find it very useful to see the variety of methods available for reliability estimates. It could be stimulating to make comparisons of conclusions under different model assumptions.

References

[1] Csörgö, M., Seshadri, V., and Yalovsky, M. (1975). In *Statistical Distributions in Scientific Work*, Vol. 2, G. P. Patil, et al., eds. Reidel, Dordrecht, pp. 79–90.

[2] David, H. A. (1981). *Order Statistics*, 2nd ed. Wiley, New York.

[3] Davis, D. J. (1952). *J. Amer. Statist. Assoc.*, **47**, 113–150.

[4] Durbin, J. (1975). *Biometrika*, **62**, 5–22.

[5] Epstein, B. (1960, a, b). *Technometrics*, **2**, 83–101; *ibid.* **2**, 167–183.

[6] Epstein, B. and Sobel, M., (1953). *J. Amer. Statist. Assoc.*, **48**, 486–502.

[7] Gail, M. H. and Gastwirth, J. L. (1978). *J. Amer. Statist. Assoc.*, **73**, 787–793.

[8] Galambos, J. and Kotz, S. (1978). *Characterizations of Probability Distributions*. Lecture Notes in Mathematics, Vol. 675, Springer Verlag, Heidelberg. (This book covers a large variety of problems on the theory of the exponential distribution. On pp. 1–5, there is a detailed historical account; a large variety of characterizations leading to the exponential distribution are discussed; the relation of point processes to exponentiality is given in Chapter 4; and it has an up-to-date bibliography on characterizations.)

[9] Guenther, W. C., Patil, S. A. and Uppuluri, V. R. R. (1976). *Technometrics*, **18**, 333–340.

[10] Hahn, G. J. and Nelson, W. B. (1973). *J. Qual. Tech.*, **5**, 178–188.

[11] Johnson, N. L. and Kotz, S. (1970). *Continuous Univariate Distributions I*. Wiley, New York. (Chapter 18 is devoted to the exponential distribution. Main emphasis is on the statistical problems related to the exponential distribution—such as estimation of parameters, tests of significance, generation of random numbers, and goodness of fit tests. It contains a good historical account and a detailed bibliography on the statistical aspects of exponentiality.)

[12] Kaminsky, K. S. (1977). *Technometrics*, **19**, 83–86.

[13] Lawless, J. F. (1971). *Technometrics*, **13**, 725–730.

[14] Likes, J. (1974). *Technometrics*, **16**, 241–244.

[15] Lilliefors, H. W. (1969). *J. Amer. Statist. Assoc.*, **64**, 387–389.

[16] Moran, P. A. P. (1951). *J. R. Statist. Soc. Ser. B*, **13**, 147–150.

[17] Perng, S. K. (1977). *Comm. Statist. A*, **6**, 1399–1407.

[18] Pyke, R. (1965). *J. R. Statist. Soc. Ser. B*, **27**, 395–436; Discussion 437–449.

[19] Rasch, D. (1977). *Biom. Z.*, **19**, 521–528.

[20] Sherman, B. (1950). *Ann. Math. Statist.*, **21**, 339–361.

[21] Shimi, I. N. and C. P. Tsokos (1977). In *The Theory and Applications of Reliability*, Vol. I, C. P. Tsokos and I. N. Shimi, eds., Academic Press, New York, pp. 5–47.

Acknowledgment

This work was supported by the Air Force Office of Scientific Research under a grant (number 78-3504) to Temple University.

(CHARACTERIZATIONS OF DISTRIBUTIONS
CHI-SQUARE DISTRIBUTION
EXPONENTIAL FAMILIES
GAMMA DISTRIBUTION
HAZARD RATE CLASSIFICATION OF DISTRIBUTIONS
KOLMOGOROV-SMIRNOV TESTS
MULTIVARIATE EXPONENTIAL DISTRIBUTION
NORMAL DISTRIBUTION
ORDER STATISTICS)

JANOS GALAMBOS

EXPONENTIAL FAMILIES

There are two main types of parametric families of distributions in statistics, the *transformation* (or *group*) *families* and the *exponential families*. The transformation families are those generated from a single probability measure by a group of transformations on the sample space, the key example being any location-scale family* $\sigma^{-1}f((x-\mu)/\sigma)$, where f is a known probability density function*. The exponential families are characterized by having probability (point or density) function of the form

$$p(x;\omega) = a(\omega)b(x)\exp\{\theta(\omega)\cdot t(x)\}, \quad (1)$$

where ω is a parameter (both x and ω may, of course, be multidimensional), $\theta(\omega)$ and $t(x)$ are vectors of common dimension, k

say, and $\cdot$ denotes inner product, i.e., $\theta(\omega) \cdot t(x) = \sum_{i=1}^{k} \theta_i(\omega) t_i(x)$. Remarkably, the basics of the present statistical theories of both location-scale families and exponential families were given in a single paper by Fisher [19]. Distributional families which are both exponential and of the transformation type possess particularly nice properties. Their general structure has been studied by Roy [26], Rukhin [27, 28], and Barndorff–Nielsen et al. [6].

A great many of the commonly occurring families of distributions are exponential. Examples of such families are the binomial*, multinomial*, Poisson*, geometric*, logarithmic*, (multivariate) normal*, Wishart*, gamma*, beta*, Dirichlet*, and von Mises–Fisher*. Moreover, if $x_1, \ldots, x_m$ are independent observations, each following an exponential family, then the model for $x = (x_1, \ldots, x_m)$ is also exponential, so that, for instance, factorial and regression experiments often have an exponential model.

