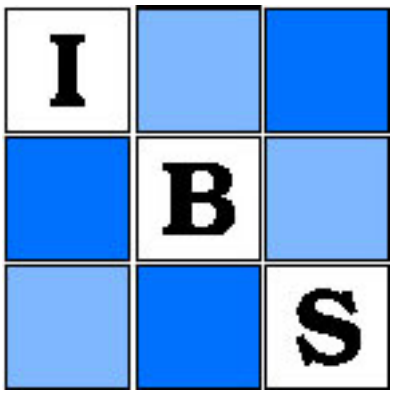

BIOMETRICS *32*, 159–172
March, 1976

ESTIMATING MEAN TIME TO REACH A MILESTONE, USING RETROSPECTIVE DATA

CORWIN L. ATWOOD[1] AND ADAM TAUBE[2]

Haile Selassie I University,[3] Addis Ababa, Ethiopia

SUMMARY

In surveys to estimate the mean age at menarche (or another milestone reached by the whole population), interviewed girls in the age range can respond that menarche (a) has not occurred or (b) has occurred or (c) occurred at a certain age t. Answers of type (a) and (b) are called *status quo data*. Answers of type (a) and (c) are called *retrospective data*. One kind of data is assumed. The distribution of age at menarche may also be assumed to be *normal* or *not necessarily normal*. This gives four possible sets of assumptions. Estimators, with their asymptotic distributions and optimal sampling allocations, are found for the case of retrospective data and non-normal distribution. These estimators are compared in examples with previously proposed estimators based on the other sets of assumptions. In these examples, retrospective data should certainly be used if available and reliable.

1. INTRODUCTION

The mean age at menarche has attracted considerable interest in the medical literature. For a list of references, see for example Bojlén and Bentzon [1968], Aw and Tye [1970] or Tekle Wold, Sterky and Taube [1972].

The data are usually of a cross sectional type: there is a set of samples of girls from various age groups, such that in the youngest age group none of the girls has experienced menarche while in the oldest group all the girls have passed their menarche. Interviewed girls may give several kinds of responses:

(a) *"Menarche has not yet occurred."*

(b) *"Menarche has occurred."*

(c) *"Menarche occurred when I was t years old."*

If the answers are all of type (a) and (b), call this *status quo data*. If the answers are all of type (a) and (c), call this *retrospective data*. This terminology can be found in medical literature. See e.g. Aw and Tye [1970]. In statistical literature, the first kind of data is also called *quantal data* while the second can be called *censored quantitative data*.

We do not consider the mixed case in which some girls give their age at menarche while others can only say that menarche has occurred.

There are two common approaches concerning the probability distribution of age at menarche: either assume (possibly after a transformation such as taking logarithms) that the distribution is *normal* or adopt a distribution free approach, assuming that the distribution is *not necessarily normal*. If the distribution is normal, the unknown parameters are

---

[1] Now with the Department of Mathematics, University of California at Davis, Davis, CA 95616.

[2] Now with the Department of Statistics, University of Uppsala, Uppsala, Sweden.

[3] Now the National University.

μ and σ. Otherwise the unknown parameters are p_i , defined as the probability that menarche occurs in the ith age group, for the various possible values of i.

Thus, the four situations listed below are commonly encountered. The fourth case, with retrospective data and no assumption of normality, will be the primary case of interest in this paper, although the estimators used for the other cases will be mentioned and compared. The cases and corresponding estimators to be considered in Section 2 are:

Status quo data, distribution normal

Probit analysis estimator

Status quo data, distribution not necessarily normal

Spearman–Kärber estimator

Retrospective data, distribution normal

Maximum likelihood estimator

Retrospective data, distribution not necessarily normal

"Retrospective" estimator
Maximum likelihood estimator

If the mean age at menarche is the only parameter of interest, the optimal sampling allocation (i.e., the proportion of girls of each age interviewed) can be found. The allocation will of course depend on the estimator being used. Since the optimal allocation also will depend on the actual values of the unknown parameters, it can at best only be approximated in practice. However the optimal allocation can be used to get a theoretical lower bound on the variance of the estimator in question.

Judicious consideration of the optimal sampling allocation is also useful for the practical experimenter. As an example, suppose that an experimenter believes that age at menarche is approximately normally distributed and that he can only get reliable status quo data. Thus he plans to use probit analysis. A glance at Fig. 1 and Section 2.3 reveals that the theoretical optimal allocation would be to interview only girls whose ages are very near μ, the mean age at menarche. He does not know μ but he is quite sure that it is between, say, 11 and 15 inclusive. He therefore decides to interview approximately equally many girls of each age from 11 through 15. This is a conservative, and not unreasonable, use of the theory. The allocation used is only an exceedingly rough approximation of the optimal allocation. However even this simple consideration has prevented the wasted effort of interviewing a lot of nine year-olds. It is unfortunate that experimenters seem in the past generally to have ignored such considerations. There are papers in which μ seems to be the only quantity of interest, where many girls were interviewed from whom very little information about μ could be expected to be obtained.

If the survey is of a multipurpose character, then a less specialized allocation must be used. In comparisons of the various estimators we will use both the optimal allocation for the estimator under consideration and the multipurpose allocation in which equal numbers of girls in each age group are interviewed.

