Niels Keiding Steffen Lauritzen

Maximum Likelihood Estimation of the Offspring Mean in a Simple Branching Process

Preprint
May
1977

4

Institute of Mathematical Statistics
University of Copenhagen

Niels Keiding and Steffen L. Lauritzen

MAXIMUM LIKELIHOOD ESTIMATION OF THE OFFSPRING MEAN IN A SIMPLE BRANCHING PROCESS

Preprint 1977 No. 4

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

May 1977

Short title: Estimation of the offspring mean

Abstract.

Assume that for a family of probability distributions $\{P_\theta, \theta = (\xi,\eta) \in \Theta\}$ on a sample space X the maximum likelihood estimate $\hat{\xi}$ of ξ is a function of some statistic $t(X)$. It is then not true in general that the maximum likelihood estimate of ξ, based on observing only $t(X)$, also is $\hat{\xi}$. This note provides simple sufficient conditions so that the above fact is true. As a corollary, it is deduced that in a large number of cases, the maximum likelihood estimate of the offspring mean in a Galton - Watson branching process is equal to the total number of children divided by the total number of parents.

Key words: cut, exponential family, Galton - Watson process, L - independence, maximum likelihood estimation.

1. Introduction.

If $Z_0 = z_0, Z_1, \ldots, Z_n$ are the n + 1 first generation sizes of a Galton - Watson branching process, then the maximum likelihood estimate of the offspring mean $m = E(Z_1)$ is in most cases given by

$$\hat{m} = (Z_1 + \cdots + Z_n) / (Z_0 + \cdots + Z_{n-1}),$$

i.e. the total number of children divided by the total number of parents. In fact, Harris (1948) showed this in the case where all individual offspring sizes are observed and the statistical model is that of a completely general family of distributions on the nonnegative integers. Since $\hat{m}$ only depends on the generation sizes $Z_0, \ldots, Z_n$, it is interesting to know whether it is also the maximum likelihood estimate of m in the general model but only based on the observation of $Z_0, \ldots, Z_n$. This is in fact so, but no complete proof existed until Feigin (1976) gave a Lagrange multiplier argument.

The purpose of this note is to point out that the result may be obtained as an elementary corollary to exponential family theory. Following up the preliminary report by Keiding (1975), we also mention some other families of offspring distributions for which $\hat{m}$ is the maximum likelihood estimate of m based on observation of $Z_0, \ldots, Z_n$. Use of this fact has in particular been made by Becker (1974, 1977 a, 1977 b). Our approach also shows more clearly why $\hat{m}$ is just the simple average alluded to above. In a final section, we make some brief comments on the relation to work by Barndorff-Nielsen on ancillarity and by Sundberg on incompletely observed exponential families.

2. Maximum likelihood estimates in marginal distributions

Consider a sample space (X,A) with a family of probability measures $P = \{P_\theta, \theta \in \Theta\}$ dominated by a σ-finite measure μ. Assume that there is given two parameter functions $\psi: \Theta \to \Omega = \psi(\Theta)$ and $\tau: \Theta \to \Lambda = \tau(\Theta)$. Define, for $\omega \in \Omega, P_\omega = \{P_\theta, \theta \in \psi^{-1}(\omega)\}$.

Let $t: X \to \Lambda$ be a statistic.

Lemma 2.1 If for all ω, t(x) is the maximum likelihood estimate of τ(θ) in the family P_ω, then this also holds in the family P.

Proof. Denote the likelihood function by L(θ). By assumption we have

$$\sup_{\theta\in\psi^{-1}(\omega)} L(\theta) = \sup_{\theta\in\psi^{-1}(\omega)\cap\tau^{-1}(t(x))} L(\theta)$$

but then

$$\sup_{\theta\in\Theta} L(\theta) = \sup_{\omega\in\Omega}\sup_{\theta\in\psi^{-1}(\omega)} L(\theta)$$

$$= \sup_{\omega\in\Omega}\sup_{\theta\in\psi^{-1}(\omega)\cap\tau^{-1}(t(x))} L(\theta) = \sup_{\theta\in\tau^{-1}(t(x))} L(\theta),$$

which was to be proved. If also t is a sufficient statistic for each fixed ω, this can be strengthened to:

Lemma 2.2 If the assumptions of Lemma 2.1 are fulfilled, and if further for all ω, t(x) is sufficient in the model P_ω, then t(x) is the maximum likelihood estimate of τ(θ) based on observing t(x) only, as well in the model P_ω as in P.