Sometimes a family is not exponential in its entirety, but the subfamilies obtained by fixing a (one- or multi-dimensional) component of the parameter are exponential. For instance, the negative binomial* family is not exponential, but for any fixed value of the shape parameter one has an exponential family. In a considerable number of other cases a distributional family for observed data y, although not necessarily exponential itself, may be thought of as derivable from an exponential family as follows. There exists a real or fictitious data set x and an exponential family of distributions for x such that y can be viewed as a function of x with the actual distributional model for y being equal to that derived from the exponential model for x. A grouped empirical distribution y obtained by grouping a sample x from the normal distribution provides an example of this. In such incompletely observed exponential situations the underlying exponential structure can often be taken to advantage in a statistical analysis; see further at the end of this article.

The exponential families share a large number of important and useful properties which often make an incisive statistical analysis feasible. The theory of exponential families, outlined in the following, is concerned with such general properties. The theory is, in fact, quite rich and only its more salient and practically useful features can be indicated here. Where no references are given below the reader may consult the monograph by Barndorff-Nielsen [5] for details and additional information. Further guidance to the literature on exponential families is given in the bibliography to the present article.

FIRST PROPERTIES

Let $\mathcal{P}$ denote the exponential family of distributions with probability functions $p(x; \omega)$. The right-hand side of (1) is said to be an exponential representation of $\mathcal{P}$ or $p(x; \omega)$, and the vectors $\theta = \theta(\omega)$ and $t = t(x)$ are called the *canonical parameter* and the *canonical statistic*. (Sometimes the term "natural" is used instead of canonical.) The smallest k for which an exponential representation of $\mathcal{P}$ with θ and t of dimension k is possible is the *order* of $\mathcal{P}$, and such a representation is said to be *minimal*. The canonical statistic t is a sufficient statistic for $\mathcal{P}$ and it is minimal sufficient if the exponential representation is minimal. (Essentially, the exponential families are the only models that allow a sufficient reduction of the data.) For theoretical purposes it is mostly convenient to work with minimal exponential representations and it will be assumed from now on that (1) is minimal. Also, θ can be assumed to be in one-to-one correspondence with ω, and when a is considered as a function of θ, we simply write $a(\theta)$ for $a(\omega(\theta))$. Similarly, we may write $p(x; \theta)$ instead of $p(x; \omega)$, etc. Set $c(\theta) = a(\theta)^{-1}$.

Example 1. The multinomial distribution* on r cells has probability function

$$p(x; \pi) = \frac{n!}{x_1! \cdots x_r!} \pi_1^{x_1} \cdots \pi_r^{x_r}.$$

Here $x_1 + \cdots + x_r = n$ and $\pi_1 + \cdots + \pi_r$

$= 1$. The family of distributions obtained by letting $\pi_1, \ldots, \pi_r$ vary freely except for the restrictions $\pi_1 > 0, \ldots, \pi_r > 0$ is the multinomial family. This family is obviously exponential with $t = x = (x_1, \ldots, x_r)$ as canonical statistic and $\theta = \ln \pi = (\ln \pi_1, \ldots, \ln \pi_r)$ as canonical parameter. However, due to the affine constraint $x_1 + \cdots + x_r = n$, the exponential representation with this t and θ as the canonical variates is not minimal. A minimal representation is obtained by taking $t = (x_1, \ldots, x_{r-1})$ and $\theta = (\ln(\pi_1/\pi_r), \ldots, \ln(\pi_{r-1}/\pi_r))$ and rewriting (1) as

$$p(x;\theta) = (1 + e^{\theta_1} + \cdots + e^{\theta_{r-1}})^{-n} \times \frac{n!}{x_1! \cdots x_r!} e^{\theta_1 x_1 + \cdots + \theta_{r-1} x_{r-1}}. \quad (2)$$

The probability functions $p(x;\omega)$ are all densities with respect to one and the same measure μ, which is, typically, either a counting measure or Lebesgue measure. Let Ω be the domain of variation for ω and let $\Theta = \theta(\Omega)$ denote the canonical parameter domain for $\mathcal{P}$. Furthermore, let $\tilde{\Theta} = \{\theta : \int b(x) e^{\theta \cdot t(x)}\, d\mu < \infty\}$, which is a convex subset of R^k. Then $\mathcal{P}$ is said to be *full* if $\Theta = \tilde{\Theta}$, and $\mathcal{P}$ is *regular* if it is full and if Θ is an open subset of R^k. All of the examples of exponential families mentioned above are regular, and regular families are particularly well behaved (see below). If Θ is a smooth manifold in $\tilde{\Theta}$, then $\mathcal{P}$ is said to be a *curved exponential family*.

Example 2. Let $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ be two independent samples, the first from the normal distribution $N(\xi, \sigma^2)$, the second from the $N(\eta, \tau^2)$ distribution. If all four parameters are unknown, then the model for the full set of observations is a regular exponential family of order 4. The submodel determined by equating the means ξ and η (which is considered in the Behrens–Fisher* situation) is still of order 4 although it involves only three independent parameters. It is nonregular, but constitutes a curved exponential family.

Example 3. A random sample of n individuals from a human population has been classified according to the ABO blood group system, the observed numbers of individuals of the various phenotypes being x_A, x_B, x_{AB}, and x_{OO}. Under the standard genetical assumptions and with p, q, and r denoting the theoretical gene frequencies, the probability of the observation is

$$\frac{n!}{x_A! x_B! x_{AB}! x_{OO}!} (p^2 + 2pr)^{x_A} \times (q^2 + 2qr)^{x_B} (2pq)^{x_{AB}} (r^2)^{x_{OO}}. \quad (3)$$

The model (3) is a curved exponential family of order 3. (It is assumed here that p, q, and r are all positive and that they vary freely except for the constraint $p + q + r = 1$.) A minimal canonical parameter is given by $\theta = (\ln\{(p/r)^2 + 2p/r\}, \ln\{(q/r)^2 + 2q/r\}, \ln\{2(p/r)(q/r)\})$, and this parameter varies over a two-dimensional manifold of $\tilde{\Theta} = R^3$.

