribution, and the log likelihood is simply the sum of (2.1) and (3.1), namely

$$logL = \sum_{x=x_0}^{x_1} \left\{ m_x log\Pi_x + s_x log(1-\Pi_x) + \sum_{a=a_0}^{x-1} m_{ax} log(\pi_{a|x}) \right\}. \tag{4.12}$$

Under the model nuptiality schedule we introduce the joint restrictions (2.3) and (3.3),

$$\Pi_x = F(x+\tfrac{1}{2}) \quad \text{and} \quad \pi_{a|x} = \frac{G(a+1)-G(a)}{G(x)} \tag{4.13}$$

The log-likelihood under the model becomes the sum of (2.4) and (3.4),

$$\sum_{x=x_0}^{x_1} \left\{ m_x log[F(x+\tfrac{1}{2})] + s_x log[1-F(x+\tfrac{1}{2})] + \sum_{a=a_0}^{x-1} m_{ax}(log[G(a+1)-G(a)] - logG(x)) \right\}. \tag{4.14}$$

This function depends on the two sets of data $\{m_x, s_x\}$ and $\{m_{ax}\}$ as well as the parameters μ, σ and c, through F and G, and may be optimized numerically.

For the cohort aged 25 to 29 in the Colombian survey we obtain

$$\hat{\mu} = 21.40, \ \hat{\sigma} = 6.11 \text{ and } \hat{c} = .838 \tag{4.15}$$

Comparison of these estimates with those obtained using individual data only shows that the household information has changed slightly the estimates of μ and σ. The estimate of c, on the other hand, is practically the same as that obtained earlier.

All the developments in Sections 2 and 3 extend naturally to the combined estimation procedure. Observed and fitted values pertaining to the household and individual data are defined as in (2.2), (2.6), (3.2) and (3.7), and the likelihood ratio goodness of fit criterion becomes simply the sum of (2.7) and (3.9), namely

$$\chi_1^2 = 2\sum_{x=x_0}^{x_1} \left\{ m_x log(\frac{P_x}{\hat{\Pi}x}) + s_x log(\frac{1-P_x}{1-\hat{\Pi}x}) + \sum_{a=a_0}^{x-1} m_{ax} log(\frac{P_{a|x}}{\pi_{a|x}}) \right\} \tag{4.16}$$

with degrees of freedom given by

$$\nu = (x_1 - x_0 + 1) + \sum_{x=x_0}^{x_1} (x-1-a_0) - 3 \tag{4.17}$$

The chi-squared statistic may easily be partitioned into components reflecting the contributions from the household and individual data. In assigning degrees of freedom to these components it would seem reasonable to consider the parameter c as estimated from the household data and the parameters μ and σ as estimated from the individual data.

For the cohort aged 25 to 29 we obtain the following results

	χ_1^2	ν	p-value
h	4.2	4	.383
i	79.4	73	.283
h+i	83.6	77	.284	(4.18)

indicating a good fit to both the household and the individual data.

The test for homogeneity of cohorts developed in Section 3.4 may still be applied to the individual data, but no analogous test exists for the household component.

Table 4.3 shows estimates of μ, σ and c obtained by applying these procedures to the six 5-year cohorts in the Colombian survey.

TABLE 4.3: Estimates of parameters of the model fitted to grouped marriage data from both the Colombia household and individual surveys (1976).

Ages	Estimates			Standard Errors			Goodness of Fit		
(1) x_0-x_1	(2) $\hat{\mu}$	(3) $\hat{\sigma}$	(4) $\hat{c}$	(5) s.e.$\hat{\mu}$	(6) s.e.$\hat{\sigma}$	(7) s.e.$\hat{c}$	(8) $\chi^2 1$	(9) ν	(10) p
20–24	21.80	6.14	.808	.524	.398	.046	60.0	52	.106
25–29	21.40	6.11	.838	.376	.314	.021	83.6	77	.284
30–34	20.70	5.07	.856	.250	.216	.012	130.4	102	.031
35–39	20.44	5.38	.846	.253	.213	.010	148.3	131	.143
40–44	21.23	5.76	.866	.265	.224	.011	135.9	149	.771
45–49	21.69	6.12	.851	.306	.254	.011	168.9	176	.636

TABLE 4.4: Estimates of the parameters of the model fitted to grouped marriage data from both the Colombia household and individual surveys (1976), when c is fixed at a preassigned level.

Age	c Fixed	Estimates			Standard Errors			Goodness of Fit		
(1) x_0-x_1	(2)	(3) $\hat{\mu}$	(4) $\hat{\sigma}$	(5) $\hat{c}$	(6) s.e.$\hat{\mu}$	(7) s.e.$\hat{\sigma}$	(8) s.e.c	(9) $\chi^2 1$	(10) ν	(11) p
15–19	No	41.39	15.18	14.192	8.276	3.910	16.553	32.3	32	.450
	Yes	24.18	7.10	.90	.346	.317	—	45.3	33	.075
	Yes	23.88	6.93	.85	.342	.316	—	46.4	33	.060
20–24	No	21.80	6.14	.808	.524	.398	.046	65.0	52	.106
	Yes	22.77	6.84	.90	.157	.170	—	68.6	53	.073
	Yes	22.24	6.46	.85	.180	.181	—	66.1	53	.107
25–29	No	21.40	6.11	.838	.376	.314	.021	83.6	77	.284
	Yes	22.34	6.89	.90	.215	.193	—	91.6	78	.139
	Yes	21.57	6.26	.85	.217	.188	—	84.1	78	.298

4.4 Fixing the Value of c

Examination of the results shown in Table 4.3 reveals that the estimates of c are quite low, specially for the younger cohorts. Because of previous work on the data from the Colombia National Fertility Survey, we know that there are mis-statements of marital status in the household survey which result in under estimation of proportions ever-married by age. To reduce the error introduced into the estimates of the mean and standard deviation of age at marriage by errors in the household data, the value of c can be fixed as it was in Section 2.6. Results of this exercise are shown in Table 4.4 for two values of c, .85 and .90.

For the age group 20-24 raising the value of c from its unconstrained estimate of .81 to .85 and then .90 raises the estimate of the mean from 21.8 to 22.2 to 22.8. Raising the value of c effectively rotates the fitted cumulative schedule about the current age of the cohort in question; it increases fitted proportions at older ages and depresses fitted proportions at younger ages, thereby raising the mean (and standard deviation). The range of the estimates of μ for different values of c (in this case a range of μ of one year produced by changing c by .1) is large enough so that one cannot place too much faith in the estimates unless one is fairly confident about the value of c.

The option of fixing c dramatically improves the estimate of μ and σ for the youngest cohort aged 15-19 at the time of the survey. As is shown in Table 4.4, the unconstrained estimates are quite absurd: a mean of 41.4 and a proportion ever-marrying of 14.2. Fixing c at either .85 or .90 produces estimates of μ which are much more reasonable, 23.9 and 24.2, respectively.

Note that for the cohort 15-19 the range in estimates of μ produced by a range in c of .05 (from .85 to .90) is only .3 year while for the cohorts aged 20-24 and 25-29 the ranges are .5 year and .8 year respectively. This result is due to the rotation effect produced on the fitted cumulative curve mentioned earlier; the effect on the mean will be greater for older than younger cohorts since for older cohorts more frequences at youngest ages are depressed in addition to more frequencies at the oldest ages being raised. The magnitude of the range in estimates of μ, and hence the degree of uncertainty about the estimate, when c is changed will depend both on the magnitude of the change in c and on the data. For some data sets, the range can be rather small. Hence, the option of fixing c can be valuable, but its value cannot be determined with precision in advance.

5. ESTIMATION FROM INDIVIDUAL DATA ON ALL WOMEN

5.1 The Data-Notation

The WFS individual interview is sometimes applied to an all-women sample; that is, a sample of women between the ages of 15 and 49, or a similar age range, selected irrespective of marital status. This has often been the case in WFS surveys in Latin America.

In such cases data on marital status by age at interview for all women and data on age at marriage by age at interview, are available for the *same* sample of women, a feature which simplifies estimation procedures. These data are often tabulated in completed years.

Table 5.1 presents such a set of data for the cohorts aged 25 to 29 in the Colombian individual survey. For each cohort the numbers marrying at each age are the same as shown earlier in Table 3.1, but this information has now been complemented by the numbers remaining single at the date of the interview. (Appendix Table 1 shows such data for all cohorts in the survey.)

TABLE 5.1: Tabulation of age at marriage by age at interview for women aged 25−29 at the time of the survey, Colombia (1976).

Age at Marriage	Age at Interview x				
(1) a	(2) 25	(3) 26	(4) 27	(5) 28	(6) 29
10	0	0	0	0	0
11	0	1	1	1	1
12	2	4	0	8	2
13	4	4	4	6	3
14	8	5	8	8	4
15	14	10	7	13	8
16	14	12	9	16	12
17	8	10	15	13	7
18	15	13	11	16	12
19	17	19	9	10	16
20	13	18	9	12	9
21	12	8	12	15	11
22	1	11	12	6	10
23	10	8	4	7	5
24	8	6	11	4	3
25	1	7	6	3	4
26		1	1	4	5
27			2	2	4
28				2	5
29					2
Total ever-married	127	137	121	146	123
Total never married	57	42	31	35	23
Total ever married by exact age x	126	136	119	144	121
Total never married by exact age x	58	43	33	37	25

An important feature of this type of data for an all-women sample is that although the experience of each cohort is incomplete, the cohort itself is complete, in the sense that it is represented by a sample of all its members. In this case the distribution of age at marriage is said to be *censored* (rather than truncated) by age at the interview.

From the point of view of estimation censoring does not present any special problems, and we shall be able to work directly with marriage frequencies, and thus estimate all three parameters of the model schedule.

Let us introduce the following notation with reference to Table 5.1:

m_{ax} = number of women married at age a completed years and aged x completed years at the interview

$m_x = \sum_{a=a_0}^{x} m_{ax}$ = number of ever-married women aged x completed years at the interview

s_x = number of single women aged x completed years at the interview

$n_x = m_x + s_x$ = total number of women aged x completed years at the interview

We now consider briefly a minor difficulty that arises in the treatment of women married at their current age, m_{xx}. As noted earlier the cohort aged x completed years has experienced a full year of exposure to marriage at each age a<x but less than a year at age x.

One possibility is to assume that women aged x completed years are on the average x+½ years, and to treat the number married at age x as married between exact ages x and x+½, and the number single at age x as not married by exact age x+½.

A simple alternative, which avoids any bias introduced by the above assumption, is to combine women married at age x completed years with women remaining single at age x completed years, and to treat the sum as the number remaining single at exact age x.

For this purpose we redefine

$m_x = \sum_{a=a_0}^{x-1} m_{ax}$ (=old $m_x - m_{xx}$)   number of women married by exact age x among women now aged x completed years

$s_x = n_x - m_x$ (=old $s_x + m_{xx}$)   number of women remaining single at exact age x among women now aged x completed years.

Note that the number of cases remains n_x, as we have just reclassified m_{xx} observations.

In the following discussion we adopt this simpler procedure. Extensions to treat m_{xx} as married by exact age x+½ are relatively simple, although details are cumbersome and will not be given.

5.2 Maximum Likelihood Estimation

We now consider fitting the model to a real cohort aged x_0 to x_1 completed years at the interview.

We shall treat the numbers $\{m_{ax}\}$ marrying at each age a<x and the number $\{s_x\}$ single at exact age x, for the cohort aged x, as having a multinomial distribution with parameters $\{\pi_{ax}\}$ $a=a_0,...,x-1$ where

π_{ax} = probability of marrying between exact ages a and a+1 for the cohort aged x.

Only $x-a_0$ parameters are required for each cohort, as the remaining parameter is

$1 - \sum_{a=a_0}^{x-1} \pi_{ax}$ = probability of remaining single at exact age x for the cohort aged x.

Note that we have introduced a different set of marriage probabilities for each single year cohort in the age group x_0 to x_1.

