Cambridge University Press & Assessment
978-1-108-48890-7 — Exponential Families in Theory and Practice
Bradley Efron
Excerpt
More Information

1

One-parameter Exponential Families

The basic unit of probability theory is a probability distribution. The basic unit of statistical inference is a *family* of probability distributions. Dating from the time of Laplace and Gauss, the one-dimensional normal family[1]

$$x \sim \mathcal{N}(\mu, \sigma^2), \tag{1.1}$$

[1] Equation (1.1) means that the real-valued random variable x has density $\exp\{-(x-\mu)^2/\sigma^2\} \cdot (2\pi\sigma^2)^{-1/2}$ on the real line.

with $\mu \in (-\infty, \infty)$ and σ^2 positive, has played a dominant role in both theory and practice. A strong desire to go beyond normal models fueled the development of exponential family theory. One-parameter exponential families are useful in their own right, and crucial to understanding the multiparameter exponential families of Parts 2 through 5. Here we will present the general one-parameter family theory, and show how it plays out in familiar contexts such as the Poisson, binomial, normal, and gamma distributions.

1.1 Definitions, Notation, and Terminology

This section reviews the basic definitions for exponential families. An exponential family is a set of probability densities $\mathcal{G}$, "density" here including the possibility of discrete atoms (as in the family of binomial densities). A *one-parameter exponential family* has densities $g_\eta(y)$ of the form

$$\mathcal{G} = \left\{ g_\eta(y) = e^{\eta y - \psi(\eta)} g_0(y) m(dy),\ \eta \in A,\ y \in \mathcal{Y} \right\}, \tag{1.2}$$

where A and $\mathcal{Y}$ are subsets of the real line $\mathcal{R}^1$.

There is a more-or-less standard terminology for the elements of (1.2):

- η is the *natural* or *canonical* parameter; in familiar families like the Poisson and binomial, it often isn't the parameter we are used to working with.
- y is the *sufficient* or *natural* statistic, a name that will be more meaningful when we discuss repeated sampling situations; in many cases (the more interesting ones) $y = y(x)$ is a function of an observed data set x (as in the binomial example below); y takes values in its sample space $\mathcal{Y}$.
- The densities in $\mathcal{G}$ are defined with respect to some *carrying measure* $m(dy)$, such as the uniform measure on $[-\infty, \infty]$ for the normal family, or the discrete measure putting weight 1 on the non-negative integers ("counting measure") for the Poisson family. Usually $m(dy)$ won't be indicated in our notation. We will call $g_0(y)$ the *carrying density*.
- $\psi(\eta)$ in (1.2) is the *normalizing function* or *cumulant generating function*; it scales the densities $g_\eta(y)$ to integrate to 1 over sample space $\mathcal{Y}$,

$$\int_{\mathcal{Y}} g_\eta(y) m(dy) = \int_{\mathcal{Y}} e^{\eta y} g_0(y) m(dy) / e^{\psi(\eta)} = 1. \tag{1.3}$$

- The *natural parameter space* A consists of all η for which the integral on the right is finite,

$$A = \left\{\eta : \int_{\mathcal{Y}} e^{\eta y} g_0(y) m(dy) < \infty\right\}. \tag{1.4}$$

Homework 1.1 Use convexity to prove that if η_1 and $\eta_2 \in A$ then so does any point in the interval $[\eta_1, \eta_2]$ (implying that A is a possibly infinite interval in $\mathcal{R}^1$).

Homework 1.2 We can reparameterize $\mathcal{G}$ in terms of $\tilde{\eta} = c\eta$ and $\tilde{y} = y/c$. Explicitly describe the reparameterized densities $\tilde{g}_{\tilde{\eta}}(\tilde{y})$.

Suppose $g_0(y)$ is any given positive function on a subset $\mathcal{Y}$ of the real line. We can construct an exponential family $\mathcal{G}$ through $g_0(y)$ by "tilting" it exponentially,

$$g_\eta(y) \propto e^{\eta y} g_0(y), \tag{1.5}$$

and then renormalizing $g_\eta(y)$ to integrate to 1,

$$g_\eta(y) = e^{\eta y - \psi(\eta)} g_0(y), \qquad \text{where } e^{\psi(\eta)} = \int_{\mathcal{Y}} e^{\eta y} g_0(y) m(dy). \tag{1.6}$$

The space A is all values of η such that the integral is finite. It seems like we might employ other tilting functions, say

$$g_\eta(y) \propto \frac{1}{1 + \eta|y|} g_0(y), \tag{1.7}$$

but only exponential tilting gives convenient properties under independent sampling.

If η_0 is any point in A we can write

$$g_\eta(y) = \frac{g_\eta(y)}{g_{\eta_0}(y)} g_{\eta_0}(y) = e^{(\eta-\eta_0)y - (\psi(\eta)-\psi(\eta_0))} g_{\eta_0}(y). \tag{1.8}$$

This is the same exponential family, now represented with

$$\eta \longrightarrow \eta - \eta_0, \quad \psi \longrightarrow \psi(\eta) - \psi(\eta_0), \quad \text{and} \quad g_0 \longrightarrow g_{\eta_0}. \tag{1.9}$$

Any member $g_{\eta_0}(y)$ of $\mathcal{G}$ can be chosen as the carrier density, with all the other members as exponential tilts of g_{η_0}. *Important*: the sample space $\mathcal{Y}$ is the *same* for all members of $\mathcal{G}$, and all put positive probability on every point in $\mathcal{Y}$. The members of $\mathcal{G}$ are absolutely continuous with respect to each other, which greatly reduces the opportunities for pathologies in exponential families.