Suppose that $\mathcal{P}$ is full and let $\theta \in \operatorname{int} \Theta$, the interior of Θ. The function $c(\theta + \zeta)/c(\theta)$, considered as a function of ζ, is then the Laplace transform for the statistic t under the distribution determined by $p(x;\theta)$. It follows that the cumulants* of t can be obtained by differentiation of $\kappa(\theta) = -\ln a(\theta)$, and in particular one has

$$E_\theta t = \frac{\partial \kappa}{\partial \theta} \quad \text{and} \quad V_\theta t = \frac{\partial^2 \kappa}{\partial \theta\, \partial \theta'} \quad (4)$$

(vectors are taken to be row vectors and transposition is indicated by $'$). The mean value mapping defined on $\operatorname{int} \Theta$ by $\theta \to \tau$, where $\tau(\theta) = E_\theta t$, is one-to-one and both ways continuously differentiable. The range $\mathcal{T}$ of τ is a subset of $\operatorname{int} C$, where C denotes the closed convex hull of the support S of the distribution of t. The sets Θ and C (and also S and $\mathcal{T}$) are important for visualizing the theory of exponential families. It is best to think of Θ and C as being subsets of two different k-dimensional Euclidean spaces, with the exponential character of the statistical model establishing a duality relation between Θ and C. A particular aspect of this is that the powerful theory of convex analysis can be fruitfully applied to the study of

exponential families, the most central reason being that $\kappa(\theta)$ is a strictly convex function on Θ. *See* GEOMETRY IN STATISTICS: CONVEXITY.

Example 1 (continued). The multinomial family is a regular exponential family of order $r-1$, and with $\Theta = R^{r-1}$ and C equal to the simplex $\{(w_1, \ldots, w_{r-1}) : w_1 \geqslant 0, \ldots, w_{r-1} \geqslant 0, w_1 + \cdots + w_{r-1} \leqslant n\}$. Moreover, by (2),

$$\kappa(\theta) = n \ln(1 + e^{\theta_1} + \cdots + e^{\theta_{r-1}}).$$

Example 4. The inverse Gaussian distribution* is a distribution on the positive part of the real axis with probability function

$$P(x; \chi, \psi) = \frac{\sqrt{\chi}}{\sqrt{2\pi}} e^{\sqrt{\chi\psi}} x^{-3/2} e^{-(1/2)(\chi x^{-1} + \psi x)}.$$

Taking $t = (-\frac{1}{2}x^{-1}, -\frac{1}{2}x)$, $\theta = (\chi, \psi)$, and $\Theta = \{(\chi, \psi) : \chi > 0, \psi \geqslant 0\}$, one has an exponential family of order 2 which is full, but not regular, since Θ includes some of its own boundary points. Here $C = \{(w_1, w_2) : w_1 < 0, w_2 < 0, w_1 w_2 \leqslant \frac{1}{4}\}$ and

$$\kappa(\chi, \psi) = -\tfrac{1}{2}\ln\chi - \sqrt{\chi\psi}\,. \qquad (5)$$

Hence, for $\psi > 0$ one finds by differentiation of (5) formulae such as

$$E_{(\chi,\psi)} x = \sqrt{\chi/\psi}\,, \quad E_{(\chi,\psi)} x^{-1} = \chi^{-1} + \sqrt{\psi/\chi}$$

and

$$V_{(\chi,\psi)}(x, x^{-1}) = \begin{bmatrix} \chi^{1/2}\psi^{-3/2} & -(\chi\psi)^{-1/2} \\ -(\chi\psi)^{-1/2} & 2\chi^{-2} + \chi^{-3/2}\psi^{1/2} \end{bmatrix};$$

see (4).

Various important operations on exponential families lead again to exponential families. In particular, the distribution of the canonical statistic t is exponential, and if $x_1, \ldots, x_n$ is a random sample from a population governed by $\mathcal{P}$, then the joint distribution of $x_1, \ldots, x_n$ is exponential with exponential representation

$$a(\theta)^n b(x_1) \cdots b(x_n) \times \exp\{\theta \cdot (t(x_1) + \cdots + t(x_n))\}. \qquad (6)$$

Moreover, if $\theta = (\theta^{(1)}, \theta^{(2)})$ and $t = (t^{(1)}, t^{(2)})$ are similar partitions of θ and t, then the conditional distribution of $t^{(2)}$ given $t^{(1)}$ is exponential with $\theta^{(2)}$ and $t^{(2)}$ as canonical parameter and statistic, respectively. Thus conditioning on $t^{(1)}$ eliminates $\theta^{(1)}$, a fact that is extremely important from the viewpoint of conditional inference*.