Assuming that the cohorts are mutually independent, the likelihood of the data is given by a product multinomial distribution, with log-likelihood

$$
\log L = \sum_{x=x_0}^{x_1} \left\{ \sum_{a=a_0}^{x-1} m_{ax} \log(\pi_{ax}) + s_x \log(1 - \sum_{a=a_0}^{x-1} \pi_{ax}) \right\}
\tag{5.1}
$$

The unrestricted m.l.e.'s of the $\{\Pi_{ax}\}$ are the sample proportions married at each age.

$$
p_{ax} = \frac{m_{ax}}{n_x} ,
\tag{5.2}
$$

with the proportion single estimated by s_x/n_x.

Under Coale's model nuptiality schedule we have

$$
\pi_{ax} = F(a+1) - F(a),
\tag{5.3}
$$

with the probability of remaining single at age x given by $1 - F(x)$, where F is the cumulative frequency of first marriages with parameters μ, σ and c introduced in Section 1.2.

Note that we are fitting the same model schedule F to all the single-year cohorts in the age group x_0 to x_1.

The log-likelihood (5.1) under the model (5.3) becomes

$$
\log L = \sum_{x=x_0}^{x_1} \left\{ \sum_{a=a_0}^{x-1} m_{ax} \log[F(a+1) - F(a)] + s_x \log[1 - F(x)] \right\}
\tag{5.4}
$$

This function depends on the data $\{m_{ax}\}$ and $\{s_x\}$, and on the parameters μ, σ and c through F, and may be optimized numerically in the usual fashion.

Applying this procedure to the cohort aged 25 to 29 in the Colombian individual survey we obtain the estimates

$$
\hat{\mu} = 21.27, \quad \hat{\sigma} = 6.02 \quad \text{and} \quad \hat{c} = .910
\tag{5.5}
$$

with estimated standard errors, based on an approximation to the information matrix,

$$
\text{s.e.}\hat{\mu} = .363, \quad \text{s.e.}\hat{\sigma} = .304 \quad \text{and} \quad \text{s.e.}\hat{c} = .025
\tag{5.6}
$$

TABLE 5.2: Estimates of parameters of the model fitted to grouped marriage data from the Colombia individual survey (1976). All-women sample.

Ages	Estimates			Standard Errors			Goodness of Fit			Homogeneity of Cohorts		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
$x_0 - x_1$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{c}$	s.e.$\hat{\mu}$	s.e.$\hat{\sigma}$	s.e.$\hat{c}$	$\chi^2 1$	ν	p	$\chi^2 1$	ν	p
20–24	21.62	6.01	.887	.609	.459	.064	61.6	52	.170	44.0	42	.388
25–29	21.27	6.02	.910	.363	.304	.025	80.3	77	.376	67.3	62	.302
30–34	20.64	5.02	.915	.238	.205	.014	124.7	102	.063	90.9	82	.235
35–39	20.44	5.38	.885	.252	.217	.013	143.5	132	.233	111.2	106	.346
40–44	21.22	5.75	.919	.270	.221	.013	127.6	152	.926	99.5	122	.932
45–49	21.68	6.12	.908	.305	.252	.015	166.9	182	.783	136.1	146	.710

We note that although the estimates of μ and σ are similar to those obtained earlier using the individual data for the ever-married women, the estimate of c is much more reasonable than that obtained using the household data (even though there are many fewer observations in the individual data), a clear indication of the better quality of the individual data. Results for all 5-year cohorts in the Colombian survey are summarised in Table 5.2 and confirm the above conclusion.

Another indication of the better quality of these data is shown by the results of fitting the model just to data on single and ever-married women by age at interview as was done in Section 2. Results, which are shown in Appendix Table 4, are much more stable.

The option of fixing c at a value believed to reflect the proportion of women who will eventually marry and re-estimating μ and σ can be used to advantage even with an all women sample. The unconstrained estimates for the cohort aged 15-19 are $\hat{\mu}=29.8$, $\hat{\sigma}=10.3$, and $\hat{c}=2.7$; these values are clearly absurd. If c is fixed at the value .90, the estimates of μ and σ fall to $\hat{\mu}=23.7$ and $\hat{\sigma}=7.0$; if c is fixed at .85 the estimates are slightly lower: $\hat{\mu}=23.4$, $\hat{\sigma}=6.8$. Although the range of estimates of the mean produced by fixing c at .85 and .90 is not so small that we could predict with confidence a precise value of the ultimate mean, either choice of c (or any other plausible one) produces estimates which are far more plausible than those obtained when c is not fixed.

5.3 Goodness of Fit of the Model

The unrestricted m.l.e.'s of the parameters $\{\pi_{ax}\}$ are the observed proportions married at each age defined in (5.2).

The restricted m.l.e.'s of the same parameters under the model, or fitted proportions marrying at each age, are given by

$$\hat{\pi}_{ax}=\hat{F}(a+1)-\hat{F}(a), \tag{5.7}$$

with the fitted proportion single given by $1-\hat{F}(x)$, where $\hat{F}$ denotes F evaluated at the m.l.e.'s $\hat{\mu}$, σ and c.

Observed and fitted proportions marrying at each age and remaining single at their current age for the cohorts aged 25 to 29 in the Colombian survey are given in Table 5.3 (Columns 2-6 and 8).

The likelihood ratio and Pearson chi-squared statistics for testing the goodness of fit of the model are given by

$$\chi_l^2 = 2\sum_{x=x_0}^{x_1}\left\{\sum_{a=a_0}^{x-1} m_{ax}\log[p_{ax}/\hat{\pi}_{ax}]+s_x\log[\frac{s_x/n_x}{1-\hat{F}(x)}]\right\} \tag{5.8}$$

and

$$\chi_p^2 = \sum_{x=x_0}^{x_1} n_x\left\{\sum_{a=a_0}^{x-1} \frac{(p_{ax}-\hat{\pi}_{ax})^2}{\hat{\pi}_{ax}} + \frac{[s_x/n_x - 1+\hat{F}(x)]^2}{1-F(x)}\right\} \tag{5.9}$$

In large samples both criteria are distributed as chi-squared statistics with degrees of freedom ν given by

$$\nu = \sum_{x=x_0}^{x_1} (x-a_0)-3, \tag{5.10}$$

which is the total number of independent cells less the number of parameters estimated.

TABLE 5.3: Proportions marrying at each age among all women 25—29, Colombia (1976). All-women sample.

Age at Marriage	Age at Interview x					Pooled	Fitted	Difference
	25	26	27	28	29			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
a			$p_{a\vert x}$			$\overline{\pi}_{a\vert x}$	$\hat{\pi}_{a\vert x}$	$\overline{\pi}_{a\vert x}-\hat{\pi}_{a\vert x}$
11	.000	.006	.007	.006	.007	.005	.005	.000
12	.011	.022	.000	.044	.014	.019	.013	.006
13	.022	.022	.026	.033	.021	.025	.027	—.003
14	.043	.028	.053	.044	.027	.039	.045	—.005
15	.076	.056	.046	.072	.055	.062	.060	.001
16	.076	.067	.059	.088	.082	.075	.072	.003
17	.043	.056	.099	.072	.048	.063	.077	—.014
18	.082	.073	.072	.088	.082	.080	.077	.002
19	.092	.106	.059	.055	.110	.084	.074	.011
20	.071	.101	.059	.066	.062	.072	.068	.005
21	.065	.045	.079	.083	.075	.069	.060	:009
22	.005	.061	.079	.033	.068	.048	.053	—.005
23	.054	.045	.026	.039	.034	.040	.045	—.005
24	.043	.034	.072	.022	.021	.038	.039	—.001
25		.039	.039	.017	.027	.031	.033	—.001
26			.007	.022	.034	.022	.027	—.006
27				.011	.027	.019	.023	—.004
28					.034	.035	.019	.016
Probability of remaining single at exact age x	.315	.240	.217	.204	.171	.175	.183	
Number of cases	184.	179.	152.	181.	146.			

For the cohort aged 25 to 29 we obtain

$$\chi^2_l = 80.3 \qquad \text{p value} = .376$$
$$\chi^2_p = 74.9 \qquad \text{p value} = .547 \qquad (5.11)$$
$$\nu = 77$$

indicating an excellent fit to the data.

Results of the likelihood ratio goodness of fit test for all cohorts in the Colombian survey are given in Table 5.2.

5.4 Homogeneity of Cohorts

We now introduce a test for homogeneity of cohorts for all-women samples which is analogous to that introduced for ever-married samples in Section 3.4.

We assume that all cohorts have followed the same nuptiality schedule $\{\pi_a\}$ which is otherwise unrestricted, so that the probability of marrying between exact ages a and a+1 is

$$\pi_{ax}=\pi_a \quad \text{for all cohorts x, } x_0 \leqslant x \leqslant x_1 \tag{5.12}$$

with the probability of remaining single at exact age x being simply

$$1-\sum_{a=a_0}^{x-1} \pi_a.$$

The likelihood function (5.1) under the homogeneous model (5.12) is given by

$$\log L = \sum_{x=x_0}^{x_1} \left\{ \sum_{a=a_0}^{x-1} m_{ax}\log(\pi_a)+s_x\log(1-\sum_{a=a_0}^{x-1} \pi_a) \right\} \tag{5.13}$$

It can be shown that the estimates which maximise the likelihood are given by

$$\bar{\pi}a = \begin{cases} \dfrac{\displaystyle\sum_{x=x_0}^{x_1} m_{ax}}{\displaystyle\sum_{x=x_0}^{x_1} n_x} \, , a<x_0 \\[4em] \dfrac{\displaystyle\sum_{x=a+1}^{x_1} m_{ax}}{\displaystyle\sum_{x=a+1}^{x_1} [\sum_{a=a}^{x-1} m_{ax}+s_x]} [1-\sum_{a=a_0}^{a-1} \bar{\pi}_a] , x_0 \leqslant a < x_1 \end{cases} \tag{5.14} \tag{5.15}$$

The expression for $a<x_0$, where there is no censoring, follows from a straightforward binomial argument. The expression for $a \geqslant x_0$ follows from a conditional probability argument. Note that (5.15) estimates the probability of marrying between ages a and a+1 as the product of two quantities: (1) the number married between ages a and a+1 divided by the number single at exact age a, which estimates the conditional probability of marrying between ages a and a+1 conditional on being single at age a, and (2) a previously obtained estimate of the probability of being single at exact age a.

The estimates given at (5.14)-(5.15) are identical to those that would be obtained by constructing a life table where

$$\sum_{x=\max(x_0,a+1)}^{x_1} m_{ax}$$

represents the number married between exact ages a and a+1, and s_x represents the number censored at exact age x. We refer to these estimates as the *pooled* (or life table) estimates of the first marriage frequencies. Pooled estimates for the cohorts aged 25 to 29 in the Colombian survey are shown in Table 5.3 (Column 7).

The likelihood ratio and Pearson chi-squared statistics for testing the hypothesis that all cohorts in the group have followed the same nuptiality schedule are given by

$$\chi_1^2 = 2\sum_{x=x_0}^{x_1} \left\{ \sum_{a=a_0}^{x} m_{ax}\log\left(\frac{Pax}{\overline{\pi}a}\right) + s_x\log\left(\frac{s_x/n_x}{x-1} \right) \right\}, \tag{5.16}$$

$$\frac{s_x/n_x}{1-\sum_{a=a_0}\overline{\pi}_a}$$

and

$$\chi_p^2 = \sum_{x=x_0}^{x_0} n_x \left\{ \sum_{a=a_0}^{x-1} \frac{(P_{ax}-\overline{\pi}a)^2}{\overline{\pi}a} + \frac{(\frac{s_x}{n_x}-1+\sum_{a=a_0}^{x-1}\overline{\pi}_a)^2}{1-\sum_{a=a_0}^{x-1}\overline{\pi}_a} \right\} \tag{5.17}$$

and are distributed in large samples as chi-squared statistics with degrees of freedom

$$\nu = \sum_{x=x_0}^{x_1-1} (x-a_0), \tag{5.18}$$

which is the number of independent cells, $x-a_0$ for each cohort aged x, less the number of parameters in the homogeneous model which is x_1-a_0.