Proof. By the Neyman factorisation theorem we have that the density of P_θ factorises as

$$\frac{dP_\theta}{d\mu}(x) = g(\psi(\theta),x)h(\theta,t(x)).$$

But then the marginal density of t becomes

$$\frac{dP_\theta t^{-1}}{d\mu t^{-1}}(t) = g^*(\psi(\theta),t)h(\theta,t),$$

and for fixed value of $\psi(\theta)$ we see that the likelihood functions based on x and on t are proportional and therefore yield the same maximum likelihood estimates. Thus t is also the "marginal" maximum likelihood estimate of $\tau(\theta)$ in $\mathcal{P}_\omega$ and Lemma 2.1 shows that this also holds in $\mathcal{P}$.

A typical situation in which Lemmas 2.1 and 2.2 apply is the following: let $\Theta \subseteq R^k \times \Omega$, $\theta = (\xi,\omega) \in \Theta$ and the density be of the form

$$\frac{dP_\theta}{d\mu}(x) = \phi(\theta)^{-1} g(x,\omega)\, e^{\xi\cdot t(x)}, \qquad (2.1)$$

where $t: X \to R^k$, and $g > 0$ a.s.μ. Assume further that for each fixed ω the sections $\Xi_\omega = \{\xi \in R^k \mid (\xi,\omega) \in \Theta\}$ are given as

$$\Xi_\omega = \{\xi \in R^k \mid \int g(x,\omega)e^{\xi\cdot t(x)}\mu(dx) < +\infty\}$$

and that for all $\omega \in \Omega$, Ξ_ω are open subsets of R^k (such a family has been called partly exponential by Barndorff-Nielsen (1971, Sec.5.2). Let C denote the closed convex hull of the support of $P_\theta t^{-1}$ and assume this to be independent of θ.

Lemma 2.3 With the assumptions above, if $t(x) \in \text{int } C$, we let $\psi(\theta) = \omega$ and $\tau(\theta) = E_\theta[t(X)]$, Lemmas 2.1 and 2.2 apply and

$\tau(\hat{\theta}) = t(x)$, whether X or t(X) only has been observed.

Proof. The assumptions imply that P_ω are regular canonical exponential families with t(X) as canonical statistic, cf. Barndorff-Nielsen (1970). Thus t(X) is sufficient in P_ψ and by Theorem 7.1 of the same reference, $\tau(\hat{\theta}) = t(x)$ in P_ψ. Thus the lemmas apply and the proof is complete.

Remark. Note that to derive the results we have not assumed the basic parameter θ to be even identifiable, which it might often not be from the observation of t(X) alone.

Corollary Let $X_1, X_2, \ldots$ be independent identically distributed according to (2.1) and let N be a stopping time not depending on θ, that is, there exist statistics $f_n: X^n \to \{0,1\}$ such that $I\{N = n\} = f_n(X_1, \cdots, X_n)$, $n = 1, 2, \cdots$. Then the maximum likelihood estimate of $E_\theta[t(X_1)]$ based on $X_1, \cdots, X_N$ is $\Sigma_1^n t(x_i)/n$.

Proof. The likelihood function is the density of $(N, X_1, \cdots, X_N)$ which at the point $(n, x_1, \cdots, x_n)$ is given by

$$\phi(\theta)^{-n} \prod_{i=1}^{n} g(x_i, \omega)\, e^{\xi \cdot (\Sigma_1^n t(x_i))} f_n(x_1, \cdots, x_n).$$

This is proportional to the likelihood corresponding to the stopping rule N = n (this general phenomenon is conventionally termed "the likelihood is independent of the stopping rule"). The latter is, however, itself of the form (2.1) so that Lemma 2.3 applies to show that $\Sigma_1^n t(x_i)$ is the maximum likelihood estimate of

$$E_\theta[\Sigma_1^n t(X_i)] = n\, E_\theta[t(X_1)].$$

Example 2.1 The negative binomial distribution.
Let $X_i \in N = \{0,1,2,\cdots\}$, μ be counting measure and

$$\frac{dP_\theta}{d\mu}(x) = \binom{\eta+x-1}{x}(1-e^\xi)^\eta \, e^{\xi x}$$

If we observe $X_1,\cdots,X_n$ independent replications of a random variable from this distribution, where both η and ξ are unknown, it follows from the corollary that $\bar{x}$ is maximum likelihood estimate of $E_\theta(X) = \eta\, e^\xi/(1-e^\xi)$ also if only $\bar{X}$ was observed.