The problem has been expressed in terms of mean age. It may be that an experimenter is less interested in the mean and more interested in the values of p_i , the probability that menarche occurs in the ith age class. If normality is assumed, then the p_i's can be estimated based on the estimates of μ and σ^2. If normality is not assumed, one must first

estimate the p_i's; then one may calculate the corresponding estimate of μ (and σ^2). Thus in either case estimates of all these unknown quantities can be found. In this paper the mean will be regarded as the quantity of chief interest because it is easier to compare estimators of a one-dimensional quantity μ than of a multidimensional quantity $(p_1, \cdots, p_k)$.

The numerical examples of Section 3 indicate that if retrospective data are available and reliable they should definitely be used. In particular the maximum likelihood estimators using retrospective data are considerably better in these examples than any of the other estimators considered.

Although the problem is phrased here in terms of menarche, the analysis of this paper applies equally well to any milestone which is reached by the whole population and for which retrospective data is available.

Note added in revision: Several related papers have come to our attention; all assume retrospective data and a distribution which is not necessarily normal. Kaplan and Meier [1958] find the MLE if the data are ungrouped. Peto [1973] uses a computer interation to find the MLE if the data are grouped in possibly overlapping classes with arbitrary possibly unequal widths. Elveback [1958] gives an estimator which is mathematically equivalent to our MLE of Section 2.6 although the parametrization and technique of solution are different.

2. ESTIMATION OF μ

2.1 Formulation of the Problem.

Let $\mathbf{T}$ be the time of menarche, a random variable with unknown cumulative distribution F, and mean $\mu = E(\mathbf{T})$. At this point no assumptions will be made about the form of F.

The time axis is divided into k classes (age groups), each of width h and with midpoint of the ith class at $a + ih$. Let p_i be the probability that menarche occurs in the ith age class. We will frequently use the phrase "at age i" to mean "in the ith age class."

For the purpose of later comparisons we formulate the problem using retrospective data and treat the status quo case as a special case in which part of the information is ignored. Suppose that n_i girls are interviewed from the ith age class and let $n = \sum n_i$, $i = 1, \cdots, k$. Each girl responds by giving her age at the time of her menarche or by saying that her menarche has not yet occurred. If her age is i, the probability that she will give j as her age at menarche is p_j, for $j < i$.

The probability that she will say her menarche occurred at age i, her present age, is more complicated. Her exact age at the time of interview, $\mathbf{A}$, may be considered as a random variable which is uniformly distributed between $a + (i - 1/2)h$ and $a + (i + 1/2)h$. The probability that she will say menarche occurred at age i is

$$P(a + (i - 1/2)h < \mathbf{T} < \mathbf{A}). \tag{2.1}$$

Every estimator considered in this paper makes some simplifying approximation here. If $\mathbf{T}$ is not assumed to be normal, (2.1) is approximated by $p_i/2$; this approximation would be exact if $\mathbf{T}$ were uniformly distributed in each age class. If $\mathbf{T}$ is assumed to be normal, (2.1) is approximated by $P(a + (i - 1/2)h < \mathbf{T} < a + ih)$. The inaccuracy introduced by these simplifications is not serious if h is small. Moreover, in the non-normal case the alternative to this approach would be to estimate (2.1) for each i; this would increase the number of unknown parameters from $k - 1$ to $2k - 1$, and not necessarily improve the

accuracy of $\hat{\mu}$. (For a discussion of the effect of the approximation in the probit analysis case, see Finney [1971, Sec. 10.7] and Tocher [1949].)

In some studies it is possible to learn each girl's exact age. For each girl this value can then be inserted in (2.1). In Section 2.6 the MLE for this case will also be given.

2.2 General Considerations.

Consider the non-normal case and approximate (2.1) by $p_i/2$. Let n_i girls of age i be interviewed. The probability of a set of responses is

$$p_1^{x_{i1}} \cdots p_{i-1}^{x_{i,i-1}} (p_i/2)^{x_{ii}} \bar{p}_i^{y_i} \tag{2.2}$$

where

$$\begin{aligned} x_{ij} &= \text{number of responses "menarche at age } j\text{"}, \\ y_i &= \text{number of responses "no menarche yet" and} \\ \bar{p}_i &= p_i/2 + p_{i+1} + \cdots + p_k . \end{aligned}$$

Multiplying expressions of the form (2.2) together for $i = 1, \cdots, k$ yields the probability of a set of responses from all n girls.

$$L = p_1^{x_{\cdot 1}} \cdots p_k^{x_{\cdot k}} (1/2)^{\sum x_{ii}} \bar{p}_1^{y_1} \cdots \bar{p}_k^{y_k} \tag{2.3}$$

where

$$x_j = \sum_{i=j}^{k} x_{ij} = \text{number of girls with menarche at age } j.$$

To avoid ambiguity when some $p_i = 0$, (2.3) should be understood only to include terms $p_i^{x_{\cdot i}}$ and $\bar{p}_i^{y_i}$ for which the exponent is nonzero.

Let us now consider the allocation question with retrospective data. If reliable retrospective data can be obtained then clearly the most information is obtainable from a girl who is interviewed after she has passed menarche. Thus the optimal allocation for estimating μ is to interview n girls who are all old enough to have passed menarche. If this is done then $y_i = 0$ for all i and the maximum likelihood estimator is easily found to be

$$\hat{\mu} = \sum (a + jh)\hat{\mathbf{p}}_j \quad \text{where} \quad \hat{\mathbf{p}}_j = \mathbf{x}_{ij}/n.$$

Here i only takes one value since the interviewed girls are all in one age group. Thus $\hat{\mu}$ is simply the sample mean of the responses based on grouped data. Sheppard's correction gives

$$\begin{aligned} E(\hat{\mu}) &\doteq E(\mathbf{T}) = \mu \\ V(\hat{\mu}) &\doteq n^{-1}[V(\mathbf{T}) + h^2/12]. \end{aligned}$$

All of the estimators of Sections 2.5 and 2.6 reduce to this estimator when this allocation is used, the optimal allocation for retrospective data.