For the cohort aged 25 to 29 we obtain

$$\chi_1^2 = 67.3 \qquad \text{p value} = .302$$
$$\chi_p^2 = 61.3 \qquad \text{p value} = .501$$
$$\nu = 62 \tag{5.19}$$

indicating, as we would have expected from the good fit found earlier, that the cohorts are fairly homogeneous.

Results of the likelihood ratio test for other cohorts are given in Table 5.2.

Since Coale's model (5.3) is a restricted case of the homogeneous model (5.12), we can obtain a chi-square test comparing the two models by direct subtraction of the goodness of fit chi-squares and the degrees of freedom corresponding to each model.

For the cohort aged 25 to 29 we obtain from (5.11) and (5.19)

$$\chi_1^2 = 13.1 \qquad \text{p value} = .598$$
$$\chi_p^2 = 13.6 \qquad \text{p value} = .558$$
$$\nu = 14 \tag{5.20}$$

The likelihood ratio chi-square statistic is the same that would be obtained by direct use of the ratio $\overline{\pi}_a/\hat{\pi}_a$.

5.5 Fitting and Forecasting

In fitting a model schedule to a cohort still undergoing the marriage process we obtain estimates of μ and σ which best reproduce the experience of the cohort up to the date of the interview. The goodness of fit criteria considered so far pertain only to this incomplete experience.

As noted earlier, a model that fits the experience of a cohort to date well will not necessarily forecast its future behaviour accurately. Yet one of the purposes of fitting the model may be to estimate the mean age at marriage, which involves an element of forecasting for all but the oldest cohorts.

TABLE 5.4: Estimates of the mean age at first marriage and the proportion ever marrying obtained by artificially censoring the available data.

	Current Age of Cohort									
Cohort Last Observed When Aged $x_0 - x_1$ (1)	35–39					40–44				
	Estimate		SE Estimates		p	Estimates		SE Estimates		p
	(2) $\hat{\mu}$	(3) c	(4) $\hat{s.e.\mu}$	(5) $\hat{s.e.c}$	(6)	(7) $\hat{\mu}$	(8) $\hat{c}$	(9) $\hat{s.e.\mu}$	(10) $\hat{s.e.c}$	(11)
20–24	20.16	.875	.507	.051	.256	22.58	1.090	.929	.113	.993
25–29	20.00	.852	.315	.023	.727	21.87	.981	.554	.041	.922
30–34	20.13	.864	.256	.017	.592	21.40	.935	.365	.019	.826
35–39	20.44	.885	.255	.014	.233	21.17	.917	.298	.015	.870
40–44						21.22	.919	.284	.013	.926

If the model is true, of course, and there are no errors in the data, the procedures described herein will produce estimates of the parameters which will be correct within the limits of sampling variability. In fitting the model to data generated from the standard with $\mu=21.36$ and $\sigma=6.58$ we have been able to recover the correct parameter values by truncating or censoring the data as early as ages 15 to 19.

It is therefore interesting to examine whether we would have obtained the same estimates of the parameters for the cohort now aged for example 40-44, if we had observed them at an earlier point in time. To accomplish this task we assume that women now aged, for example 40-44, who reported an age at marriage of 20 would have reported the same age at marriage 5, 10 or 15 years ago. In short, we must assume that dates (both of birth and of marriage) are reported correctly. We then re-estimate the parameters for a cohort by utilizing data which would have been gathered 5, 10, 15, . . ., years earlier. Results for two cohorts are presented in Table 5.4.

Consider first the cohort aged 35-39 at the time of the survey. Estimates of the mean age at marriage and proportion ever-marrying are 20.44 and .885 respectively. If the same women had been interviewed five years earlier, when they were aged 30-34, their reports would have produced estimates of μ and c of 20.13 and .864 respectively. Even 15 years earlier their experience to date would have produced quite similar estimates of 20.16 and .875. Figure 5.1 shows the pooled estimates for this cohort, as well as the fitted schedules based on the experience of the cohort up to date, and based on the experience censored at ages 20-24.

We conclude that in this case, estimates which would have been produced earlier are remarkably similar to those actually resulting from the survey. The biggest difference arises between estimates based on the current data and those which would have resulted had the cohort been interviewed 10 years earlier, when aged 25-29; the mean would have been underestimated by .44 and the proportion ever-marrying underestimated by .033. The implication of this finding is that while the model would have predicted well in this example, the actual prediction error is higher than the estimated standard errors of the estimates. Hence, one must expect rather less precision in the estimates of the eventual mean age at first marriage and proportion ever-marrying for young cohorts than would be implied by the estimated standard errors.

The cohort aged 40-44 reveals a more dismal picture. Estimates of the parameters based on the current data are almost identical to those which would have been obtained five years earlier. After this point however, estimates of both the mean age at marriage and the proportion ever-marrying rise monotonically the further back in time one assumes the survey was taken. The

FIGURE 5.1: Pooled estimates of proportions marrying at each age for the cohort aged 35–39 and fitted schedules based on the experience up to ages 20–24 and up to ages 35–39; all-women sample.
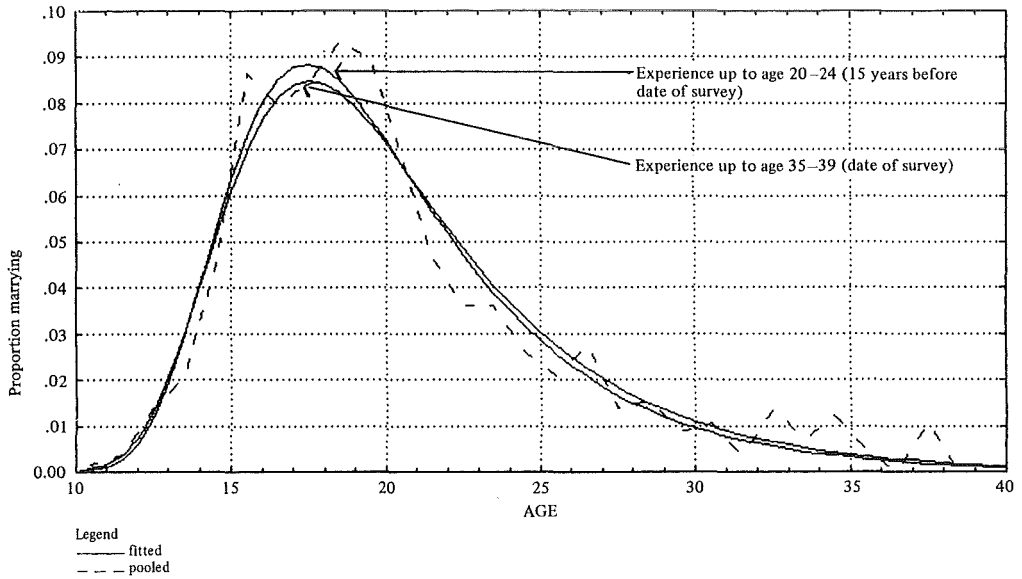
FIGURE 5.2: Pooled estimates of proportions marrying at each age for the cohort aged 40–44 and fitted schedules based on the experience up to ages 20–24 and up to ages 40–44; all-women sample.

45
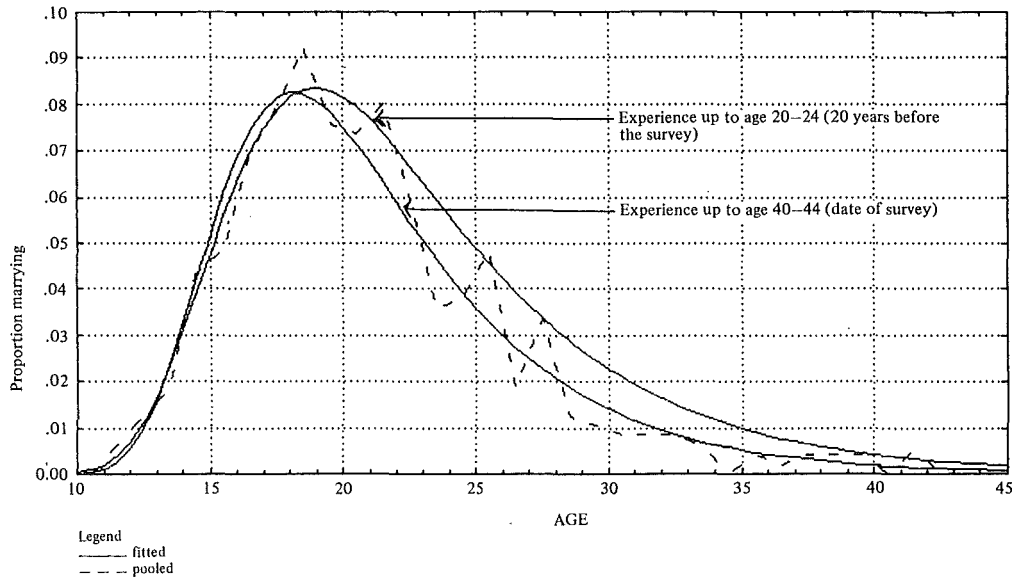
estimates based on data which would have been collected 20 years earlier are clearly inconsistent with the current estimates; the mean is over-estimated by nearly 1.4 years and the proportion ever-marrying over-estimated by .171. We conclude that in some cases the model may not predict well.

It must be emphasized that if the data do conform to the model, artificial censoring or truncation will not affect the estimates of the parameters. The problem with the data for the cohort aged 40-44 at the time of the survey (and to a much lesser extent with the data for the cohort aged 35-39) is that they simply do not conform well to the model. This lack of conformity is evident in a plot of the pooled estimates for the cohort aged 40-44, shown in Figure 5.2. The data are clearly irregular and do not form a smooth curve with a single peak. There are big positive outliers at ages 18, 21, 25 and 27 and big negative outliers at ages 23, 26 and 28. When the experience of the cohort up to ages 40-44 is used the fitted schedule is anchored at the upper tail by a large number of points which conform to the model. As one successively discards the points at ages 35-39, 30-34 and 25-29, the outliers acquire more prominence and the best fitting curve (in a maximum likelihood sense) moves steadily to the right, thereby implying a larger mean, as clearly seen from Figure 5.2.

The lesson to be learned is straightforward. If one is fitting the model to a series of points which are highly erratic (due to random or non-random variations in age reporting such as caused for example by digit preference) especially in the central age groups, then the predictive power of the model is likely to be small indeed. One can best use the model in such a case as a diagnostic or smoothing device. If, on the other hand, the data form a series which is smooth and single-peaked, then one can place more faith in the predictive power of the model. Nevertheless, period effects can modify the predictive power even when the model to date fits well. The model cannot foresee war, famine, social change or revolution; its predictions are limited by the assumption that past behaviour reveals something about future behaviour.

6. ESTIMATION FROM UNGROUPED DATA

6.1 The Data

In Sections 3 and 5, dealing with estimation using individual data from ever-married or all-women samples, we have used age at marriage and age at interview tabulated in completed years. We refer to this type of data as grouped data.

In WFS individual surveys these ages are calculated from three dates — namely date of respondent's birth, date of first marriage and date of interview — all available or imputed in month/year form. Thus the ages under reference are 'known' to the nearest month and may be taken to represent exact years. We refer to this type of data as ungrouped data.

We now consider fitting the model using ungrouped data or exact ages, and discuss estimation and goodness of fit procedures appropriate for ever-married and all-women samples.

6.2 Estimation from All-women Samples

Consider first a sample or cohort of n respondents, of whom m are ever-married. For convenience let i=1,...,m index those ever-married and let i=m+1,...,n index those single. For the i-th respondent let

$$x_i = \text{age at interview in exact years } (i=1,...,n)$$

$$a_i = \text{age at marriage in exact years } (i=1,...,m)$$

Note that in an all-women sample the distribution of age at marriage is *censored* by age at the interview ($a_i \leqslant x_i$ for i$\leqslant$m but a_i is undefined for i$>$m).