Example 2.2 The completely general family on $\{0,\cdots,k\}$. Let Y be a random variable on $\{0,\cdots,k\}$. Observing Y is equivalent to observing $(I\{Y = 0\},\cdots, I\{Y = k\}) \in X_k$ $= \{(x_0,\cdots,x_k), x_i = 0 \text{ or } 1, \Sigma x_i = 1\}$. The family of distributions with density

$$\eta_0^{x_0}\cdots\eta_k^{x_k}\, e^{\xi t(x)}$$

with $t(x) = \Sigma\, i\, x_i$ and $(\xi,\eta) \in R \times \{(\eta_0,\cdots,\eta_k), 0 \leq \eta_i \leq 1, \Sigma\, \eta_i = 1\}$ is (an overparameterized version of) the completely general family on X_k. Notice that

$$\int \eta_0^{x_0}\cdots\eta_k^{x_k}\, e^{\xi \Sigma i x_i}\, \mu(dx) = \sum_{i=0}^{k} \eta_i\, e^{i\xi} < \infty$$

for all $\xi \in R$, so that $\Xi_\eta = R$. The closed convex support of $t(X)$ is $[0,k]$. It now follows from Lemma 2.3 and the Corollary that if $Y_1,\cdots,Y_n$ are independent identically distributed on $\{0,\cdots,k\}$ according to the general model, then the maximum likelihood estimate $\hat{\tau}$ of $\tau = E(Y)$ based on $\bar{Y}$ is $\bar{y}$, provided $0 < \bar{y} < k$. However, it is readily seen that if $\bar{y} = 0$ (resp.k) then $\hat{\eta}_0 = 1$ (resp. $\hat{\eta}_k = 1$) and hence $\hat{\tau} = 0$ (resp.k) so that in all cases $\hat{\tau} = \bar{y}$.

Example 2.3 The completely general family on N. Let $Y_1,\cdots,Y_n$ be independent identically distributed on $N = \{0,1,2,\cdots\}$ in the statistical model given by all distributions on N and consider the problem of deriving the maximum likelihood estimate of $\tau = E(Y)$ based on observing $\bar{Y}$ only. With similar notation as in Example 2.2 above, the likelihood function when $n\bar{Y} = y$ is

$$L(\eta_0,\eta_1,\cdots) = \sum_{\Sigma x_i=y} \eta_0^{x_0}\eta_1^{x_1}\cdots,$$

$0 \leq \eta_i \leq 1$, $\Sigma\, \eta_i = 1$. Since L does not depend on $\eta_{y+1},\eta_{y+2},\cdots$ and is clearly increasing in each of the $y + 1$ first coordinates separately, it is obvious that the maximization problem is identical to that of the general distribution on $\{0,\cdots,y\}$. And for that problem we showed in Example 2.2 above that $\hat{\tau} = \bar{y}$. Notice that the estimated mean is always finite even though the statistical model allows the theoretical mean to be infinite.

3. Maximum likelihood estimation of the offspring mean in a branching process.

To define a Galton-Watson process, it is convenient to use the following representation. Let $Y_1,Y_2,\cdots$ be independent identically distributed random variables on N.