Now let us consider the estimators and corresponding optimal sampling allocations for the four situations described in Section 1.

2.3 Status Quo Data, Distribution Normal.

The maximum likelihood estimator in this case is the probit analysis estimator $\hat{\mu}_P$ detailed in Finney [1971]. There is no explicit formula for $\hat{\mu}_P$. Rather it is found iteratively as the solution of the maximum likelihood equations. The asymptotic distribution of the MLE is well known (Cramér [1946]). Under conditions which hold here, it is asymptot-

ically normal and unbiased. In the notation of Section 2.1, the asymptotic covariance matrix of $n^{1/2}(\hat{\mathfrak{y}}_P , \hat{\vartheta}_P)$ is the inverse of

$$M = n^{-1}\begin{bmatrix} \sum n_i f_i^2[F_i(1 - F_i)]^{-1} & \sum n_i\sigma^{-1}(x_i - \mu)f_i^2[F_i(1 - F_i)]^{-1} \\ \sum n_i\sigma^{-1}(x_i - \mu)f_i^2[F_i(1 - F_i)]^{-1} & \sum n_i\sigma^{-2}(x_i - \mu)^2 f_i^2[F_i(1 - F_i)]^{-1} \end{bmatrix}$$

where $x_i = a + ih$, the midpoint of the ith class, $F_i = F(x_i)$ and $f_i = F'(x_i)$.

We wish to find an allocation which minimizes the asymptotic variance of $\hat{\mathfrak{y}}_P$, i .e., which minimizes the upper left entry of M^{-1}. This theoretically optimal allocation turns out to be suitable for practical use only in modified form as indicated in Section 1. We treat here only the case with an even number of cells symmetrically placed around μ. This is sufficient to get the entries of Table 1 in Example 2.

If k is even and the cells are symmetrically placed around μ, then any allocation can be symmetrized as follows. For any i, if there are n_i and n_{k+1-i} girls interviewed in cells i and $k + 1 - i$, let the symmetrized allocation take $(n_i + n_{k+1-i})/2$ observations in each of the two cells. If $n_i + n_{k+1-i}$ is odd, assign the last observation arbitrarily. Symmetrizing the allocation leaves the diagonal elements of M unchanged but it makes the off-diagonal elements 0. (If some $n_i + n_{k+1-i}$ are odd, then the off-diagonal elements are nearly 0 for n large). This decreases the diagonal elements of M^{-1}, as an elementary computation shows. Thus the upper left entry of M^{-1} is minimized if the design is symmetrical in which case the asymptotic variance of $\hat{\mathfrak{y}}_P$ is $(\sum n_i f_i^2[F_i(1 - F_i)]^{-1})^{-1}$. This is minimized by taking half of the interviewed girls each in the two age groups on opposite sides of μ where $f_i^2[F_i(1 - F_i)]^{-1}$ is maximized.

In Example 2 of Section 3, there are eight cells symmetrically placed around μ. The optimal allocation for $\hat{\mathfrak{y}}_P$ is to interview $n/2$ girls each from age groups four and five. The asymptotic variance is listed in Table 1.

2.4 Status Quo Data, Distribution Not Necessarily Normal.

The unknown parameters are $p_1 , \cdots , p_k$. Assume that $\sum p_i = 1$, i.e. the age classes sampled cover the entire age range in which menarche ever occurs. If only status quo data is available, the probability of a set of responses analogous to (2.3) is

$$L = \pi\,(1 - \bar{p}_i)^{n_i - \nu_i}\bar{p}_i^{\nu_i}$$

where as before $\bar{p}_i = p_i/2 + p_{i+1} + \cdots + p_k$.

From this it follows that the MLE of $\bar{p}_i$ is $\mathbf{y}_i/n_i$. The corresponding estimator of μ is

$$\hat{\mathfrak{y}}_{SK} = a + \tfrac{1}{2}h + h\sum \mathbf{y}_i/n_i ,$$

the "Spearman-Kärber estimator." (See Finney [1964]). Since the $\mathbf{y}_i$ are independent binomial with parameters n_i and $\bar{p}_i$, it is immediate that

$$E(\hat{\mathfrak{y}}_{SK}) = a + \tfrac{1}{2}h + h\sum_1^k \bar{p}_i \tag{2.4}$$

so that

$$E(\hat{\mathfrak{y}}_{SK}) \doteq \mu.$$

The equality is not exact because the ages were grouped. Similarly

$$V(\hat{\mathfrak{y}}_{SK}) = h^2\sum_1^k \bar{p}_i(1 - \bar{p}_i)/n_i . \tag{2.5}$$

Use of Lagrange multipliers shows that the choice of n_i's which minimizes $V(\hat{\mathfrak{y}}_{SK})$ is

$$n_i = n[\bar{p}_i(1 - \bar{p}_i)]^{1/2} \Big/ \sum_1^k [\bar{p}_i(1 - \bar{p}_i)]^{1/2}. \tag{2.6}$$

Of course the $\bar{p}_i$'s are unknown so the optimal n_i can only be approximated, based on any available prior information about the distribution. However this does not seem difficult when dealing with age at menarche. Note that since $\bar{p}_i(1 - \bar{p}_i)$ is largest when $\bar{p}_i = 1/2$, the value of n_i should be largest when i is somewhere in the middle between 1 and k. Substitution of (2.6) in (2.5) yields a lower bound on $V(\hat{\mathfrak{y}}_{SK})$:

$$V(\hat{\mathfrak{y}}_{SK}) \geq h^2 n^{-1}\left\{\sum_1^k [\bar{p}_i(1 - \bar{p}_i)]^{1/2}\right\}^2$$

with equality attained if and only if the values of n_i are given by (2.6).