Example 5. In the item analysis model the observations x_{ij} $(i = 1, \ldots, r; j = 1, \ldots, s)$ are assumed to be independent 0–1 variates (*see* PSYCHOLOGICAL TESTING THEORY) with

$$\Pr[x_{ij} = 1] = e^{\alpha_i + \beta_j}/(1 + e^{\alpha_i + \beta_j}).$$

For instance, i may indicate a person and j may indicate a question, while x_{ij} is 1 or 0 according as person i answers question j correctly or not. The parameters α_i and β_j describe, respectively, the ability of person i and the difficulty of question j. This is an exponential model with $(x_{1\cdot}, \ldots, x_{r\cdot}, x_{\cdot 1}, \ldots, x_{\cdot s})$, the set of row and column sums, as a canonical statistic and with $(\alpha_1, \ldots, \alpha_r, \beta_1, \ldots, \beta_s)$ as the corresponding canonical parameter. Inference on the row parameters, say, may be performed in the conditional distribution of $(x_{1\cdot}, \ldots, x_{r\cdot})$ given $(x_{\cdot 1}, \ldots, x_{\cdot s})$, which depends on $(\alpha_1, \ldots, \alpha_r)$ only. For most purposes, conditional inference on $(\alpha_1, \ldots, \alpha_r)$ will be superior to unconditional inference, especially if the number of columns s in the table of x_{ij}'s is large compared to the number of rows r. In particular, if r is fixed while s tends to infinity, then the unconditional maximum likelihood estimate* of $(\alpha_1, \ldots, \alpha_r)$ is not even consistent, whereas the conditional estimate is consistent and asymptotically normal (under mild regularity assumptions).

The conditionality phenomenon discussed in Example 5 is not particular to that example

but similar conclusions hold in wide generality; see Andersen [2–4].

The *conjugate family** of a full exponential family $a(\theta)b(x)\exp\{\theta \cdot t\}$ is the family of distributions on Θ having probability functions

$$p(\theta;\gamma,\chi) = d(\gamma,\chi)e^{\chi\cdot\theta-\gamma\kappa(\theta)};$$

i.e., θ is here considered as a random variable, γ and χ are parameters, and $d(\gamma,\chi)$ is a norming constant which makes the integral of $p(\theta;\gamma,\chi)$, relative to Lebesgue measure, equal to 1. Clearly, the conjugate family is also exponential. When it is appropriate to view θ as a random variable and $a(\theta)b(x)\exp\{\theta \cdot t\}$ as the conditional distribution of x given θ, it will in many cases be convenient to choose the conjugate family as the model for θ. The classical construction of the negative binomial distribution as a mixture of the Poisson family with respect to a gamma distribution can be taken as an exemplification of this. Note also that if θ follows a distribution from the conjugate family, then so does the conditional distribution of θ given x (the posterior distribution of θ).

If Θ belongs to the interior of $\tilde{\Theta}$, in particular if $\mathcal{P}$ is regular, then the mean value parameter τ can be used instead of ω or θ as the parameter of $\mathcal{P}$. Also, in this case, the mixed parameter $(\tau^{(1)},\theta^{(2)})$, where $\tau^{(1)} = E_\theta t^{(1)}$, affords a parametrization of $\mathcal{P}$. The latter parametrization has the property that if $\mathcal{P}$ is regular, then $\tau^{(1)}$ and $\theta^{(2)}$ are variation independent.

Any subfamily of an exponential family is, of course, exponential. For a full exponential family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ those subfamilies $\mathcal{P}_0 = \{P_\theta : \theta \in \Theta_0\}$ that are *affine* (*linear*), i.e., for which Θ_0 is of the form $\Theta_0 = \Theta \cap L$ with L an affine (linear) subspace of R^k, are of special importance. In fact, the log-linear models* are precisely of this kind. By a suitable choice of the exponential representation of $\mathcal{P}$ it can always be arranged that L is a linear subspace and moreover, if convenient, that Θ_0 is determined by $\Theta_0 = \{\theta : \theta^{(2)} = 0\}$. If $\mathcal{P}_0$ is linear and if $\mathbb{P}$ denotes the projection onto L, then $\mathcal{P}_0$ has a minimal exponential representation such that $\mathbb{P}\theta$ and $\mathbb{P}t$, interpreted as vectors of the dimension of L, are the canonical parameter and the canonical statistic in this representation.

Example 6. Most of the manageable models and submodels for one- or multidimensional normal random variables are affine in the sense described above.

Example 7. Let x_{ij} $(i = 1, \ldots, r;\ j = 1, \ldots, s)$ be independent 0–1 random variables and denote the probability that x_{ij} equals 1 by $\pi_{ij} (0 < \pi_{ij} < 1)$. The matrix x of these observations follows an exponential model of order $r \cdot s$ and with $\Theta = R^{r\cdot s}$. The item analysis model discussed in Example 5 is the linear subfamily determined by $\Theta_0 = \Theta \cap L = L$, where L is the linear subspace of $R^{r\cdot s}$ such that a point θ belongs to L if and only if $\theta_{ij} = \alpha_i + \beta_j$ for some $\alpha_i \in R$ and $\beta_j \in R (i = 1, \ldots, r;\ j = 1, \ldots, s)$. The projection $\mathbb{P}x$ of x on L is given by $(\mathbb{P}x)_{ij} = x_{i\cdot} + x_{\cdot j} - \bar{x}_{\cdot\cdot}$. Thus $(\bar{x}_{\cdot\cdot}, x_{1\cdot} - \bar{x}_{\cdot\cdot}, \ldots, x_{r-1\cdot} - \bar{x}_{\cdot\cdot}, x_{\cdot 1} - \bar{x}_{\cdot\cdot}, \ldots, x_{\cdot s-1} - \bar{x}_{\cdot\cdot})$ is a minimal canonical statistic for the item analysis model. This statistic is in one-to-one affine correspondence with the canonical (but not minimal canonical) statistic $(x_{1\cdot}, \ldots, x_{r\cdot}, x_{\cdot 1}, \ldots, x_{\cdot s})$.