Let $Z_0 = z_0$, $Z_1 = Y_1 +\cdots+ Y_{z_0}$, $Z_2 = Y_{z_0+1}+\cdots+ Y_{z_0+Z_1},\cdots,$ $Z_n = Y_{z_0+Z_1+\cdots+Z_{n-2}+1}+\cdots+ Y_{z_0+Z_1+\cdots+Z_{n-1}}$. Then $Z_0,\cdots,Z_n$ form the first generation sizes of a Galton-Watson process with the common distribution of the Y's as offspring distribution. Observation of all individual offspring sizes of the n first generation sizes amounts to observing $Y_1,\cdots,Y_{z_0+Z_1+\cdots+Z_{n-1}}$. We notice that observation of the sequence $Y_1,Y_2,\cdots$ is stopped at the stopping time $z_0+Z_1+\cdots+Z_{n-1}$ which does not depend on any parameter. Therefore the likelihood function is obtained from the likelihood of k independent replications of Y_i by replacing k by $z_0+Z_1+\cdots+Z_{n-1}$. If the maximum likelihood estimate of E(Y) based on observing certain partial sums $Y_1+\cdots+Y_{k_1}, Y_{k_1+1}+\cdots+Y_{k_2},\cdots,Y_{k_m+1}+\cdots+Y_k$ only is given by $\bar{y}$, then the maximum likelihood estimate of the offspring mean in the Galton-Watson process based on observing only $Z_1,\cdots,Z_n$ is given by

$$\frac{y_1+\cdots+y_{z_0+z_1+\cdots+z_{n-1}}}{z_0+z_1+\cdots+z_{n-1}} = \frac{z_1+\cdots+z_n}{z_0+\cdots+z_{n-1}} .$$

By a reasoning analogous to that of the proof of the Corollary, it is immediate that the same maximum likelihood estimate is valid if $Z_1,\cdots Z_N$ is observed, where N is a stopping time not depending on the parameters, cf. Becker (1974) and Keiding (1975).

Sufficient conditions for statistical models to satisfy these demands are easily obtained from Section 2. It follows that offspring distributions such as the completely general distributions (Harris 1948, Jagers 1973, 1975, Dion 1974), the power series distributions (Becker 1974, Eschenbach & Winkler 1975, Heyde 1975, Heyde & Feigin 1975), binary splitting (Jagers 1975), the modified geometric distribution (Keiding 1975) and the negative binomial distributions will all lead to $\hat{m}$ as maximum likelihood estimate of m. A nontrivial counterexample is the zeta-distributions with probabilities $p_x = (x+1)^{-\theta}/\zeta(\theta)$, $x = 0,1,2,\cdots$, $1 < \theta < \infty$, with ζ the Riemann zeta function, for which $\hat{m}$ is not the maximum likelihood estimate (Keiding 1975).

For the multitype Galton-Watson process, Asmussen and Keiding (1977) derived for the statistical model containing all offspring distributions the maximum likelihood estimate of the offspring mean matrix $M = (m_{ij})$ based on observation of all parent-offspring combinations $U^{\nu}_{k,i}$: the vector of offspring of the k'th individual of type i alive at time ν, $\nu = 0,\cdots,n-1$. This is given by the obvious average

$$\hat{m}_{ij} = \sum_{\nu=0}^{n-1} z^{i}_{\nu+1}(j) / \sum_{\nu=0}^{n-1} z_{\nu}(i)$$

where $z^{i}_{\nu+1}(j) = \Sigma_{k=1}^{z_\nu} u^{\nu}_{k,i}(j)$ is the number of individuals of type j in the $(\nu+1)$'st generation whose parents were of type i and $z_{\nu}(j) = \Sigma\, z^{i}_{\nu}(j)$. It is now seen that $\hat{m}_{ij}$ is also the maximum likelihood estimate of m_{ij} if only based on observation of the $z^{i}_{\nu}(j)$, provided the class of offspring distributions is of the type discussed.

Similar results concerning estimation of the offspring mean for Markov branching processes can also be derived, cf. Keiding (1975).