Suppose now that girls are interviewed in too few age classes, say in classes 1 to k, when menarche can also occur (although with small probability) in classes 0 and $k + 1$. In this case, formulas (2.4) and (2.5) are still correct, where k is the number of age groups sampled, and now $\bar{p}_i = p_i/2 + p_{i+1} + \cdots + p_{k+1}$. Thus the 0 and $k + 1$ terms are missing from the summation in (2.4) and a bias is introduced (which is small if p_0 and p_{k+1} are small). On the other hand the variance (2.5) is smaller than it would be if girls in more age groups had been interviewed. To avoid this complication the experimenter may wish to take a few observations outside the range where he believes menarche occurs or he may decide that the complication does little damage and he can live with it.

2.5 Retrospective Data, Distribution Normal.

Swan [1969] assumes that one can observe an upper and lower bound on the normal random variable. Our situation here is thus a special case of his; indeed, so is Section 2.3. Swan then finds the maximum likelihood equations for μ and σ which can be solved numerically by the Newton–Raphson iterative method. If the underlying assumption of normality is correct, then the estimator is asymptotically normal with asymptotic mean μ. The asymptotic covariance matrix of the estimator of (μ, σ) can be worked out in a routine way, although it is rather cumbersome to write down.

A numerical example is presented in Table 1.

The optimal allocation for this estimator, or any estimator using retrospective data, was discussed at the end of Section 2.2.

2.6 Retrospective Data, Distribution Not Necessarily Normal.

Two estimators will be considered for this case, the "retrospective estimator" introduced in Tekle Wold *et al.* [1972] and the maximum likelihood estimator (MLE).

Retrospective Estimator.

The retrospective estimator is easily motivated. The estimator of p_j is taken to be a multiple of $x_{\cdot j}$ where $x_{\cdot j}$ is the number of girls (of all ages $\geq j$) who had menarche at age j. To make the estimates unbiased, set

$$\hat{\mathbf{p}}_j = c_j \mathbf{x}_{\cdot j}$$

where

$$c_j^{-1} = n_j/2 + \sum_{j+1}^k n_i .$$

These $\hat{\mathbf{p}}_i$'s do not necessarily sum to 1. This will be corrected later but for now define the *uncorrected retrospective estimator* as

$$\hat{\mathfrak{y}}_{UR} = \sum (a + jh)\hat{\mathbf{p}}_i .$$

Since the $\hat{\mathbf{p}}_i$'s are unbiased, $E(\hat{\mathfrak{y}}_{UR}) = \sum (a + jh)p_i \doteq \mu$. Unfortunately the variance computations are not so simple because the $\hat{\mathbf{p}}_i$'s are correlated. The derivation of $V(\hat{\mathfrak{y}}_{UR})$ is only sketched since in the examples of Section 3 $\hat{\mathfrak{y}}_{UR}$ is definitely poorer than the MLE. We have

$$V(\ _{UR}) = V\Big[\sum_i (a\hat{\mathfrak{y}} + jh) \sum_{i=j}^{k} \mathbf{x}_{ii}c_i\Big]$$

$$= \sum_i V\Big[\sum_{j=1}^{i} (a + jh)\mathbf{x}_{ii}c_i\Big].$$

Expand the sum over j in terms of the variances and covariances of the $\mathbf{x}_{ij}$'s and get

$$V(\hat{\mathfrak{y}}_{UR}) = a^2A + 2aB + C \tag{2.7}$$

where

$$A = \sum_i [\sum_{j\le i} c_i{}^2V(\mathbf{x}_{ij}) + \sum_{j\neq l\le i}\sum c_ic_l \operatorname{cov}(\mathbf{x}_{ij} , \mathbf{x}_{il})]$$

and B and C are similar expressions. Note here the surprising fact that $V(\hat{\mathfrak{y}}_{UR})$ depends on the location parameter a.

The terms A, B and C can be evaluated using the fact that $(\mathbf{x}_{i1} , \cdots , \mathbf{x}_{ii} , \mathbf{y}_i)$ has a multinomial distribution. After some manipulation in the case $n_i = n/k$ we obtain

$$A = n^{-1}k \sum_i [\sum d_i{}^2p_i - (\sum d_ip_i)^2 + \tfrac{1}{2}d_i{}^2(1 - \tfrac{1}{2}p_i)p_i - \sum d_id_ip_ip_i]$$

$$B = n^{-1}kh \sum_i [\sum jd_i{}^2p_i - (\sum d_ip_i)(\sum jd_ip_i) + \tfrac{1}{2}d_i{}^2i(1 - \tfrac{1}{2}p_i)p_i - \tfrac{1}{2} \sum d_id_i(i + j)p_ip_i]$$

$$C = n^{-1}kh^2 \sum_i [\sum j^2d_i{}^2p_i - (\sum jd_ip_i)^2 + \tfrac{1}{2}d_i{}^2i^2(1 - \tfrac{1}{2}p_i)p_i - \sum d_id_ijipjp_i]$$

where $d_i = k - j + 1/2$ and all the summations within the brackets are over $j < i$. This result will be used in the numerical comparisons of Section 3.