Example 8. A full m-dimensional Poisson contingency table x is a set of independent Poisson random variables $x_{i_1 i_2 \ldots i_m}$ $(i_\nu = 1, \ldots, d_\nu,\ \nu = 1, \ldots, m)$ with freely varying mean value parameters $\lambda_{i_1 i_2 \ldots i_m}$. In this exponential model x is a minimal canonical statistic with θ given by $\theta_{i_1 i_2 \ldots i_m} = \ln \lambda_{i_1 i_2 \ldots i_m}$ as the corresponding canonical parameter. The hierarchical submodels are linear exponential subfamilies such that $\mathbb{P}t$ is in one-to-one affine correspondence with the so-called minimal set of fitted marginals of the table x.

If Θ contains an open subset of R^k, then the family of distributions of t is complete*, a

fact that is useful, in particular, in relation to Basu's theorem*.

Let u be an arbitrary statistic such that u possesses a probability function $p(u;\theta)$ with respect to some measure ν. Then, assuming for simplicity that $0 \in \Theta$ and $a(0) = 1$ (as can always be arranged), one has

$$p(u;\theta) = a(\theta)E_0(e^{\theta \cdot t} \mid u)p(u;0). \quad (7)$$

The problem of determining an explicit expression for the probability function for u is thus in effect solved if one can find such an expression for just one element of $\mathcal{P}$.

Example 9. The von Mises–Fisher distribution* is the distribution on the unit sphere in d-dimensional Euclidean space whose density with respect to the uniform distribution on the sphere is of the form

$$a(\kappa)e^{\kappa\mu \cdot x} \quad (8)$$

where $\kappa \geqslant 0$ while x and μ are unit vectors in R^d. The norming constant depends on κ only and the class of von Mises–Fisher distributions on the unit sphere in R^d is an exponential family of order d and with x and $\theta = \kappa\mu$ as canonical variates.

Let $x_1, \ldots, x_n$ be a sample of n observations from the distribution (8). The model for $x_1, \ldots, x_n$ is again exponential with the resultant vector $x_\cdot = x_1 + \cdots + x_n$ as canonical statistic and with $\theta = \kappa\mu$ as canonical parameter. By (7) the length r of the resultant $x_\cdot$ has a density with respect to Lebesgue measure of the form

$$a(\kappa)^n E_0(e^{\kappa\mu \cdot x_\cdot} \mid r)p(r;0). \quad (9)$$

The advantage here is that the second and third factors in this expression are to be calculated under the uniform distribution for $x_1, \ldots, x_n$. In particular, observing that the unit vector in the direction of the resultant $x_\cdot$ must also follow the uniform distribution when $\theta = 0$, one sees that $E_0(e^{\kappa\mu \cdot x_\cdot} \mid r) = a(\kappa r)^{-1}$.

The log-likelihood function for ω based on a single observation x with distribution (1) is

$$l(\omega) = l(\omega;x) = \theta(\omega) \cdot t(x) - \kappa(\theta(\omega)), \quad (10)$$

or, in shorthand notation,

$$l(\theta) = l(\theta;t) = \theta \cdot t - \kappa(\theta).$$

This is a strictly concave function of θ provided that the canonical parameter domain Θ is convex.

Note that for a random sample of n observations $x_1, \ldots, x_n$ the log-likelihood function is of the form

$$l(\theta) = n\big(\theta \cdot \bar{t} - \kappa(\theta)\big), \quad (11)$$

where $\bar{t} = (t(x_1) + \cdots + t(x_n))/n$. By (4) the first- and second-order derivatives of (10) may be written

$$\frac{\partial l}{\partial \omega} = (t - \tau)\frac{\partial \theta'}{\partial \omega} \quad (12)$$

and

$$\frac{\partial^2 l}{\partial \omega\, \partial \omega'} = -\frac{\partial \theta}{\partial \omega'} V_\theta t \frac{\partial \theta'}{\partial \omega} + (t - \tau) \cdot \frac{\partial^2 \theta}{\partial \omega\, \partial \omega'}. \quad (13)$$

[The second term on the right-hand side of (13) is to be interpreted as the sum over i of $(t_i - \tau_i)\partial^2\theta_i/(\partial\omega\, \partial\omega')$.] From the latter formula it follows that the Fisher (or expected) information* function $i(\omega)$ is given by

$$i(\omega) = \frac{\partial \theta}{\partial \omega'} V_\theta t \frac{\partial \theta'}{\partial \omega}$$

and that this is related to the observed information function $j(\omega)$, i.e., minus the left-hand side of (13), by

$$i(\omega) = j(\omega) + (t - \tau) \cdot \frac{\partial^2 \theta}{\partial \omega\, \partial \omega'}. \quad (14)$$

In general, the maximum likelihood estimate $\hat{\theta}$ (or $\hat{\omega}$) has to be found by numerical iteration, although in a number of important special cases $\hat{\theta}$ can be expressed explicitly in terms of t. The Newton–Raphson algorithm* and the Davidon–Fletcher–Powell algorithm are both, ordinarily, very efficient for determining $\hat{\theta}$. For certain special types of models, notably log-linear models for contingency tables*, the so-called method of iterative scaling (or Deming–Stephan algorithm) provides a convenient procedure for the computation of $\hat{\theta}$ (see Darroch and Ratcliff [13]). *See* ITERATED MAXIMUM LIKELIHOOD ESTIMATOR.

The exponential families met in practice are mainly either regular or a curved subfamily of a regular family; and for the further discussion $\mathcal{P}$ will be assumed to be of one of these types and the two types will be considered in turn.