Under Coale's model nuptiality schedule the probability of marrying between exact ages a and a+da is f(a)da, where f(a) is the frequency of first marriages defined at (1.1). Hence, the contribution to the likelihood of a women married at exact age a_i is simply

$$f(a_i), \qquad i=1,...,m \tag{6.1}$$

On the other hand, the probability of remaining single at exact age x is 1-F(x), where F(x) denotes the cumulative frequency of first marriages defined at (1.6). Hence, the contribution to the likelihood of a women single at exact age x_i is simply

$$1-F(x_i), \qquad i=m+1,...,n \tag{6.2}$$

The logarithm of the likelihood function under the model is then

$$\log L = \sum_{i=1}^{m} \log[f(a_i)] + \sum_{i=m+1}^{n} \log[1-F(x_i)] \tag{6.3}$$

This function depends on the data $\{a_i, x_i\}$ and the parameters μ, σ and c through f and F, and may be optimized numerically using the procedures mentioned in Section 8.

For the cohort aged 25 to 29 completed years in the Colombian individual survey we obtain

$$\hat{\mu}=21.17, \qquad \hat{\sigma}=5.97 \text{ and } \hat{c}=.904 \tag{6.4}$$

which are similar to the estimates obtained from grouped data at (5.5).

Estimates of the standard errors of the estimates, obtained from a numerical approximation to the information matrix, are

$$\text{s.e.}\hat{\mu}=.332, \qquad \text{s.e.}\hat{\sigma}=.276 \text{ and } \text{s.e.}\hat{c}=.023 \tag{6.5}$$

which are also comparable to those obtained using grouped data at (5.6).

TABLE 6.1: Estimates of the parameters of the model fitted to ungrouped marriage data from the Colombia individual survey (1976). All-women sample.

Cohort	Estimates			Standard Errors			Goodness of Fit
(1) $x_0 - x_1$	(2) $\hat{\mu}$	(3) $\hat{\sigma}$	(4) $\hat{c}$	(5) s.e.$\hat{\mu}$	(6) s.e.$\hat{\sigma}$	(7) s.e.c.	(8) D
20—24	21.57	6.01	.895	.480	.363	.050	.024
25—29	21.17	5.97	.904	.332	.276	.023	.023
30—34	20.61	5.06	.917	.253	.211	.014	.020
35—39	20.45	5.42	.889	.244	.212	.014	.041
40—44	21.15	5.75	.919	.269	.233	.013	.033
45—49	21.69	6.40	.913	.333	.273	.014	.042

Our experience indicates that estimates of standard errors obtained from ungrouped data are generally more stable and reliable than those obtained from grouped data.

Estimates of all three parameters and their standard errors for the six 5-year cohorts in the Colombian individual survey are given in Table 6.1.

Let $\hat{F}(a)$ denote the fitted nuptiality schedule obtained by evaluating the function F(a) at the m.l.e.'s $\hat{\mu}$, $\hat{\sigma}$ and $\hat{c}$. Note that $\hat{F}(a)$ is a maximum likelihood estimator of the cumulative frequency of first marriages under the assumption that the latter has the parametric form introduced in Section 1.2.

6.3 The Kaplan-Meier Estimate

We now consider a procedure for assessing the goodness of fit of the model which is based on a comparison of the fitted nuptiality schedule with a non-parametric estimate of the cumulative frequency of first marriages, which maximizes the likelihood of the data over the class of all distribution functions.

The non-parametric estimate in question, which will be denoted $\bar{F}(a)$, is the product-limit estimate of a distribution function from censored data developed by Kaplan and Meier (1958), and represents an extension to continuous data of basic life table concepts.

Let $a_{(1)} < a_{(2)} < ... < a_{(k)}$ denote the distinct ages at marriage observed in the sample, with $k \leq n$ and define $a_{(0)} = -\infty$ and $a_{(k+1)} = \infty$. Let m_i denote the number of women married at exact age $a_{(i)}$, and let l_i denote the number of single women at exact age x where $a_{(i)} \leq x < a_{(i+1)}$, for $i = 1, ..., k$.

In life table terminology m_i represents the number of "deaths" at exact age $a_{(i)}$, and l_i represents the number of "losses" or observations censored between exact ages $a_{(i)}$ and $a_{(i+1)}$, including losses at $a_{(i)}$ but not at $a_{(i+1)}$.

For each age $a_{(i)}$ define the risk set

$$R_i = \sum_{j=i}^{k} (m_j + l_j)$$

This set comprises all women remaining single just before age $a_{(i)}$ and thus "at risk" of first marrying at exact age $a_{(i)}$.

The product-limit estimate of the probability of marrying by exact age $a_{(i)}$ is then

$$\bar{F}[a_{(i)}] = 1 - \prod_{j=1}^{i} [1 - \frac{m_j}{R_j}]. \tag{6.7}$$

Note that m_j/R_j estimates the probability of marrying at exact age $a_{(j)}$ conditional on being single just before that age; the quantity in brackets estimates the probability of remaining single at age $a_{(j)}$; the product from $j=1$ to i estimates the probability of remaining single from ages $a_{(1)}$ to $a_{(i)}$; and thus $\bar{F}[a_{(i)}]$ estimates the probability of marrying by exact age $a_{(i)}$.

The estimate may be extended to any age $a < x_{(n)}$, the largest censoring age, and other than the sample points $a_{(i)}$, by setting $F[a_{(0)}]=0$ and

$$\bar{F}(a) = \bar{F}[a_{(i)}] \quad \text{for } a_{(i)} \leqslant a < a_{(i+1)} \tag{6.8}$$

Details of the derivation of $\bar{F}$ using a maximum likelihood argument may be found in Kaplan and Meier (1958, p.475).

In the grouped data case the estimate (6.7) turns out to be the same as the pooled estimate introduced in Section 5.4, which is also based on a life table argument.

Note that we have two estimates of the cumulative frequency of first marriages, an estimate $\bar{F}(a)$ from the class of all distribution functions, and an estimate $\hat{F}(a)$ from the subclass of functions having the parametric form proposed by Coale and McNeil (1972).

Since the two estimates are m.l.e.'s one might expect these developments to lead to a likelihood ratio test of the goodness of fit of the model. Unfortunately such is not the case, because the ratio of the likelihoods does not give a fair comparison between a discrete function such as $\bar{F}(a)$ — which assigns positive probability to the actual observed values and zero probability to any other value — and a continuous function such as $\hat{F}(a)$ — which assigns positive probability density to any possible value whether observed or not.

Since the two estimates are consistent, however, it is possible to assess the goodness of fit of the model by a direct comparison of $\bar{F}(a)$ and $\hat{F}(a)$ for all ages a. In particular, a summary measure of the goodness of fit of the model is given by the largest difference between the two estimates.

$$D = \text{supremum}_{a_{(1)} \leqslant a \leqslant a_{(k)}} \quad | \hat{F}(a) - \bar{F}(a) |. \tag{6.9}$$

It can be shown that the maximum must occur at one of the sample points, so that

$$D = \max_{i} [\max\{|\hat{F}[a_{(i)}] - \bar{F}[a_{(i)}]| , |\hat{F}[a_{(i)}] - \bar{F}[a_{(i-1)}]| \}], \tag{6.10}$$

The statistic D is a censored-sample analog of Kolmogorov-Smirnov's goodness of fit statistic. The distribution of D is known for complete samples, but its properties under censoring have — to our knowledge — not been established. Thus D may be used as a descriptive measure of goodness of fit but not as a formal test.

Figure 6.1 shows the parametric and non-parametric estimates $\bar{F}(a)$ and $\hat{F}(a)$ of the cumulative frequencies of first marriage for the cohort age 25 to 29 in the Colombian survey.

The closeness of the two curves indicates a good fit of Coale's model nuptiality schedule to the data. The largest distance between the curves is $D=.023$. There are two problems in interpreting this statistic, one of a general nature and one specific to WFS data. First, it is affected by the proportion who ever marry; if the cumulative curve reached only half the level shown in Figure 6.1, then *ceteris paribus,* D would be only half as big. The second problem, specific to WFS

FIGURE 6.1: Kaplan-Meier [$\bar{F}(a)$] and fitted [$\hat{F}(a)$] proportions ever-married among women aged 25–29 at the time of the survey; all-women sample.

Legend
– – – fitted [$\hat{F}(a)$]
——— Kaplan-Meier [$\bar{F}(a)$]
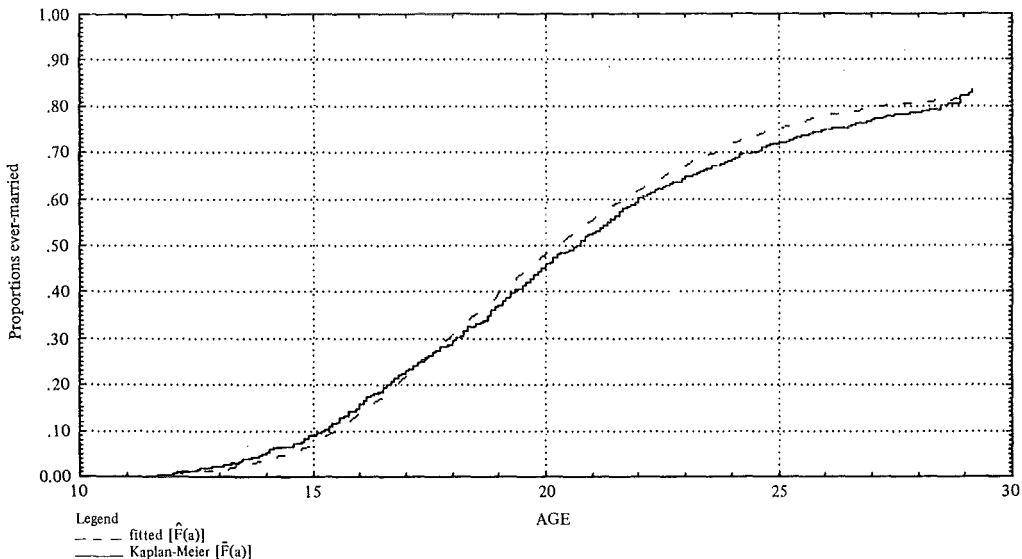
data, is that the ages at marriage are not really distinct since they are all expressed in twelfths of a year. Hence heaping on these fractions is inevitable. Since $\hat{F}(a)$ is continuous and $\bar{F}(a)$ is a step function more heaping will invariably increase D. Thus, both the plot and the statistic tend to make the goodness of fit appear worse than it would be if exact ages were used.

6.4   Estimation from Ever-married Samples

We now adapt these procedures to the case of an ever-married sample.

For the i-th respondent in a sample or cohort of m ever-married respondents let

$$x_i = \text{age at interview in exact years}$$
$$a_i = \text{age at marriage in exact years}$$

Note that in a sample of ever-married women the distribution of age at marriage is *truncated* by age at the interview ($a_i \leqslant x_i$). We therefore argue in terms of conditional probabilities of marriage.

The probability of marrying between exact ages a and a+da conditonal on marrying by exact age x, under Coale's model nuptiality schedule, is given by g(a|x)da where

$$g(a|x) = \frac{g(a)}{G(x)} , \tag{6.11}$$

where g(a) and G(x) denote the probability density and the cumulative distribution functions of age at marriage, defined at (1.2) and (1.8).

The logarithm of the likelihood function under the model is then

$$\log L = \sum_{i=1}^{m} \left\{ \log[g(a_i)] - \log[G(x_i)] \right\}. \tag{6.12}$$

50

TABLE 6.2: Estimates of the parameters of the model fitted to ungrouped marriage data from the Colombia individual survey (1976). Ever-married women sample.