To correct for the fact that $\sum \hat{\mathbf{p}}_i$ is not necessarily 1, set $\hat{\mathbf{p}}_i{}^* = \mathbf{p}_i/\sum_i \hat{\mathbf{p}}_i$ and define the *corrected retrospective estimator*

$$\hat{\mathfrak{y}}_{CR} = \sum (a + jh)\mathbf{p}_i{}^*.$$

Exact results are difficult because $\mathbf{p}_i{}^*$ is a quotient of correlated random variables. However the asymptotic distribution of $\hat{\mathfrak{y}}_{CR}$ can be found as follows.

Since $(\hat{\mathbf{p}}_1 , \cdots , \hat{\mathbf{p}}_k)$ can be expressed as a sample mean, it is asymptotically multivariate normal and therefore $\hat{\mathfrak{y}}_{UR}$ is also asymptotically normal. Since $\sum \hat{\mathbf{p}}_i \to 1$ in probability, it follows (by Rao [1965], Theorem 2.c.4.x.) that

$$\hat{\mathfrak{y}}_{UR} - \mu$$

and

$$\frac{\hat{\mathfrak{y}}_{UR} - \mu}{\sum \hat{\mathbf{p}}_i} = \hat{\mathfrak{y}}_{CR} - \mu + \mu[1 - 1/\sum \hat{\mathbf{p}}_i]$$

have the same asymptotic distribution. If $\mu = 0$, i.e. if a is chosen to equal $-h \sum ip_i$, then the expression on the right of the equality reduces to $\hat{\mathfrak{y}}_{CR} - \mu$. Therefore $n^{1/2}(\hat{\mathfrak{y}}_{CR} - \mu) \to N(0, \sigma^2)$ in distribution where σ^2 is expression (2.7) with $a = -h \sum ip_i$. This forms the basis for part of Table 1 in Section 3.

Maximum Likelihood Estimator.

We now derive the MLE. Those who only wish to use the estimator will find a summary of the procedure for calculating it given at the end of the derivation.

We want to maximize (2.3), or equivalently to maximize

$$\log L = c + \sum_1^{k-1} x_{\cdot i} \log p_i + \sum_1^{k-1} y_i \log \bar{p}_i + x_{\cdot k} \log p_k + y_k \log \bar{p}_k \tag{2.8}$$

where c is a term which does not depend on the p_i's. To avoid ambiguity when $p_i = 0$, no terms are included in which the logarithm is multiplied by zero. There are $k - 1$ parameters to be estimated. Let p_k be the dependent parameter by setting $p_k = 1 - \sum_1^{k-1} p_i$. Differentiating (2.8) with respect to p_i, $i = 1, \cdots, k - 1$ and setting the result equal to zero gives

$$0 = x_{\cdot i}/p_i - y_i/2\bar{p}_i - \sum_{i+1}^{k-1} (y_i/\bar{p}_i) - (x_{\cdot k} + y_k)/p_k \tag{2.9}$$

for $i = 1, \cdots, k - 1$. The summation vanishes if $i = k - 1$. Note that if $p_1, \cdots, p_k$ are any nonnegative numbers which satisfy (2.9), then $cp_1, \cdots, cp_k$ also satisfy (2.9) for any constant $c > 0$. (Here

$$\bar{p}_i = \tfrac{1}{2}p_i + p_{i+1} + \cdots + p_k \tag{2.10}$$

not $1 - p_1 - \cdots - p_{i-1} - \tfrac{1}{2}p_i$!). Therefore the method of solution will be to find some $p_1, \cdots, p_k$ which satisfy the $k - 1$ equations (2.9) and then to multiply them by the appropriate c so that $\sum cp_i = 1$.

We will find the values of p_i one at a time, beginning at the top by choosing an arbitrary $p_k > 0$. If at a certain point we have found $p_k, \cdots, p_{i+1}$, then (2.9) gives

$$Ap_i^2 + (AB + y_i - x_{\cdot i})p_i - Bx_{\cdot i} = 0 \tag{2.11}$$

where A and B are defined by

$$\begin{aligned} A &= \sum_{i+1}^{k-1} (y_i/\bar{p}_i) + (x_{\cdot k} + y_k)/p_k \\ B &= 2(p_{i+1} + \cdots + p_k). \end{aligned} \tag{2.12}$$

If $x_{\cdot k} + y_k > 0$, then A is nonzero and equation (2.11) has a unique nonnegative solution (since $ABx_{\cdot i} \geq 0$) which can be found by using the quadratic formula.

If $x_{\cdot k} + y_k = 0$, then for some i we may have $A = 0$ and perhaps many solutions or no solution for p_i. As a fairly general example, suppose $x_{\cdot k} = x_{\cdot k-1} = y_k = y_{k-1} = 0$ and $x_{\cdot k-2} > 0$. When $i = k - 1$ then (2.11) reduces to $0 = 0$. Thus p_{k-1} is not determined. Essentially, the data cannot distinguish between the classes $k - 1$ and k but treat them as a single class. Arbitrary values must initially be assigned to both p_k and p_{k-1}; in particular they may be set equal, or either may be made zero based on considerations other than the data. It will eventually be clear that $\hat{p}_{k-1} + \hat{p}_k$ is determined, but if this sum is nonzero then the relative weights of the two components are not determined.