4. Remarks.

a. L-independence

If in the factorisation used to prove Lemma 2.2, h depends on θ via $\tau(\theta)$ only, i.e.

$$f(x,\theta) = g(\psi(\theta),x)h(\tau(\theta),t(x))$$

and if $\Theta = \psi^{-1}(\Omega) \times \tau^{-1}(\Lambda)$, then the parameter functions ψ and τ were termed L-independent by Barndorff-Nielsen (1971), who went on to give conditions ensuring that g is in fact the conditional density of X given t(X). In this case t(X) is called a cut and is termed S-ancillary with respect to ψ. It is a special case of Lemma 2.3 that if t(X) is a canonical cut in an exponential family, (cf.loc.cit.Section 3.3), then t(x) is the maximum likelihood estimate of $E[t(X)]$. It is, however, easily seen that already in the trinomial family on $\{0,1,2\}$, the statistic X_1+2X_2 is not a cut, (this fact seems to be implicitly stated by Barndorff-Nielsen (1976)), so that this approach will probably not be able to yield the result of Section 3 (cf. Examples 2.2 and 2.3).

b. Maximum likelihood estimators based on the observation of a statistic only.

Let $\mathcal{P} = \{P_{(\xi,\eta)} \mid (\xi,\eta) \in \Xi \times H\}$ be a dominated family of distributions on some sample space $\mathcal{X}$. If the maximum likelihood estimate $\hat{\xi}$ of ξ based on observing $X \in \mathcal{X}$ depends on some statistic $t(X)$ only, it need not be true that the maximum likelihood estimate of ξ based on observing $t(X)$ only is also given by $\hat{\xi}$. The following example was in the present context given by T.P. Speed (cf. Feigin (1976)) as a counterexample to such a claim by Jagers (1975, Lemma (2.13.2)).

Example 4.1 The normal distribution

Let $X_1, \cdots X_n$ be iid normal $(\mu, \sigma^2), \mu \in \mathcal{R}, \sigma^2 > 0$. The maximum likelihood estimate of μ is $\bar{x}$ whether $(X_1, \cdots, X_n)$ or only $\bar{X}$ is observed, but the maximum likelihood estimate of σ^2 is SSD/n if based on $(X_1, \cdots, X_n)$, but SSD/(n-1) if based on observing SSD = $\Sigma(X_i-\bar{X})^2$ only. In accordance with these facts, we notice that $\bar{x}$ is maximum likelihood estimate of μ whether σ^2 is fixed or varying, but if μ is fixed, then the maximum likelihood estimate of σ^2 is $\Sigma(x_i-\mu)^2/n$ which depends on μ so that Lemma 2.1 does not apply. Further, SSD is not sufficient so the extra conditions of Lemma 2.2 are violated also.

c. Incompletely observed exponential families

A large class of families of distributions with densities of the form (2.1) are generated by incomplete observation of an exponential family in the following way: Suppose

$$\frac{dP_\theta}{d\mu}(x) = \frac{e^{\xi \cdot t(x) + \eta \cdot u(x)}}{\phi(\xi,\eta)}$$

and we do not observe X but only t(X). We have

$$\frac{dP_\theta t^{-1}}{d\mu t^{-1}}(t) = \frac{e^{\xi \cdot t}}{\phi(\xi,\eta)} f(\eta,t),$$

thus being of the form (2.1) It follows from Lemma 2.3 that the maximum likelihood estimate of $E_\theta[t(X)]$ based on t(X) alone is t(x). Most of this also follows from the basic paper by Sundberg (1974), who stated that the likelihood equation for θ based on observing t(X) is

$$E_\theta\left[\begin{matrix} t(X) \\ u(X) \end{matrix} \,\middle|\, t(X) = t(x)\right] = E_\theta\left[\begin{matrix} t(X) \\ u(X) \end{matrix}\right], \qquad (4.1)$$

from which it follows that $t(x) = \widehat{E_\theta[t(X)]}$. But Sundberg did not provide regularity conditions to gurarantee that there is a solution to (4.1) and that a solution maximises the likelihood. That will not be true in general.

Acknowledgement. We thank Steen Andersson for useful suggestions.

References

Asmussen, S. and Keiding, N. (1977) Martingale central limit theorems and asymptotic estimation theory for multitype branching processes. Preprint No.1, Inst. Math. Statist., Univ. Copenhagen.

Barndorff-Nielsen, O. (1970) Exponential families. Exact theory. Various Publication Series No. 19, Mathematical Institute, University of Aarhus.

Barndorff-Nielsen, O. (1971) On Conditional Statistical Inference. Aarhus.

Barndorff-Nielsen, O. (1976) Factorization of likelihood functions for full exponential families. J. Roy. Statist. Soc. B 38, 37-44.