Cohort	Estimates		Standard Errors		Goodness of Fit
(1) $x_0 - x_1$	(2) $\hat{\mu}$	(3) $\hat{\sigma}$	(4) $s.e.\hat{\mu}$	(5) $s.e.\hat{\sigma}$	(6) D
20–24	21.51	5.97	.517	.395	.053
25–29	21.10	5.91	.327	.274	.037
30–34	20.60	5.05	.245	.207	.028
35–39	20.45	5.42	.253	.216	.038
40–44	21.15	5.74	.282	.232	.037
45–49	21.69	6.40	.342	.277	.047

This function depends on the data $\{a_i, x_i\}$ and the parameters μ and σ through g and G, and may be optimized numerically as noted in Section 8.

For the cohorts aged 25 to 29 completed years in the Colombian individual survey, we have

$$\hat{\mu}=21.10 \text{ and } \hat{\sigma}=5.91, \tag{6.13}$$

which are fairly similar to those obtained from grouped data.

Estimates of the standard errors of $\hat{\mu}$ and $\hat{\sigma}$ for this cohort, obtained from a numerical approximation to the information matrix, are

$$s.e.\hat{\mu}=.327 \text{ and } s.e.\hat{\sigma}=.274, \tag{6.14}$$

which are comparable to those obtained using grouped data.

Estimates of the parameters, as well as associated standard errors, for six 5-year cohorts in the Colombian individual survey are given in Table 6.2.

For each cohort the maximum likelihood estimate of the conditional probability of marrying by exact age a given marriage by exact age x>a is given by

$$\hat{G}(a|x)=\hat{G}(a)/\hat{G}(x), \tag{6.15}$$

where $\hat{G}$ denotes the cumulative distribution function G evaluated at the m.l.e.'s $\hat{\mu}$ and $\hat{\sigma}$.

6.5   A Product-limit Estimate for Truncated Data

In order to assess the goodness of fit of the model to a sample of ever-married women we now develop a non-parametric estimate of the cumulative distribution function from a truncated sample, which maximizes the likelihood of the data over the class of all distribution functions.

The estimate, which will be denoted $\bar{G}(a|x)$, is analogous to the Kaplan-Meier product-limit estimate for censored samples, and hence will be referred to as the product-limit estimate for truncated samples. We first introduce the notation and the estimate and then proceed to its derivation.

Let $a_{(1)} < a_{(2)} < ... < a_{(k)}$ denote the distinct ages at marriage in the sample and define $a_{(0)}=-\infty$ and $a_{(k+1)}=\infty$. Let m_i denote the number of women married at exact age $a_{(i)}$ and let t_i denote the number of women interviewed at exact age x for $a_{(i)} \leq x < a_{(i+1)}$.

Here t_i represents the number of observations truncated at ages between $a_{(i)}$ and $a_{(i+1)}$, including those truncated at $a_{(i)}$ but not at $a_{(i+1)}$. Note that since all women in the sample are ever-married and interviewed

$$\sum_{i=1}^{k} m_i = \sum_{i=1}^{k} t_i = n \qquad (6.16)$$

Let us now define the quantity

$$M_i = \sum_{j=1}^{i} (m_j - t_j) \qquad (6.17)$$

which can be seen to be the number of women married at age $a_{(i)}$ or earlier and interviewed at age $a_{(i+1)}$ or later.

Then the product-limit estimate of the probability of marrying by age $a_{(i)}$ conditional on marrying by age $x_{(m)}$, the largest observed age at interview in the sample, is

$$\overline{G}[a_{(i)} | x_{(m)}] = \prod_{j=i}^{k-1} \frac{M_j}{M_j + m_j + 1} \qquad (6.18)$$

The ratio $M_j/(M_j + m_j + 1)$ is the ratio of the number of women married by age $a_{(j)}$ and interviewed at age $a_{(j+1)}$ or later, to the number of women married by age $a_{(j+1)}$ and interviewed at age $a_{(j+1)}$ or later, and thus estimates the probability of marrying by age $a_{(j)}$ conditional on marrying by age $a_{(j+1)}$.

The product of these probabilities from $j=i$ to $k-1$ gives an estimate of the probability of marrying by age $a_{(i)}$ conditional on marrying by age $a_{(k)}$. Since there are no marriages in the sample between ages $a_{(k)}$ and $x_{(m)}$, these probabilities may also be considered conditional on marrying by age $x_{(m)}$.

The estimate may be extended to any age $a < x_{(m)}$ other than the sample points $a_{(i)}$ by letting

$$G(a_{(0)} | x_{(m)}) = 0$$

and

$$G(a | x_{(m)}) = G(a_{(i)} | x_{(m)}) \text{ for } a_{(i)} \leqslant a < a_{(i+1)} \qquad (6.19)$$

We now show that G is a maximum likelihood estimator of the conditional distribution function in the class of all distribution functions.

Let $G_m(a)$ denote the probability of marrying by exact age a conditional on marrying by exact age $x_{(m)}$, considered as an arbitrary function to be determined so as to maximize the likelihood.

For a sample of m women, where the ith woman married at age a_i and was interviewed at age x_i, $[a_i \leqslant x_i \leqslant x_{(m)}]$, the likelihood is given by

$$L = \prod_{i=1}^{m} \frac{G_m(a_i) - G_m(a_i - 0)}{G_m(x_i)} \qquad (6.20)$$

where $G_m(a_i - 0)$ denotes the value of $G_m(a)$ immediately at the left of a_i.

Let $a_{(i)}$, m_i and t_i be as defined earlier, and let $x_{(ij)}$ for $j = 1, \ldots, t_i$ denote the exact ages at interview of the t_i women whose experience was truncated between $a_{(i)}$ and $a_{(i+1)}$, including those truncated at $a_{(i)}$ but not at $a_{(i+1)}$.

The likelihood function may then be written as

$$L = \prod_{i=1}^{k} \left\{ G_m[a_{(i)}] - G_m[a_{(i)} - 0] \right\}^{m_i} \left\{ \prod_{j=1}^{t_i} G_m[x_{(ij)}] \right\}^{-1} \qquad (6.21)$$

Note that (6.21) is just a restatement of (6.20) since

$$\prod_{i=1}^{m} \left\{ G_m(a_i) \cdot G_m(a_i \cdot 0) \right\} = \prod_{i=1}^{k} \left\{ G_m[a_{(i)}] \cdot G_m[a_{(i)} \cdot 0] \right\}^{m_i}$$

and

$$\prod_{i=1}^{m} G_m(x_i)^{-1} = \prod_{i=1}^{k} \prod_{j=1}^{t_i} G_m[x_{(ij)}]^{-1}$$

To maximize the likelihood we would like to make $G_m[a_{(i)}]$ as large as possible, and $G_m[a_{(i)} \cdot 0]$ and $G_m[x_{(ij)}]$ as small as possible, under the restriction that G_m is non-decreasing.

Since $a_{(i)} \leqslant x_{(ij)} \leqslant a_{(i+1)} \cdot 0$, we require for monotonicity

$$G_m[a_{(i)}] \leqslant G_m[x_{(ij)}] \leqslant G_m[a_{(i+1)} \cdot 0] \tag{6.22}$$

Subject to this constraint, the first term will be as large as possible and the other two as small as possible when they are all equal. Denoting the common value as P_i we have

$$G_m[a_{(i)}] = G_m[x_{(ij)}] = G_m[a_{(i+1)} \cdot 0] = P_i \tag{6.23}$$

with $P_0 = 0$ and $P_k = 1$.

Note that P_i is the probability of marrying by exact age $a_{(i)}$ conditional on marrying by exact age $a_{(k)}$ or $x_{(m)}$, as there are no marriages after age $a_{(k)}$.

The likelihood function (6.21) may then be written as

$$L = \prod_{i=1}^{k} [P_i \cdot P_{i-1}]^{m_i} [P_i]^{-t_i} \tag{6.24}$$

Let us now write

$$p_i = \frac{P_i}{P_{i+1}} , \quad i=1,...,k-1 \tag{6.25}$$

with $p_0 = 0$ and $p_k = 1$.

Note that p_i is the probability of marrying by exact age $a_{(i)}$ conditional on marrying by exact age $a_{(i+1)}$.

We can then write

$$P_i = \prod_{j=i}^{k} p_j \tag{6.26}$$

$$\text{and } P_i \cdot P_{i-1} = \prod_{j=i}^{k} p_j (1 - p_{i-1}).$$

The likelihood function (6.24) now becomes

$$L = \prod_{i=1}^{k} \left\{ \prod_{j=1}^{k} p_j (1 - p_{i-1}) \right\}^{m_i} \left\{ \prod_{j=1}^{k} p_j \right\}^{-t_i} \tag{6.27}$$

Collecting powers of p_i we obtain

$$L = \prod_{i=1}^{k} p_i^{\sum_{j=1}^{i} (m_j - t_j)} (1 - p_{i-1})^{m_i} \qquad (6.28)$$

The log-likelihood function is then

$$\log L = \sum_{i=1}^{k} \left\{ \sum_{j=1}^{i} (m_j - t_j) \log p_i + m_i \log(1 - p_{i-1}) \right\} \qquad (6.29)$$

Differentiating with respect to p_i for $i=1,...,k$ gives

$$\frac{\partial \log L}{\partial p_i} = \frac{\sum_{j=1}^{i} (m_j - t_j)}{p_i} - \frac{m_{i+1}}{1 - p_i} \qquad (6.30)$$

and setting the derivative to zero gives the m.l.e.

$$\hat{p}_i = \frac{M_i}{M_i + m_{i+1}}, \quad i=1,...,k-1 \qquad (6.31)$$

where M_i is as defined at (6.17).

FIGURE 6.2: Product-limit $[\bar{G}(a)]$ and fitted $[\hat{G}(a)]$ proportions ever-married among women married by age 29. 167, ever-married women sample.

Legend
– – – fitted $[\hat{G}(a)]$
——— product-limit $[\bar{G}(a)]$
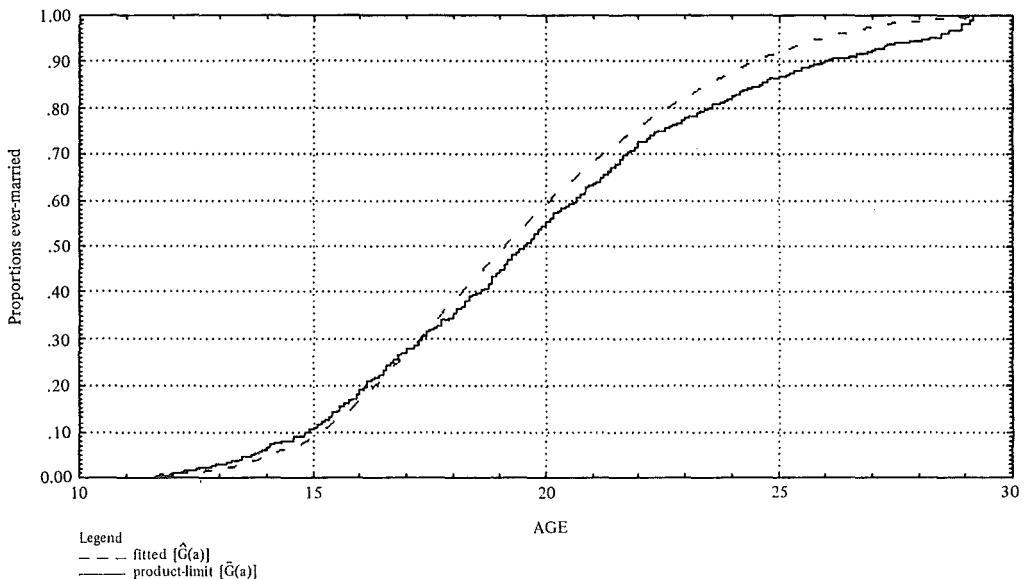
By the invariance property of m.l.e.'s and (6.26) we then obtain

$$\hat{p}_i = \prod_{j=1}^{k-1} \frac{M_i}{M_j + m_{j+1}} , \qquad i=1,\ldots,k-1 \qquad (6.32)$$

which in view of (6.23) is also the m.l.e. of $G_m[a_{(i)}]$. This step completes the derivation.