Having assigned a value to p_{k-1}, set $i = k - 2$ and try to solve (2.11) for p_{k-2}. Again

$A = 0$, and (2.11) becomes

$$(y_{k-2} - x_{.k-2})p_{k-2} = 2(p_{k-1} + p_k)x_{.k-2} .$$

If $y_{k-2} > x_{.k-2}$ then there is a unique solution and no further complication in finding the remaining p_i's. If $y_{k-2} \leq x_{.k-2}$, then there is no nonnegative solution for p_{k-2}, *i.e.*, the derivative of (2.8) is positive for all positive p_{k-2}. Therefore the maximum likelihood solution cannot have $p_{k-1} + p_k > 0$. Thus the upper two class probabilities should be set equal to 0 and the number of classes under consideration reduced by two. A simple way is to redefine k, setting it equal to the former $k - 2$. With this redefined k, begin by letting $p_k > 0$ be arbitrary, then solve (2.11) successively for $p_{k-1}, \cdots, p_1$ with no complications since now $x_{.k} > 0$.

So in summary to find the MLE, let $p_k > 0$ be arbitrary. For $i = k - 1, \cdots, 1$, successively define A and B by (2.12) and (2.10) and let p_i be the nonnegative solution of (2.11). If for any i, the solution p_i is not unique, then a nonnegative value must be assigned arbitrarily. If for any i there is no nonnegative solution p_i, then redefine k equal to this i, assign all the higher classes probability zero and return to the beginning of this paragraph with the redefined k. Finally for $i = 1, \cdots, k$ set

$$\hat{p}_i = p_i \Big/ \sum_1^k p_i .$$

The usual asymptotic theory for the MLE applies here. The $\hat{\mathbf{p}}_i$'s which correspond to nonzero p_i's are asymptotically normal, unbiased and minimum variance. If k now denotes the number of nonzero p_i's, the asymptotic covariance matrix of $(\hat{\mathbf{p}}_1, \cdots, \hat{\mathbf{p}}_{k-1})$ is the inverse of the symmetric matrix M with

$$M_{ii} = \bar{n}_i/p_i + \sum_{i+1}^{k-1} n_l/\bar{p}_l + n_i/4\bar{p}_i + n_k/p_k$$

where $\bar{n}_i$ is $\frac{1}{2}n_i$ plus the total number of girls interviewed at ages $> i$ and

$$M_{ij} = \sum_{i+1}^{k-1} n_l/\bar{p}_l + n_i/2\bar{p}_i + n_k/p_k$$

for $j < i$.

The corresponding estimator of μ is $\sum_1^k (a + jh)\hat{\mathbf{p}}_j$ which is an asymptotically unbiased and asymptotically efficient estimator of $\sum_1^k (a + jh)p_j \doteq \mu$. Since $\hat{\mathfrak{y}} = a + kh - h\sum_1^{k-1} (k - j)\hat{\mathbf{p}}_j$, the asymptotic variance of $\hat{\mathfrak{y}}$ is $h^2b'Vb$, where V is the asymptotic covariance matrix of $(\hat{\mathbf{p}}_1, \cdots, \hat{\mathbf{p}}_{k-1})$, b' is the row vector $(k - 1, \cdots, 1)$ and b is the corresponding column vector.

MLE If the Girls' Ages Are Known Exactly.

As mentioned at the end of Section 2.1, each girl's age may be known exactly. In this case, the probability that a girl of age i will say that her menarche occurred at age i is given by (2.1) with **A** replaced by her exact age at the time of interview. This can be approximated by $p_i e_{il}$ where e_{il} is the time that the lth girl has been in the ith age class divided by the class width. The likelihood equations to be solved are then

$$0 = x_{.i}/p_i - \sum_{l=1}^{n_i} y_{il}e_{il} \Big/ \left[\sum_{i+1}^{k} p_i + p_i(1 - e_{il})\right] - \sum_{j=i+1}^{k-1} \sum_{l=1}^{n_j} y_{jl} \Big/ \left[\sum_{i+1}^{k} p_i + p_i(1 - e_{il})\right] - (x_{.k} + y_k)/p_k \qquad (2.13)$$

for $i = 1, \cdots, k-1$, where $y_{il} = 1$ if the lth girl of age i has not had menarche and $y_{il} = 0$ otherwise. The method of solution is similar to that for (2.9): choose $p_k > 0$ arbitrarily, solve successively for $p_{k-1}, \cdots, p_1$ and then set $\hat{p}_i = p_i / \sum p_i$. The difference is that each p_i must be found numerically using a computer rather than as the solution of a quadratic equation.

3. NUMERICAL EXPENSES

3.1 Estimation Methods.

As described in Tekle Wold *et al.* [1972], Ethiopian girls from ages nine to 17 were interviewed. The girls were classed by year of age so the class width is $h = 1$. The MLE of the p_i's corresponding to ages 12 through 16 are .0940, .3094, .4280, .1255 and .0431, respectively; the $\hat{p}_i$'s for other ages are 0.