REGULAR EXPONENTIAL FAMILIES

For a regular exponential family $\mathcal{P}$ the maximum likelihood estimate exists, and is then —by the strict concavity of $l(\theta)$—unique if and only if the likelihood equation, which may be written [see (12)]

$$E_\theta t = t,$$

has a solution $\hat{\theta}$. This happens precisely when $t \in \operatorname{int} C$.[1] (In case the distributions of $\mathcal{P}$ are discrete, there is therefore a positive probability that $\hat{\theta}$ does not exist, as t will fall on the boundary of C with positive probability.)

It follows by (14) that $i(\hat{\omega}) = j(\hat{\omega})$; i.e., at the maximum likelihood point the observed information equals the expected information.

Suppose that $\mathcal{P}_0$ is a linear (or affine) subfamily of $\mathcal{P}$. If the maximum likelihood estimate exists under $\mathcal{P}$, then it also exists under $\mathcal{P}_0$. The converse does not hold true in general. The precise condition for the maximum likelihood estimate to exist under $\mathcal{P}_0$ is that, in the previously estabished notation, the likelihood equation $\mathbb{P}\tau(\theta) = \mathbb{P}t$ has a solution in Θ_0.

In repeated sampling the maximum likelihood estimate $\hat{\theta}$ is determined from $E_\theta t = \bar{t}$ [see (11)]. Since $\tau(\theta) = E_\theta t$ is a one-to-one smooth function it follows trivially from the asymptotic normality of $\bar{t}$ and from (4) that $\hat{\theta}$ is asymptotically normal with mean θ and variance (matrix) $(nV_\theta t)^{-1}$. More refined approximations to the distributions of $\bar{t}$ and $\hat{\theta}$ may be obtained from Edgeworth or saddlepoint expansions (see Barndorff-Nielsen and Cox [7]). When the order of $\mathcal{P}$ is 1, simple transformations are available for improving the approximate normality or for variance or spread stabilization. Specifically, for a fixed $\lambda \in [0, 1]$, let

$$\phi(\theta) = \int \{\kappa''(\theta)\}^\lambda d\theta$$

(indefinite integration), and note that $\bar{t} = \hat{\tau}$. Then

- **a.** For $\lambda = \frac{1}{3}$, the transformation $\phi(\theta)$ improves normality relative to that of $\hat{\theta}$, and it symmetrizes the log-likelihood function by making the third derivative of this function equal to 0 at the maximum likelihood point $\hat{\phi}$.
- **b.** For $\lambda = \frac{1}{2}$, $\phi(\theta)$ stabilizes the variance relative to that of $\hat{\theta}$, and it makes the second derivative of the log-likelihood function at $\hat{\phi}$—and hence the information function $j(\phi) = i(\phi)$—constant.
- **c.** For $\lambda = \frac{1}{2}$, $\phi(\tau)$ stabilizes the variance relative to that of $\hat{\tau}$.
- **d.** For $\lambda = \frac{2}{3}$, $\phi(\tau)$ improves normality relative to that of $\hat{\tau}$.

[Recall that $\phi(\tau)$ is an abbreviated notation for $\phi(\theta(\tau))$.]

Example 10. In the case of the gamma distribution

$$p(x; \theta) = \frac{\theta^\kappa}{\Gamma(\kappa)} x^{\kappa - 1} e^{-\theta x}$$

with known shape parameter κ, these transformations are (a) $\phi = \theta^{1/3}$, (b) $\phi = \ln \theta$, (c) $\phi = \ln \tau$, and (d) $\phi = \tau^{1/3}$. Here $\tau = E_\theta x$ and multiplicative constants have been dropped from the expressions for ϕ.

CURVED EXPONENTIAL FAMILIES

Curved exponential models occur not seldomly in practice, and they are of great importance in discussions of various key concepts and methods of statistical inference (see, e.g., Efron [16] and Efron and Hinkley [18]). Instances of such models are given in Examples 2, 3, and 11.

In exponential models of this kind the canonical parameter domain Θ is a smooth manifold in R^k. Assume more specifically

that Ω is an open subset of R^m where $m < k$, and that the $k \times m$ matrix $\partial\theta/\partial\omega'$ of partial derivatives of $\theta(\omega)$ is a continuous function of ω and has rank m for all $\omega \in \Omega$. The form of the score function (12) shows that, in general, the maximum likelihood estimate is determined as that value $\hat{\theta} \in \Theta$ for which $t - \tau(\hat{\theta})$ is orthogonal to the tangent plane of Θ at $\hat{\theta}$. The set of values of t that give rise to one and the same estimate $\hat{\theta}$ thus belongs to a hyperplane of dimension $k - m$. Without further assumptions it may be shown that asymptotically $\hat{\omega}$ exists, and is consistent and normally distributed with asymptotic variance $(ni(\omega))^{-1}$. Moreover, if $\mathcal{P}_0 = \{P_\theta : \theta \in \Theta_0\}$is a curved subfamily of $\mathcal{P}$ with Θ_0 a manifold as above, the dimension of the manifold being $q < m$, then the likelihood ratio test statistic for $\mathcal{P}_0$ under $\mathcal{P}$ is asymptotically χ^2-distributed with $m - q$ degrees of freedom.

Example 11. A pure birth process* x_t, with birth intensity $\lambda \in (0, \infty)$ and initial population size 1, has been observed continuously over a fixed time interval $[0, t]$. The family of distributions for $\{x_s : 0 \leqslant s \leqslant t\}$ is curved exponential, of order 2, with $(\ln\lambda, -\lambda)$ as canonical parameter and (b_t, s_t) as the corresponding canonical statistic. Here b_t is the number of births in the interval $[0, t]$ and s_t is the total lifetime,

$$s_t = \int_0^t x_s \, ds.$$

The log-likelihood function is

$$l(\lambda) = b_t \ln\lambda - \lambda s_t$$

and thus

$$\partial l/\partial\lambda = b_t/\lambda - s_t.$$

On comparing with formula (12) one sees that the mean vector $\tau = E_\lambda(b_t, s_t)$ must be proportional to $(\lambda, 1)$; indeed,

$$\tau = \{(e^{\lambda t} - 1)/\lambda\}(\lambda, 1).$$

Moreover, the pairs (b_t, s_t) which give rise to one and the same maximum likelihood estimate $\hat{\lambda}$ are those on the half-line $\{w(\hat{\lambda}, 1) : w > 0\}$. (For details on this model, see the contribution by Keiding to the discussion of Efron [16].)