In the case of grouped data the product-limit estimate just developed reduces to the pooled estimate introduced in Section 3.4.

In order to assess the goodness of fit of the model to data truncated at age $x_{(m)}$ we can now compare the parametric estimate $\hat{G}(a|x_{(m)})$ of Section 6.4 with the non-parametric estimate $\bar{G}(a|x_{(m)})$, for all ages $a \leqslant x_{(m)}$.

A summary measure of the goodness of fit of the model is given by the statistic

$$D=\max\left[\max\left\{|\hat{G}[a_{(i)}|x_{(m)}]-\bar{G}[a_{(i)}|x_{(m)}]|\,,\,|\hat{G}[a_{(i)}|x_{(m)}]-\bar{G}[a_{(i-1)}|x(m)|\right\}\right] ,$$
$$(6.33)$$

which is a truncated-sample analog of the Kolmogorov-Smirnov statistic.

Figure 6.2 shows the two estimates $\hat{G}$ and $\bar{G}$ for the cohort aged 25-29 completed years in the Colombian individual survey. As the largest observed age at marriage is 29.167 both curves represent cumulative probabilities of marriage conditional on marrying by exact age 29.167.

The closeness of the two curves indicates a fairly good fit of Coale's model nuptiality schedule to the data. The largest difference between the two curves is D=.037. The same reservations stated earlier about the tendency of D to reflect an understatement of the goodness of fit since ages at marriage are confined to twelfths of a year apply in the ever-married sample as well. However, the statistic is obviously not affected by the proportion who can marry, since the sample is of ever-married women only.

7. FITTING THE MODEL TO FIRST BIRTH DATA

Recall that the standard probability density function which forms the basis of the model nuptiality schedule very closely approximates the convolution of a normal and three exponentials. Suppose that the risk of pregnancy leading to a live first birth were constant over time and across women. Then, the waiting time from initiation of intercourse to first birth would be distributed exponentially. If one regards marriage as the entry into the risk of exposure to pregnancy, then the above discussion would imply that a model first birth schedule could be constructed as the convolution of a normal and four exponentially distributed delays, or equivalently as the convolution of the age at first marriage and an exponential delay till the first birth. However, since Coale and McNeil found that a convolution of a normal and four exponential delays could be very closely approximated by a convolution of a normal and only three exponential delays, it follows that the marriage model should itself replicate first birth schedules adequately.

An initial analysis conducted by Trussell (Trussell, Menken and Coale, 1979) confirmed both that the marriage model fits first birth (and even second and third births) schedules well, and that the four parameter model (i.e. the marriage model and another parameter for the exponential delay) fits the data no better than the three parameter model. This analysis showed that period first birth schedules were replicated more closely than cohort schedules (at least for American data), because period effects appeared to be considerable. In recent extensions of this preliminary investigation in a Ph.D. Thesis, David Bloom (1980) has confirmed that the model does fit well when applied to data from a variety of countries and that period effects are indeed important.

The model has been fitted to the first birth data from the Colombia individual survey, and the parameter estimates are presented in Tables 7.1 and 7.2. In Table 7.1, the results for data on women who ever had a first birth are presented, while Table 7.2 extends the analysis by presenting estimates of the proportion ever having a first birth as well. Perusal of these tables indicates that the model does not fit the first birth data as well as the marriage data. This result could be due to the fact that first births (if we extrapolate from experience in other countries) appear to display more period effects than marriages, or could be attributed to errors in the dating of the first birth, or could be a consequence of genuine lack of fit of the model. Examination of the pooled estimates for each 5-year cohort reveals that the first birth schedules are very irregular, thus tending to lend heavier support to the first two explanations. Nevertheless, we are encouraged by these results, since poor overall fits are usually accompanied by a finding that the cohorts (20-24, 25-29, 35-39) are not homogeneous.

TABLE 7.1: Estimates of the parameters of the model fitted to grouped first birth data from the Colombia individual survey (1976). Women who had a first birth.

Ages	Estimates		Standard Errors		Goodness of Fit			Homogeneity of Cohorts		
(1) $x_0 - x_1$	(2) μ	(3) σ	(4) $s.e.\hat{\mu}$	(5) $s.e.\hat{\sigma}$	(6) χ_1^2	(7) ν	(8) p	(9) χ_1^2	(10) ν	(11) p
20–24	24.04	6.69	.904	.615	76.8	48	.005	58.1	38	.019
25–29	22.40	6.00	.375	.308	119.5	73	.000	94.1	58	.002
30–34	21.59	4.96	.247	.211	104.6	93	.194	82.0	74	.245
35–39	21.70	5.58	.263	.226	159.4	122	.013	134.2	98	.009
40–44	22.02	5.60	.271	.220	159.9	153	.335	116.1	122	.633
45–49	22.51	6.51	.321	.260	168.4	168	.476	112.0	136	.934

TABLE 7.2: Estimates of paramters of the model fitted to grouped first birth data from the Colombia individual survey (1976). All-women sample.

Ages	Estimates			Standard Errors			Goodness of Fit			Homogeneity of Cohorts		
(1) $x_0 - x_1$	(2) $\hat{\mu}$	(3) $\hat{\sigma}$	(4) $\hat{c}$	(5) $s.e.\hat{\mu}$	(6) $s.e.\hat{\sigma}$	(7) $s.e.\hat{c}$	(8) χ_j^2	(9) ν	(10) p	(11) χ_1^2	(12) ν	(13) p
20—24	23.49	6.36	1.06	.655	.468	.090	81.2	52	.006	59.6	42	.038
25—29	22.43	6.02	.925	.373	.307	.029	121.6	77	.001	96.2	62	.003
30—34	21.62	4.98	.936	.244	.209	.014	109.9	97	.176	87.2	78	.224
35—39	31.69	5.57	.899	.261	.221	.013	161.4	127	.021	136.1	102	.013
40—44	22.02	5.61	.909	.272	.220	.014	162.9	167	.358	119.6	126	.644
45—49	22.51	6.15	.924	.319	.263	.014	173.7	182	.657	117.3	146	.961

TABLE 7.3: Estimates of the average delay between first marriage and first birth in Colombia.

Cohort $x_0 - x_1$ (1)	c Not Estimated* (2)	c Estimated† (3)	Calculated Directly (4)
20—24	2.53	1.87	1.18
25—29	1.18	1.16	1.10
30—34	0.07	0.98	0.86
35—39	1.27	1.25	1.08
40—44	0.81	0.80	0.86
45—49	0.82	0.83	0.69

*Based on Tables 3.3 and 7.1.
†Based on Tables 4.1 and 7.2.

In Table 7.3 we present the implied average delay between first marriage and first birth — obtained by subtracting the estimated mean age at first birth from the estimated mean age at first marriage. For these results to be meaningfully interpretable, it must be the case that marriage is a true signal of initiation of exposure to the risk of childbearing. With the exception of the cohort 35-39 (which was already identified as an outlier), these results seem to indicate a lengthening over time of the delay between first marriage and first birth, a finding which is internally consistent with the raw data (shown in the fourth column of Table 7.3) and consistent with the observed fall in fertility.

Comparison with the raw data shows clearly that they are affected by mis-statement of date of birth of respondent or date of the respondent's first birth; the low values at ages 30-34 and 45-49 are clearly inconsistent with the other mean intervals. If there has been no change in age at marriage or age at first birth one would expect to see a declining trend (steeper at first) in the mean intervals calculated from the raw data due to the truncated nature of the data; women at older ages can, *ceteris paribus,* have longer intervals from marriage to first birth. Undoubtedly, this truncation partially explains why the estimate of the interval based on the model (which corrects for truncation) is higher. Although we could not recommend fitting the model to both sets of data in order to compute the mean dealy, we feel that the estimates based on this procedure are quite reasonable for Colombia.

8. COMPUTATIONAL CONSIDERATIONS

8.1 Optimization Procedures

Maximization of the log-likelihood function requires numerical techniques, since no analytical expressions are available for the m.l.e.'s. We employed two algorithms, the Davidon-Fletcher-Powell (DFP) method (Powell, 1971), and a quadratic hill climbing algorithm (GRADX) developed by Goldfeld and Quandt (1972).

The first algorithm, DFP (Powell, 1971), converged in almost every case, the only exceptions being for the age group 15-19 when the option to fix c was chosen. This algorithm is relatively fast and always converged to the same parameter estimates (when it converged). Furthermore, the estimates of the standard errors obtained from the inverse of the negative of the matrix of second partial derivatives (the inverse of the information matrix) seemed to be relatively stable. This finding was encouraging since DFP does not calculate second derivatives directly but builds up a matrix, initially the identity matrix, which eventually converges to the inverse of the information matrix if enough iterations occur. Bad estimates of the standard errors will result if the starting values are too close to the m.l.e., but one experience showed that starting values that differed by as little as .05 from the m.l.e. still gave very reasonable estimates of their standard errors.

The second algorithm GRADX (Goldfeld and Quandt, 1972) was used whenever DFP (rarely) failed. GRADX used alone proved to fail more often than DFP, though fortunately we never found a case where *both* failed to converge. GRADX is (about 20%) faster than DFP, but estimates of the standard errors proved to be unstable. GRADX employs directly the matrix of second derivatives, so estimates of standard errors are obtained even when the starting values are the m.l.e.'s.

We found that the likelihood function, though it appears in many cases to be flat near the maximum, was nevertheless easy to maximize. In no case did at least one algorithm fail to converge, even when the starting values were far from the m.l.e.'s.

The choice of starting values did not prove critical. In our work we used as default starting values $\mu=20$, $\sigma=6$ and $c=.9$. When fitting the model to the six 5-year cohorts in Colombia we used the default starting values for the cohort 20-24, and the final estimates of the previous cohort as starting values for each of the cohorts 25-29 to 45-49.

8.2 Evaluation of the Incomplete Gamma Function

The main problem we had in computing the function was discovering a way to compute the cumulative distribution function G(a). Recall that

$$G_0(z) = 1 - I(e^{-\lambda(z-\theta)}; a/\lambda - 1) = 1 - I(w,p) \tag{8.1}$$

where $w = e^{-\lambda(z-\theta)}$, $p = a/\lambda - 1$, z is the standardized aged $(x-\mu)/\sigma$, and $I(w,p)$ is the incomplete gamma function.

We experimented with several methods for evaluating the incomplete gamma function. We finally settled on an extremely fast version which involves creating a table of the values of $G_0(z)$ at regular intervals of z (of .005) and interpolating quadratically for values of z in between tabulated values. This procedure was modified slightly for very small or large values of z as will be explained below.

To calculate $I(w,p)$ we employed the well known series first derived by Pearson (1922):

$$I(w,p) = e^{-w} \sum_{j=0}^{\infty} \frac{w^{p+1+j}}{\Gamma_{(p+2+j)}} = \frac{e^{-w}w^{p+1}}{\Gamma_{(p+2)}} \left[1 + \frac{w}{(p+2)} + \frac{w^2}{(p+2)(p+3)} + \ldots \right]$$

$$= c(s_1 + s_2 + s_3 + \ldots) \tag{8.2}$$

58

In practice we considered the series to have converged when $c.s_i < 10^{-10}$. One problem with this series expansion is that the number of terms required for convergence becomes very large as z becomes small (w becomes large), as the following table shows:

z	-2.4	-2.0	-1.5	-1.0	-0.5	0	.5	1.0	1.5	2	2.5	3	3.5
No. of terms	54	35	21	15	11	8	7	6	5	4	4	4	3

This consideration is not so important if one wants to evaluate a table only once and inter-polate thereafter, but it is overwhelming if one wanted to calculate I(w,p) directly for each value of z.

The main problem for very small values of z, say as z becomes more negative than -2.37, is that the individual members (the s_i) of the series (both numerators and denominators and their ratios) become so huge and the constant $c (= e^{-w} w^{p+1} / \Gamma(p+2))$ becomes so small that all precision is lost from the computed answer.