As Example 1, suppose that these are the true class probabilities and that the distribution is uniform within each class so that (2.1) equals $\frac{1}{2}p_i$. We consider the estimators which do not assume normality and three sampling allocations: interview $n/9$ girls from each of the ages nine through 17 (as was actually done), interview $n/5$ girls from each of the ages 12 through 16 and for the estimator being used interview girls according to the best allocation for that estimator. We will call these the "nine cell allocation," the "five cell allocation" and the "optimal allocation for the estimator." (The optimal allocations were described at the ends of Sections 2.2, 2.3 and 2.4). Estimators which assume a normal

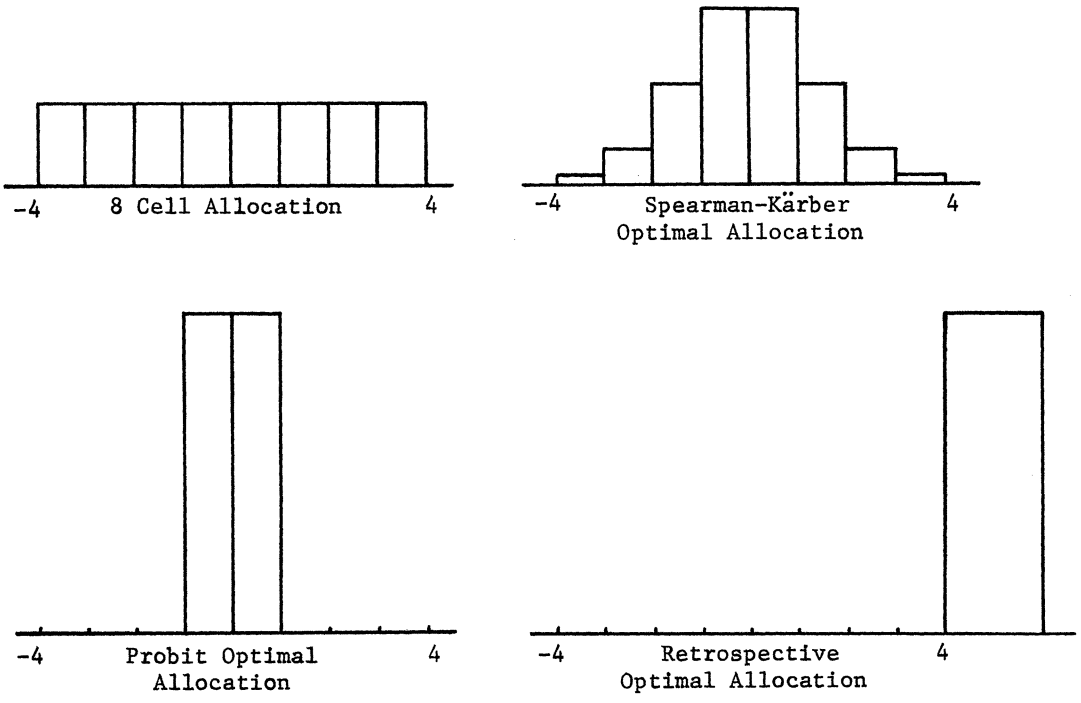

FIGURE 1

SAMPLING ALLOCATIONS USED FOR EXAMPLE 2

TABLE 1
$nV(\hat{\mu})$

	Example 1			Example 2	
	9 cell allocation	5 cell allocation	optimal allocation	8 cell allocation	optimal allocation
Probit analysis	-	-	-	$\doteq$ 4.43	$\doteq$ 1.72
Spearman-Kärber	5.65	3.14	2.80	4.70	2.80
Retrospective data, distr. normal, MLE	-	-	-	$\doteq$ 1.69	1.08
Uncorrected retrospective	$\geq$ 2.14	$\geq$ 1.79	.90	$\geq$ 2.14	1.08
Corrected retrospective	$\doteq$ 2.80	$\doteq$ 3.37	.90	$\doteq$ 2.97	1.08
Retrospective data, distr. not normal, MLE	$\doteq$ 1.93	$\doteq$ 1.45	.90	$\doteq$ 1.89	1.08

distribution are not considered because it is difficult to find the asymptotic distribution of the MLE when the assumptions on which the MLE was based do not hold.

As Example 2, suppose that **T** is Normal(0, 1) and the time axis is divided into eight cells of width 1, from -4 to 4. We consider all the estimators mentioned in Section 2 and two allocations: the "eight cell allocation," in which $n/8$ girls are interviewed from each of the eight classes and the optimal allocation for the estimator being used. All the allocations used in this example are illustrated schematically in Fig. 1.

Table 1 shows $nV(\hat{\mu})$ for the various estimators and allocations considered. The symbol $\doteq$ indicates an asymptotic result, approximately correct for large n. The figure in the "uncorrected retrospective" row is the minimum possible value, attained for the proper choice of the location parameter. Note that all of the estimators which use retrospective data coincide when the optimal allocation for these estimators is used.

In Example 1, it was assumed that expression (2.1) equals $\frac{1}{2}p_i$ so no approximations are involved in Table 1 except for the usual approximation involved in any asymptotic result.

In Example 2, the tabulated variances obtained by large sample theory for the MLE's are the asymptotic values which would be correct if (2.1) were exactly equal to its approximation. To give an idea of the inaccuracy introduced by this approximation, the variances of the retrospective and the Spearman-Kärber estimators were computed exactly and also

TABLE 2

COMPARISON OF ESTIMATES

Method	$\hat{\mu}$	95% Confidence Limits
Probit analysis	13.58	13.34, 13.82
Spearman-Kärber	13.60	13.36, 13.84
Retrospective data, distr. not norm., MLE	13.71	13.57, 13.85

by using the simplification for (2.1). Some of the approximate variances then differed from the exact variances by as much as .2 or .3. This suggests that the tabulated asymptotic variances of the MLE's in Example 2 may be inaccurate by .2 or .3 units.