INCOMPLETELY OBSERVED EXPONENTIAL SITUATIONS

Suppose that the exponential model (1) holds but that only the value u of some function of x—and not x itself—has been observed. A variety of often-occurring statistical situations can usefully be viewed in this way, (see Sundberg [31], Haberman [20], and Pedersen [24]). In particular, many frequency tables for which some of the individuals or items studied are only partly classified fall within the present framework. Phenotype classifications in genetics are commonly of this kind.

By (7), the log-likelihood function based on u is

$$l(\theta) = \kappa(\theta \mid u) - \kappa(\theta)$$

where $\kappa(\theta \mid u) = \log E_0(\exp\{\theta \cdot t\} \mid u)$. If the exponential model is regular, then the first- and second-order derivatives of l with respect to θ take the form

$$\frac{\partial l}{\partial\theta} = E_\theta(t \mid u) - E_\theta t$$

and

$$\frac{\partial^2 l}{\partial\theta\,\partial\theta'} = V_\theta(t \mid u) - V_\theta t.$$

Thus the likelihood equation is

$$E_\theta t = E_\theta(t \mid u). \tag{15}$$

(It may furthermore be noted that the conditional distribution of t given u is again exponential.) Cyclic iteration in this likelihood equation will usually converge to the maximum likelihood estimate $\hat{\theta}$, but convergence is generally slow (see Dempster, et al. [14]). However, in genetical applications the algorithm is often convenient. (In genetics*, this whole procedure for determining maximum likelihood estimates is known as the gene counting method and was developed by Cepellini, et al. [9] and Smith [29, 30]).

Example 3 (continued). Let x_{AA}, x_{AB}, x_{BB}, x_{BO}, x_{AO}, and x_{OO} denote the numbers of individuals in the sample of the various possible genotypes. These numbers follow a regular exponential family of order 2 and one may view the actually observed phenotype

numbers $x_A = x_{AA} + x_{AO}$, $x_B = x_{BB} + x_{BO}$, x_{AB}, x_{OO} as deriving from incomplete observation of the genotypes. It now follows from (15), essentially without calculation, that the likelihood equations are

$$2np = x_A + x_{AB} + x_A \frac{p}{p + 2r},$$

$$2nq = x_B + x_{AB} + x_B \frac{q}{q + 2r}.$$

CONCLUDING REMARKS

The larger part of parametric statistical models occurring in the theory and applications of statistics are exponential or have a partially exponential structure, and the many useful properties shared by exponential families enhance their methodological importance. For obtaining a further appreciation of the role of exponential models in developing and illustrating principles and methods of statistics, the reader is referred to the books by Barndorff-Nielsen [5], Cox and Hinkley [11], and Lehmann [23]; see also Efron and Hinkley [18]. The shared properties mentioned above also imply that wide classes of exponential or partially exponential models can be handled by compact, integrated computer programs, such as in GLIM* or GENSTAT (*see* STATISTICAL SOFTWARE).

Further Reading

In addition to the references given in the article itself, some further bibliographical notes will be presented here. The many important properties of exponential families, beyond that of yielding sufficient reduction, have been discovered rather gradually and through the contributions of many research workers. Only a short bibliography will be presented here, but guidance to the large majority of works in the field is available via the references actually given here; see in particular Barndorff-Nielsen [5], Barndorff-Nielsen and Cox [7], Efron and Hinkley [18], and Lehmann [23].

Fisher's [19] indication that exponential families are the only families of distributions that yield nontrivial sufficiency reductions was quickly taken up by Darmois [12], Koopman [21], and Pitman [25], who sought rigorous mathematical formulations of this indication. Sometimes, therefore, the terms Darmois–Koopman family or Darmois–Koopman–Pitman family* are used instead of exponential family. The first fully satisfactory result in this direction is due to Dynkin [15].

In statistical mechanics* certain exponential families appear, by derivation from various assumptions, as models of the probability distributions of local properties in large physical systems; see, for instance, Kubo [22]. Models of this kind originated with Maxwell*, Boltzmann*, and Gibbs at the end of the last century.

A rather comprehensive account of the exact, as opposed to asymptotic, theory of exponential families is given in Barndorff-Nielsen [5]; see also Chentsov [10] and Efron [17]. Asymptotic properties, including the limiting behavior of maximum likelihood estimators, are discussed in Andersen [1], Berk [8], Sundberg [31], and Barndorff-Nielsen and Cox [7].

The relationships between ancillarity and sufficiency and exponential families are treated in Barndorff-Nielsen [5] and Efron and Hinkley [18].

An interesting class of models, the *factorial series* families*, which possess a number of properties similar to those of exponential families, have been introduced and studied by Berg; see Barndorff-Nielsen [5]. The observation x and parameter ω corresponding to a factorial series family are both k-dimensional vectors whose coordinates are nonnegative integers, and the probability function is of the form

$$p(x; \omega) = a(\omega)b(x)\omega^{(x)},$$

where $\omega^{(x)} = \omega_1^{(x_1)}\omega_2^{(x_2)} \cdots \omega_k^{(x_k)}$ [the notation $n^{(m)}$ indicating the descending factorial, i.e., $n^{(m)} = n(n-1)\cdots(n-m+1)$]. Such distributions arise as the result of certain sampling procedures employed to obtain data for inference on the sizes of various classes of elements or individuals, the parameters $\omega_1, \ldots, \omega_k$ denoting these sizes.