Becker, N. (1974) On parametric estimation for mortal branching processes. Biometrika 61, 393-399.

Becker, N. (1977a) Estimation for a Galton-Watson process with application to epidemics. Biometrics 33, to appear.

Becker, N.G. (1977b) On a general stochastic epidemic model. Theor. Pop. Biol. 11, 23-36.

Dion, J.-P. (1974) Estimation of the mean and the initial probabilities of a branching process. J. Appl. Prob. 11, 687-694.

Eschenbach, W. and Winkler, W. (1975) Maximum-Likelihood-Schätzungen beim Verzweigungsprozess von Galton-Watson. Math. Operationsforsch. u. Statist. 6 213-224.

Feigin, P. (1976) A note on maximum likelihood estimation for simple branching processes. Manuscript, Technion-Israel Institute of Technology, Haifa.

Harris, T.E. (1948) Branching processes. Ann. Math. Statist. 19, 474-494.

Heyde, C.C. (1975) Remarks on efficiency in estimation for branching processes. Biometrika 62, 49-55.

Heyde, C.C. and Feigin, P.D. (1975) On efficiency and exponential families in stochastic process estimation. In: Statistical distributions in scientific work (ed. G.P. Patil, S. Kotz, J.K. Ord) Dordrecht: Reidel, Vol. 1, 227-240.

Jagers, P. (1973) A limit theorem for sums of random numbers of i.i.d. random variables. In: Jagers, P. and Råde, L. (ed.) Mathematics and Statistics. Essays in honour of Harald Bergström. Göteborg, pp. 33-39.

Jagers, P. (1975) Branching processes with biological applications. New York: Wiley.

Keiding, N. (1975) Estimation theory for branching processes. Bull. Int.Statist.Inst. 46 (4), 12-19.

Sundberg, R. (1974) Maximum likelihood theory for incomplete data from an exponential family. Scand. J. Statist. 1, 49-58.

PREPRINTS 1976

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Becker, Niels: Estimation for an Epidemic Model.

No. 2 Kaplan, Norman: A Generalization of a Result of Erdös and Renyi and a Related Problem.

No. 3 Lauritzen, Steffen & Remmer, Ole: An Investigation of the Asymptotic Properties of Geodetic Least Squares Estimates in Certain Types of Network.

No. 4 Asmussen, Søren: Almost Sure Behavior of Linear Functionals of Supercritical Branching Processes.

No. 5 Hald, Anders & Møller, Uffe: On the SPRT of the Mean of a Poisson Process.

No. 6 Lauritzen, Steffen: Projective Statistical Fields.

No. 7 Keiding, Niels: Inference and Tests for Fit in the Birth-Death-Immigration Process.

No. 8 Møller, Uffe: OC and ASN og the SPRT for the Poisson Process.

No. 9 Keiding, Niels: Population Growth and Branching Processes in Random Environments.

No. 10 Andersen, Per K., Andersen, Søren & Lauritzen, Steffen: The Average Noise from a Poisson Stream of Vehicles.

No. 11 Hald, Anders & Møller, Uffe: Multiple Sampling Plans of Given Strength for the Poisson and Binomial Distributions.

No. 12 Johansen, Søren: Two Notes on Conditioning and Ancillarity.

No. 13 Asmussen, Søren: Some Martingale Methods in the Limit Theory of Supercritical Branching Processes.

No. 14 Johansen, Søren: Homomorphisms and general exponential families.

No. 15 Asmussen, Søren & Hering, Heinrich: Some Modified Branching Diffusion Models.

No. 16 Hering, Heinrich: Minimal Moment Conditions in the Limet theory for General Markov Branching Processes.

PREPRINTS 1977

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Asmussen, Søren & Keiding, Niels: Martingale Central Limit Theorems and Asymptotic Estimation Theory for Multitype Branching Processes.

No. 2 Jacobsen, Martin: Stochastic Processes with Stationary Increments in Time and Space.

No. 3 Johansen, Søren: Product Integrals and Markov Processes.

No. 4 Keiding, Niels & Lauritzen, Steffen L. : Maximum likelihood estimation of the offspring mean in a simple branching process.