Study of the table shows that the MLE using retrospective data are substantially better than any of the other estimators in these examples. Also in Example 2, use of the normality of the distribution results in a modest decrease in $V(\hat{\mu})$.

3.2 Values of the Estimates.

The methods of estimation have been compared above. We now compare actual values of some of the estimates for real menarche data. This does not show which estimators are most efficient; for that comparison one should look at Table 1. Rather, our one purpose for looking at the estimates is that something might be learned about which assumptions hold for menarche data. Specifically, if the MLE's of Section 2.6 differed greatly from the Spearman-Kärber estimate, then we would question the reliability of the retrospective data. If the probit analysis estimate differed greatly from the Spearman-Kärber estimate, then we would question the normality of the distribution. If there were any large discrepancy among the estimates we would also be concerned about the effect of grouping the data into one-year age classes.

Using the data reported in Tekle Wold, Sterky and Taube [1972], we get Table 2. As can be seen from a comparison of the confidence limits, the estimates are quite consistent with each other and this example does not provide grounds for questioning any of the assumptions made in this paper.

For further comparisons of estimates, based on not entirely comparable sets of data, the reader is referred to Aw and Tye [1970].

ACKNOWLEDGMENT

The authors thank the referees and the associate editor for looking at the paper so carefully and for several very helpful suggestions.

ESTIMATION DU TEMPS D'ATTENTE D'UNE PHASE DE LA VIE, UTILISANT DES DONNÉES RETROSPECTIVES

RÉSUMÉ

Dand les enquête que l'on fait pour estimer l'âge moyen de la menarche (ou toute autre phase atteinte par toute la population), les jeunes filles de cet âge interrogées peuvent répondre que la ménarche

a) né s'est pas encore produite ou

b) s'est produite ou

c) s'est produite à un certain âge t.

Des réponse du type a) et b) sont appelées des *données de statut* (status quo data). Des réponses de type a) et c) sont appelées *données rétrospectives* On suppose que les données sont de l'un de ces deux types, la distribution de l'âge de la menarche peut aussi être supposée *normale* ou non *nécessairement normale.* Cela donne quatre ensembles d'hypothèses possibles. On trouve, dans le cas de connées rétrospectives et de distribution non-normale, des estimateurs, leurs distributions asymptotiques et leurs positions d'échantillonnage optimales. Ces estimateurs sont comparés sur des exemples avec des estimateurs précédemment proposés sur la base des autres ensembles d'hypothèses. D'après ces exemples, on devrait certainement utiliser les données rétrospectives lorsqu'elles sont disponibles et utilisables.

REFERENCES

Aw, E. and Tye, C. Y. [1970]. Age of menarche of a group of Singapore girls. *Human Biol. 42*, 329–36.

Bojlén, K. and Bentzon, M. W., [1968]. The influence of climate and nutrition on age at menarche: a historical review and a modern hypothesis. *Human Biol. 40*, 69–85.

Cramér, H. [1946]. *Mathematical Methods of Statistics.* Princeton Univ. Press, Princeton.

Elveback, L. [1958]. Estimation of survivorship in chronic disease: the "actuarial" method. *J. Amer. Statist. Assoc. 53*, 420–40.

Finney, D. J. [1964]. *Statistical Method in Biological Assay* (2nd ed.), Griffin, London.

Finney, D. J. [1971]. *Probit Analysis*, (3rd ed.), Cambridge Univ. Press, Cambridge.

Kaplan, E. L. and Meier, P., [1958], Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc. 53*, 457–81.

Peto, R., [1973]. Experimental survival curves for interval-censored data. *Applied Statist. 22*, 86–91.

Rao, C. R., [1965]. *Linear Statistical Inference and its Applications.* Wiley, New York.

Swan, A., [1969]. Computing maximum likelihood estimates for parameters of the normal distribution from grouped and censored data. *Applied Statist. 18*, 65–9.

Tekle Wold, F., Sterky, G. and Taube, A., [1972]. The age of menarche in a group of school girls in Addis Ababa, *Eth. Med. J. 10*, 159–66.

Tocher, K. D., [1949]. A note on the analysis of grouped probit data. *Biometrika 36*, 9–17.

Received November 1973, *Revised June* 1975

Key Words: Menarche, Probt analysis, Spearman-Kärber estimator, Retrospective data, Censored data, Quantitative data with censoring.

APPENDIX

A summary of the data reported in Tekle Wold, Sterky and Taube [1972] is given here. "Age i" means that the true age is between i-1/2 and $i + 1/2$. In the notation of the present paper $n_i = 40$ for i= 9,..., 17 and

$$x_{.12} = 21 \qquad y_{12} = 38$$

$$x_{.13} = 57 \qquad y_{13} = 29$$

$$x_{.14} = 62 \qquad y_{14} = 11$$

$$x_{.15} = 14 \qquad y_{15} = 4$$

$$x_{.16} = 2 \qquad y_{16} = 2$$

$$y_{17} = 0.$$

All other values of $x_{.i}$ are 0; all other values of y_i are 1.