NOTE

1. More generally, if $\mathcal{P}$ is full but not necessarily regular, then the maximum likelihood estimate* $\hat{\theta}$ exists if and only if $t \in \text{int}\, C$; and $\hat{\theta}$ is unique. If, in addition, $E_\theta t = t$ has a solution in the interior of Θ, then this solution equals $\hat{\theta}$. For the maximum likelihood estimate to be always a solution of $E_\theta t = t$ it is therefore necessary and sufficient that $\mathcal{T} = \text{int}\, C$. The latter condition is equivalent to $\kappa(\theta)$ being *steep*, which means that $|\partial\kappa/\partial\theta|$, the length of the gradient of $\kappa(\theta)$, tends to ∞ for θ tending to the boundary of Θ. This occurs, in particular, for regular families. The inverse Gaussian family of distributions provides an example of a steep, nonregular exponential family.

References

[1] Andersen, A. H. (1969). *Bull. Int. Statist. Inst.*, **43**, Bk. 2, 241–242.

[2] Andersen, E. B. (1970). *J. R. Statist. Soc. B*, **32**, 283–301.

[3] Andersen, E. B. (1971). *J. Amer. Statist. Ass.*, **66**, 630–633.

[4] Andersen, E. B. (1973). *Conditional Inference and Models for Measuring*. Mentalhygiejnisk Forlag, Copenhagen.

[5] Barndorff-Nielsen, O. (1978). *Information and Exponential Families*. Wiley, Chichester, England.

[6] Barndorff-Nielsen, O., Blaesild, P., Jensen, J. L., and Jørgensen, B. (1981–1982). *Proc. R. Soc. London* A. To appear.

[7] Barndorff-Nielsen, O. and Cox, D. R. (1979). *J. R. Statist. Soc. B*, **41** 279–312.

[8] Berk, R. H. (1972). *Ann. Math. Statist.*, **43**, 193–204.

[9] Ceppellini, R., Siniscalco, M., and Smith, C. A. B. (1955). *Ann. Hum. Genet. Lond.*, **20**, 97–115.

[10] Chentsov, N. N. (1972). *Statistical Decision Rules and Optimal Conclusions*. Nauka, Moscow (in Russian).

[11] Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

[12] Darmois, G. (1935). *C. R. Acad. Sci. Paris*, **260**, 1265–1266.

[13] Darroch, J. N. and Ratcliff, D. (1972). *Ann. Math. Statist.*, **43**, 1470–1480.

[14] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). *J. R. Statist. Soc. B*, **39**, 1–38.

[15] Dynkin, E. B. (1951). *Select. Trans. Math. Statist. Prob.*, **1**, 23–41. (1961: English transl.; original in Russian).

[16] Efron, B. (1975). *Ann. Statist.*, **3**, 1189–1242.

[17] Efron, B. (1978). *Ann. Statist.*, **6**, 362–376.

[18] Efron, B. and Hinkley, D. V. (1978). *Biometrika*, **65**, 457–487.

[19] Fisher, R. A. (1934). *Proc. R. Soc. Lond. A*, **144**, 285–307.

[20] Haberman, S. J. (1974). *Ann. Statist.*, **2**, 911–924.

[21] Koopman, L. H. (1936). *Trans. Amer. Math. Soc.*, **39**, 399–409.

[22] Kubo, R. (1965). *Statistical Mechanics*. North-Holland, Amsterdam.

[23] Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York.

[24] Pedersen, J. G. (1978). *Ann. Hum. Genet. London*, **42**, 231–237.

[25] Pitman, E. J. G. (1936). *Proc. Camb. Philos. Soc.*, **32**, 567–579.

[26] Roy, K. K. (1975). *Sankhyā* A, **37**, 82–92.

[27] Rukhin, A. L. (1974). *Zap. Naucvn. Sem. Leningrad. Otdel. Mat. Inst. Steklov*, **43**, 59–87. *J. Sov. Math.*, **9**, 886–910. (1978: English transl.).

[28] Rukhin, A. L. (1975). In *Statistical Distributions in Scientific Work*, Vol. 3, G. P. Patil, S. Kotz, and J. K. Ord, eds. D. Reidel, Dordrecht, Holland.

[29] Smith, C. A. B. (1957). *Ann. Hum. Genet. Lond.*, **21**, 254–276.

[30] Smith, C. A. B. (1967). *Ann. Hum. Genet. Lond.*, **31**, 99–107.

[31] Sundberg, R. (1974). *Scand. J. Statist.*, **1**, 49–58.

O. BARNDORFF-NIELSEN

EXPONENTIAL SCORES *See* LOGRANK SCORES

EXPONENTIAL SMOOTHING

Exponential smoothing methods are widely used in operations research*, business, economic, and engineering contexts, particularly in the areas of inventory and production control. See, for example, Smith [12], McKenzie [9], Bradshaw [2], Dyer [5], Little [8], Brennan [3], Wu [13], and Pandit, Burney and Wu [10]. Perhaps the primary and most definitive reference to exponential smoothing per se is Brown [4, Chaps. 7 and 12]. In general, the concept of exponential smoothing is seen by contemporary statistical authors as a special case of a more general modeling procedure. We open our discussion with the classical motivation for exponential smoothing.