Here we employed another approximation to the cumulative gamma function I(w,p), due to Gray, Thompson and McWilliams (1969),

$$G_0(z) = 1 - I(w,p) = \frac{w^{p+1} e^{-w}}{\Gamma(p+1)} \left[1 - \frac{p}{(w-p)^2 + 2w} \right] / (w-p) \qquad (8.3)$$

We found that the two approximations (8.2) and (8.3) could be joined when z=2.1. For very large values of z, say z above 1.9 we found that the series (8.2) could be used directly, as only 4 terms are needed for convergence. Hence, $G_0(z)$ was calculated by interpolation for values of z such that $-2.1 < z < 1.9$; the simple formula was used for $z \leqslant -2.1$; and the first four terms in the series were employed for $z \geqslant 1.9$. It should be noted that the same parameter estimates were obtained in extensive trials regardless of whether the expensive or cheap method of calculating $G_0(z)$ was used.

8.3 A Computer Program

All estimates in this paper were computed using the computer package NUPTIAL, which was written by the present authors. The numerical optimization routines are contained in a separate package developed by S.M. Goldfeld and R.E. Quandt. This package, which contains not only the algorithms GRADX and DFP but also several others, is available from the Econometric Research Program, Department of Economics, Princeton University, Princeton, N.J. 08544, USA.

The package NUPTIAL contains several options from which the user may choose, among which are

(a) maximize the likelihood function or minimize the sum of squared differences between the observed and fitted schedules,

(b) discard individual data on age at marriage for women marrying at their current age,

(c) fix the value of c, and estimate only μ and σ,

(d) use household data, individual data, or both, or an all-women sample,

(e) print data, observed and fitted values, and steps in the optimization,

(f) plot observed and fitted values.

This package, and the manual which accompanies it, are available from the World Fertility Survey, 35-37 Grosvenor Gardens, London. SW1.

REFERENCES

Asano, C. (1965). On estimating multinomial probabilities by pooling incomplete samples. *Annals of the Institute of Statistical Mathematics, Tokyo*—17: 1—13.

Bloom, David (1980). *Age Patterns of Women at First Birth,* unpublished Ph.D. thesis, Princeton University.

Coale, Ansley J. (1971). Age-patterns of marriage. *Population Studies* 25: 193—214.

Coale, Ansley J. (1977). The development of new models of nuptiality and fertility. *Population,* numéro spécial: 131—154.

Coale, Ansley J. and Donald R. McNeil (1972). The distribution by age at first marriage in a female cohort. *Journal of the American Statistical Association* 67: 743—749.

Goldfield, Stephen M. and Richard E. Quandt (1972). *Nonlinear methods in econometrics.* New York: North Holland.

Gray, H. L., R. W. Thompson, and G. V. McWilliams (1969). A new approximation for the chi-square integral. *Mathematics of Computation* 23: 85—89.

Hajnal, John (1953). Age at marriage and proportions marrying. *Population Studies* 7: 111—132.

Kaplan, E. L. and Paul Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53: 457—481.

Pearson, Karl (ed) (1922). *Tables of the incomplete Γ function.* London: HMSO (since 1934: Cambridge University Press).

Powell, M. J. D. (1971). Recent advances in unconstrained optimisation. *Mathematical Programming* 1: 26—57.

Trussell, James (1980). Age at first marriage in Sri Lanka and Thailand: An Illustrative Analysis. *World Fertility Survey, Scientific Reports Series* No. 13.

Trussell, James, Jane Menken and Ansley Coale (1979). A general model for analysing the effect of nuptiality on fertility. Presented at an IUSSP Conference on Nuptiality and Fertility, Bruges, Belgium.

Verma, Vijay, Christopher Scott, and Colm O'Muircheartaigh (1980). Sample Designs and Sampling Errors for the World Fertility Survey. *Journal of the Royal Statistical Society,* Series A, *143.*

GLOSSARY OF SYMBOLS

The following is a summary of the notation used in the paper. Symbols used only in a particular section are not included.

Data		Section Reference
x	age at interview	1.1/2.1/3.1/5.1
a	age at marriage	1.1/2.1/3.1/5.1
m_{ax}	number of women married at age a and now aged x	3.1/5.1
m_x	number of ever-married women aged x	2.1/3.1/5.1
s_x	number of single women aged x	2.1/5.1
n_x	total number of women aged x	2.1/5.1
P_x	proportion of ever-married women at age x	2.1
$P_{a|x}$	proportion of women married at age a among ever-married women aged x	3.2
pax	proportion of women married at age a among all women aged x	5.2

Model		
$f(a)$	frequency of first marriages	1.2
$F(a)$	cumulative frequency of first marriages	1.2
$g(a)$	probability density function (p.d.f.) of age at first marriage	1.2
$G(a)$	cumulative distribution function (c.d.f.) of age at first marriage	1.2
g_s, G_s	Swedish standard p.d.f. and c.d.f.	1.2
g_0, G_0	standard p.d.f. and c.d.f. with mean 0 and variance 1	1.3
c	proportion of women in a cohort who eventually marry	1.2
a_0, k	parameters of the standard nuptiality schedule	1.2
μ, σ	mean and standard deviation of age at marriage	1.3
z	standardized age $(x-a_0)/k$ or $(x-\mu)/\sigma$	1.2/1.3
$\pi_{a|x}$	probability of marrying at age a conditional on marrying by age x	3.2/4.3
π_{ax}	unconditional probability of marrying at age a for cohort aged x	5.2
Π_x	probability of being ever-married by age x	2.2/4.2/4.3

Estimates and Tests		
$\wedge$	denotes maximum likelihood estimates under the model	2.2/3.2/4.2/4.3/5.2/6.2/6.4
$-$	denotes pooled or product-limit estimates	3.4/5.4/6.3/6.5
χ_l^2	likelihood ratio chi-squared statistic	2.3/3.3/3.4/5.3/5.4
χ_p^2	Pearson's chi-squared statistic	2.3/3.3/3.4/5.3/5.4
ν	degrees of freedom	2.2/3.3/3.4/5.3/5.4

APPENDIX TABLES

TABLE 1: Number of ever-married and never-married women, by age, in the Colombia individual survey (1976).

Age (1)	Ever-married (2)	Never-Married (3)
15	7.	318.
16	26.	280.
17	37.	227.
18	71.	230.
19	74.	153.
20	117.	145.
21	102.	84.
22	124.	99.
23	138.	81.
24	108.	53.
25	127.	57.
26	137.	42.
27	121.	31.
28	146.	35.
29	123.	23.
30	129.	19.
31	87.	14.
32	109.	16.
33	100.	12.
34	106.	7.
35	110.	21.
36	119.	13.
37	101.	15.
38	89.	10.
39	89.	12.
40	120.	17.
41	77.	10.
42	83.	4.
43	76.	5.
44	78.	6.
45	95.	8.
46	71.	9.
47	77.	6.
48	65.	11.
49	61.	5.

TABLE 2: Age at marriage by age at interview for women in the Colombia individual survey (1976).

Age at Inter-view	Age at Marriage																																							
	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
15.	0.	0.	1.	1.	4.	1.																																		
16.	0.	0.	1.	2.	2.	6.	15.																																	
17.	1.	1.	1.	3.	6.	7.	13.	5.																																
18.	0.	0.	4.	3.	7.	14.	9.	25.	9.																															
19.	0.	1.	3.	2.	7.	9.	12.	13.	15.	12.																														
20.	0.	0.	3.	8.	7.	16.	15.	15.	20.	18.	15.																													
21.	0.	0.	1.	6.	11.	8.	11.	16.	13.	13.	17.	6.																												
22.	0.	2.	1.	5.	8.	11.	17.	15.	19.	18.	12.	10.	6.																											
23.	0.	3.	1.	1.	3.	9.	16.	17.	21.	21.	19.	14.	7.	6.																										
24.	0.	1.	1.	6.	1.	12.	10.	18.	13.	15.	10.	8.	2.	9.	2.																									
25.	0.	0.	2.	4.	8.	14.	14.	8.	15.	17.	13.	12.	1.	10.	8.	1.																								
26.	0.	1.	4.	4.	5.	10.	12.	10.	13.	19.	18.	8.	11.	8.	6.	7.	1.																							
27.	0.	1.	0.	4.	8.	7.	9.	15.	11.	9.	9.	12.	12.	4.	11.	6	1.	2.																						
28.	0.	1.	8.	6.	8.	13.	16.	13.	16.	10.	12.	15.	6.	7.	4.	3.	4.	2.	2.																					
29.	0.	1.	2.	3.	4.	8.	12.	7.	12.	16.	9.	11.	10.	5.	3.	4.	5.	4.	5.	2.																				
30.	0.	0.	1.	5.	2.	16.	13.	9.	16.	14.	11.	8.	7.	9.	9.	3.	4.	1.	1.	0.	0.																			
31.	0.	0.	0.	3.	2.	7.	10.	8.	8.	9.	5.	4.	10.	4.	7.	2.	4.	0.	1.	1.	0.	2.																		
32.	0.	1.	2.	0.	4.	8.	12.	11.	10.	12.	9.	9.	8.	2.	6.	0.	5.	0.	4.	2.	1.	3.	0.																	
33.	0.	0.	0.	0.	3.	6.	11.	12.	13.	6.	4.	9.	14.	4.	5.	3.	1.	4.	2.	0.	0.	0.	2.	1.																
34.	0.	1.	1.	0.	6.	11.	12.	9.	13.	6.	11.	7.	5.	7.	3.	2.	3.	2.	3.	1.	0.	1.	1.	1.	0.															
35.	0.	0.	1.	5.	4.	14.	12.	6.	13.	11.	7.	8.	6.	2.	3.	2.	4.	3.	2.	3.	1.	0.	1.	1.	0.	1.														
36.	0.	1.	2.	1.	4.	8.	6.	13.	10.	12.	14.	8.	6.	5.	2.	5.	6.	1.	1.	0.	2.	1.	2.	2.	5.	1.	1.													
37.	0.	1.	3.	0.	8.	18.	9.	11.	9.	9.	7.	3.	4.	6.	3.	2.	2.	0.	2.	0.	1.	0.	2.	0.	0.	1.	0.	0.												
38.	0.	0.	1.	4.	7.	6.	10.	11.	13.	7.	5.	3.	2.	4.	4.	2.	1.	2.	3.	0.	0.	0.	1.	1.	0.	0.	0.	2.	0.											
39.	1.	0.	1.	2.	3.	4.	9.	8.	9.	13.	7.	5.	3.	4.	3.	1.	3.	2.	1.	2.	2.	1.	2.	0.	2.	0.	0.	0.	0.	1.										
40.	0.	1.	1.	2.	6.	7.	7.	10.	15.	10.	7.	8.	8.	7.	8.	5.	4.	5.	1.	1.	2.	1.	1.	1.	0.	0.	1.	1.	0.	0.	0.									
41.	0.	0.	1.	0.	5.	4.	7.	6.	7.	9.	7.	9.	3.	6.	2.	1.	0.	2.	2.	1.	0.	0.	1.	0.	0.	2.	0.	0.	1.	1.	0.	0.								
42.	0.	2.	1.	2.	5.	4.	6.	9.	10.	5.	8.	7.	6.	0.	3.	3.	1.	4.	1.	1.	0.	2.	0.	1.	0.	0.	0.	1.	1.	0.	0.	0.								
43.	0.	0.	1.	3.	1.	3.	8.	8.	5.	5.	6.	7.	6.	2.	2.	7.	2.	3.	1.	0.	1.	1.	2.	1.	0.	0.	0.	1.	0.	0.	0.	0.	0.	0.						
44.	0.	0.	2.	2.	4.	5.	5.	5.	7.	7.	7.	7.	6.	2.	3.	7.	2.	2.	1.	2.	1.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	1.	0.	0.	0.		
45.	0.	1.	1.	2.	8.	3.	11.	6.	6.	11.	7.	9.	3.	4.	9.	2.	3.	1.	0.	1.	2.	1.	0.	1.	0.	1.	1.	0.	0.	0.	0.	1.	0.	0.	0.	0.				
46.	1.	0.	1.	1.	5.	5.	2.	2.	10.	8.	8.	7.	5.	3.	1.	4.	1.	1.	1.	2.	1.	1.	0.	0.	0.	1.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.			
47.	0.	1.	0.	1.	4.	3.	4.	4.	8.	4.	7.	4.	5.	6.	5.	5.	1.	1.	2.	2.	1.	0.	2.	3.	0.	2.	0.	0.	1.	0.	0.	1.	0.	0.	0.	0.	0.	0.		
48.	0.	0.	1.	1.	3.	2.	4.	7.	4.	8.	8.	2.	7.	0.	2.	3.	3.	0.	0.	4.	0.	2.	1.	0.	2.	0.	0.	0.	1.	0.	0.	0.	1.	0.	0.	0.	0.	0.	0.	
49.	0.	0.	1.	1.	1.	4.	2.	9.	6.	4.	4.	3.	6.	1.	3.	1.	2.	2.	4.	3.	1.	1.	0.	0.	0.	0.	0.	0.	0.	0.	0.	1.	0.	0.	1.	0.	0.	0.	0.	0.

63

TABLE 3: Summary of estimates of the model fitted to grouped marriage data from the Colombia National Fertility Survey (1976). Numbers in italics indicate results when all data are included. Numbers in roman type indicate results when data on age at marriage equal to the age at interview were omitted.

Age (1) $x_0 - x_1$	Sample (2)	Estimates (3) $\hat{\mu}$	(4) $\hat{\sigma}$	(5) $\hat{c}$	Standard Errors (6) $s.e.\hat{\mu}$	(7) $s.e.\hat{\sigma}$	(8) $s.e.\hat{c}$	p-Value (9)
20–24	I	21.507	5.938		.640	.479		.121
	I	*21.626*	*6.005*		*.566*	*.427*		*.143*
	B	21.798	6.135	.808	.524	.398	.046	.106
	B	*21.859*	*6.161*	*.813*	*.505*	*.389*	*.045*	*.129*
	A	21.620	6.012	.887	.609	.459	.064	.170
	A	*21.614*	*5.996*	*.891*	*.501*	*.381*	*.053*	*.215*
25–29	I	21.224	5.980		.362	.303		.292
	I	*21.176*	*5.946*		*.353*	*.300*		*.343*
	B	21.396	6.112	.838	.376	.314	.021	.284
	B	*21.337*	*6.070*	*.835*	*.366*	*.307*	*.021*	*.328*
	A	21.272	6.017	.910	.363	.304	.025	.376
	A	*21.250*	*6.003*	*.906*	*.352*	*.296*	*.024*	*.400*
30–34	I	20.623	5.003		.247	.212		.058
	I	*20.649*	*5.026*		*.245*	*.211*		*.042*
	B	20.697	5.068	.856	.250	.216	.012	.031
	B	*20.721*	*5.089*	*.856*	*.245*	*.211*	*.011*	*.022*
	A	20.643	5.021	.915	.238	.205	.014	.063
	A	*20.669*	*5.043*	*.917*	*.273*	*.232*	*.016*	*.058*
35–39	I	20.434	5.377		.251	.217		.188
	I	*20.510*	*5.448*		*.251*	*.218*		*.182*
	B	20.441	5.383	.846	.253	.213	.010	.143
	B	*20.517*	*5.455*	*.846*	*.254*	*.219*	*.009*	*.139*
	A	20.440	5.383	.885	.252	.217	.013	.233
	A	*20.516*	*5.453*	*.890*	*.252*	*.218*	*.012*	*.209*
40–44	I	21.207	5.740		.263	.226		.917
	I	*21.194*	*5.727*		*.280*	*.237*		*.929*
	B	21.232	5.763	.866	.265	.224	.011	.771
	B	*21.218*	*5.750*	*.866*	*.271*	*.234*	*.011*	*.794*
	A	21.219	5.752	.919	.270	.221	.013	.926
	A	*21.205*	*5.738*	*.919*	*.269*	*.224*	*.013*	*.955*
45–49	I	21.685	6.117		.320	.266		.669
	I	*21.677*	*6.109*		*.318*	*.264*		*.683*
	B	21.692	6.124	.851	.306	.254	.11	.636
	B	*21.684*	*6.116*	*.851*	*.305*	*.252*	*.011*	*.651*
	A	21.683	6.115	.908	.035	.252	.015	.783
	A	*21.675*	*6.108*	*.908*	*.304*	*.251*	*.015*	*.849*

Notes: I = Individual data on ever-married women only.
         B = Both household data and individual data.
         A = All-women sample.

TABLE 4: Summary of estimates of the model fitted to data on numbers of women single and ever-married by age at interview obtained from the Colombian National Fertility Survey (1976).

Age	Sample	Estimates			Standard Errors			p-Value
(1) $x_0 - x_1$	(2)	(3) $\hat{\mu}$	(4) $\hat{\sigma}$	(5) $\hat{c}$	(6) $s.e.\hat{\mu}$	(7) $s.e.\hat{\sigma}$	(8) $s.e.\hat{c}$	(9)
15–49	HH	22.439	5.284	.858	.146	.162	.006	.011
	I	21.922	4.976	.907	.193	.224	.008	.816
15–44	HH	22.489	5.334	.861	.160	.174	.007	.011
	I	21.928	4.983	.907	.206	.234	.009	.771
15–39	HH	22,437	5.281	.858	.167	.179	.009	.071
	I	21.842	4.896	.901	.220	.241	.012	.867
15–34	HH	22.612	5.442	.872	.230	.234	.015	.126
	I	22.080	5.122	.921	.276	.297	.020	.884
15–29	HH	22.138	5.022	.830	.290	.272	.023	.286
	I	21.622	4.706	.878	.367	.359	.030	.930
15–24	HH	21.791	4.738	.794	.539	.452	.057	.135
	I	21.034	4.219	.810	.592	.509	.068	.891

Notes: HH = Household survey.

        I   = Individual (all-women) survey.

TABLE 5: G(Z), proportion ever-married at exact age Z in the standard schedule with mean 0 and standard deviation 1.

Age Z	0	1	2	3	4	5	6	7	8	9
−1.9	.0001	.0001	.0001	.0001	.0001	.0000	.0000	.0000	.0000	.0000
−1.8	.0004	.0004	.0003	.0003	.0002	.0002	.0002	.0002	.0001	.0001
−1.7	.0014	.0012	.0011	.0010	.0009	.0008	.0007	.0006	.0005	.0005
−1.6	.0038	.0035	.0031	.0029	.0026	.0023	.0021	.0019	.0017	.0015
−1.5	.0088	.0082	.0075	.0070	.0064	.0059	.0054	.0050	.0045	.0042
−1.4	.0179	.0167	.0157	.0146	.0137	.0127	.0119	.0110	.0103	.0095
−1.3	.0323	.0306	.0289	.0273	.0258	.0243	.0229	.0216	.0203	.0190
−1.2	.0532	.0508	.0485	.0462	.0440	.0419	.0398	.0379	.0359	.0341
−1.1	.0810	.0779	.0749	.0719	.0690	.0662	.0635	.0608	.0582	.0557
−1.0	.1155	.1118	.1081	.1045	.1009	.0974	.0940	.0907	.0874	.0841
−.9	.1560	.1517	.1475	.1433	.1392	.1351	.1310	.1271	.1232	.1193
−.8	.2014	.1966	.1920	.1873	.1827	.1782	.1736	.1692	.1647	.1604
−.7	.2502	.2452	.2402	.2352	.2303	.2254	.2205	.2157	.2109	.2061
−.6	.3010	.2959	.2908	.2856	.2805	.2754	.2703	.2653	.2602	.2552
−.5	.3526	.3475	.3423	.3371	.3320	.3268	.3216	.3165	.3113	.3062
−.4	.4038	.3987	.3937	.3886	.3834	.3783	.3732	.3681	.3629	.3578
−.3	.4537	.4488	.4438	.4389	.4339	.4290	.4240	.4189	.4139	.4089
−.2	.5015	.4968	.4921	.4874	.4826	.4779	.4731	.4683	.4634	.4586
−.1	.5468	.5424	.5380	.5335	.5290	.5245	.5200	.5154	.5108	.5062
−.0	.5893	.5852	.5810	.5769	.5727	.5684	.5642	.5599	.5555	.5512
.0	.5893	.5852	.5810	.5769	.5727	.5684	.5642	.5599	.5555	.5512
.1	.6288	.6250	.6211	.6173	.6133	.6094	.6055	.6015	.5974	.5934
.2	.6652	.6617	.6582	.6546	.6510	.6474	.6437	.6400	.6363	.6326
.3	.6986	.6954	.6922	.6889	.6856	.6823	.6789	.6756	.6721	.6687
.4	.7292	.7262	.7233	.7203	.7173	.7143	.7112	.7081	.7050	.7018
.5	.7569	.7542	.7516	.7489	.7461	.7434	.7406	.7378	.7349	.7321
.6	.7820	.7796	.7772	.7748	.7723	.7698	.7673	.7647	.7621	.7595
.7	.8048	.8026	.8004	.7982	.7760	.7937	.7914	.7891	.7868	.7844
.8	.8252	.8233	.8213	.8193	.8173	.8153	.8132	.8111	.8090	.8069
.9	.8437	.8419	.8401	.8384	.8365	.8347	.8329	.8310	.8291	.8272
1.0	.8602	.8587	.8571	.8555	.8538	.8522	.8505	.8488	.8471	.8454
1.1	.8751	.8737	.8723	.8708	.8694	.8679	.8664	.8649	.8633	.8618
1.2	.8884	.8872	.8859	.8846	.8833	.8820	.8806	.8793	.8779	.8765
1.3	.9004	.8992	.8981	.8969	.8957	.8946	.8934	.8921	.8909	.8897
1.4	.9110	.9100	.9090	.9080	.9069	.9058	.9048	.9037	.9026	.9015
1.5	.9206	.9197	.9188	.9178	.9169	.9159	.9150	.9140	.9130	.9120
1.6	.9291	.9283	.9275	.9267	.9258	.9250	.9241	.9233	.9224	.9215
1.7	.9368	.9360	.9353	.9346	.9338	.9330	.9323	.9315	.9307	.9299
1.8	.9436	.9429	.9423	.9416	.9409	.9403	.9396	.9389	.9382	.9375
1.9	.9497	.9491	.9485	.9479	.9473	.9467	.9461	.9455	.9448	.9442
2.0	.9551	.9546	.9540	.9535	.9530	.9524	.9519	.9513	.9508	.9502
2.1	.9599	.9595	.9590	.9585	.9581	.9576	.9571	.9566	.9561	.9556
2.2	.9643	.9638	.9634	.9630	.9626	.9622	.9617	.9613	.9608	.9604
2.3	.9681	.9678	.9674	.9670	.9666	.9662	.9659	.9655	.9651	.9647
2.4	.9716	.9712	.9709	.9706	.9702	.9699	.9695	.9692	.9688	.9685
2.5	.9746	.9743	.9741	.9738	.9735	.9731	.9728	.9725	.9722	.9719
2.6	.9774	.9771	.9769	.9766	.9763	.9760	.9758	.9755	.9752	.9749
2.7	.9798	.9796	.9794	.9791	.9789	.9786	.9784	.9781	.9779	.9776
2.8	.9820	.9818	.9816	.9814	.9812	.9809	.9807	.9805	.9803	.9801
3.	.9857	.9872	.9886	.9899	.9909	.9919	.9928	.9936	.9943	.9949
4.	.9954	.9959	.9964	.9968	.9971	.9974	.9977	.9980	.9982	.9984
5.	.9986	.9987	.9988	.9990	.9991	.9992	.9993	.9994	.9994	.9995
6.	.9995	.9996	.9996	.9997	.9997	.9997	.9997	.9998	.9998	.9998
7.	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
8.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000