The Poisson Family

As an important first example we consider the Poisson family. A Poisson random variable Y having expectation $\mu > 0$ takes values on the non-negative integers $\mathcal{Z}_+ = \{0, 1, \dots\}$,

$$\Pr_\mu\{Y = y\} = e^{-\mu}\mu^y/y!, \qquad \text{for } y \in \mathcal{Z}_+. \tag{1.10}$$

The densities $e^{-\mu}\mu^y/y!$, taken with respect to counting measure on $\mathcal{Y} = \mathcal{Z}_+$, can be written in exponential family form as

$$g_\eta(y) = e^{\eta y - \psi(\eta)} g_0(y) \begin{cases} \eta = \log\mu & (\mu = e^\eta) \\ \psi(\eta) = e^\eta & (= \mu) \\ g_0(y) = 1/y!. \end{cases} \tag{1.11}$$

(Here $g_0(y)$ is not a member of $\mathcal{G}$, and is not even a proper density.)

Homework 1.3 (a) Rewrite $\mathcal{G}$ so that $g_0(y)$ corresponds to the Poisson distribution with $\mu = 1$.

(b) Carry out the numerical calculations that tilt Poi(12), seen in Figure 1.1, into Poi(6).

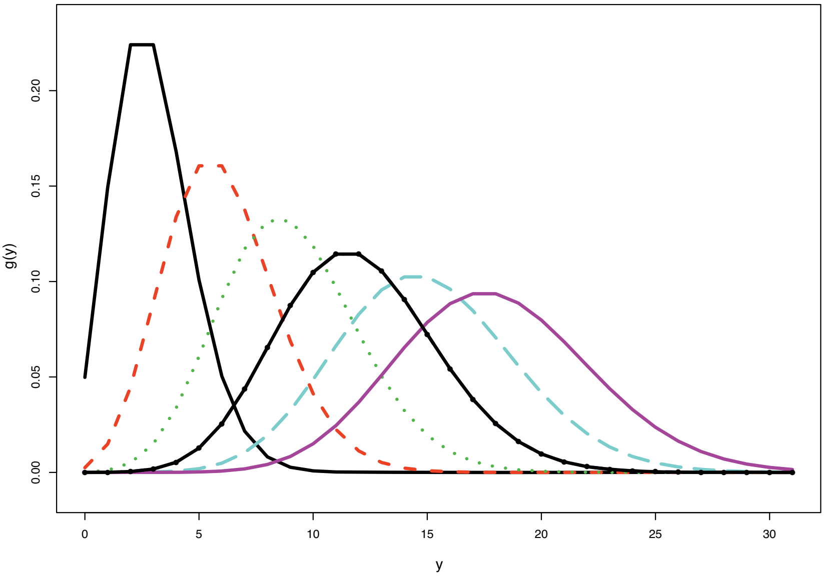


Figure 1.1 Poisson densities for $\mu = 3, 6, 9, 12, 15, 18$; heavy curve with dots for $\mu = 12$.

Even though the mathematics in (1.11) is straightforward, it is still a little surprising to see that any Poisson density is a simple exponential tilt of any other.

1.2 Moment Relationships

The name *cumulant generating function* for the normalizer $\psi(\eta)$ reflects an older methodology for finding expectations, variances, and higher-order moments. The methodology is particularly useful and easy to apply within exponential families.

Expectation and Variance

Differentiating $\exp\{\psi(\eta)\} = \int_{\mathcal{Y}} e^{\eta y} g_0(y) m(dy)$ with respect to η, and indicating differentiation by dots, gives

$$\dot{\psi}(\eta) e^{\psi(\eta)} = \int_{\mathcal{Y}} y e^{\eta y} g_0(y) m(dy) \tag{1.12}$$

and

$$\left(\ddot{\psi}(\eta) + \dot{\psi}(\eta)^2\right) e^{\psi(\eta)} = \int_{\mathcal{Y}} y^2 e^{\eta y} g_0(y) m(dy). \tag{1.13}$$

(The dominated convergence conditions for differentiating inside the integral are always satisfied inside exponential families; see Theorem 2.2 of Brown, 1986.) Multiplying by $\exp\{-\psi(\eta)\}$ gives expressions for the expectation μ_η and variance V_η of Y,

$$\dot{\psi}(\eta) = \mu_\eta = E_\eta\{Y\}, \tag{1.14}$$

$$\ddot{\psi}(\eta) = V_\eta = \mathrm{Var}_\eta\{Y\}, \tag{1.15}$$

where E_η and Var_η indicate expectation and variance under density g_η. V_η is greater than 0, implying that $\psi(\eta)$ has a positive second derivative everywhere, in other words, that $\psi(\eta)$ is convex. Except in trivial cases, the variance V_η is positive for all $\eta \in A$.

Notice that

$$\dot{\mu} = \frac{d\mu}{d\eta} = V_\eta > 0.$$

The mapping from η to μ is 1:1 increasing and infinitely differentiable. We can index the family $\mathcal{G}$ just as well with μ, the *expectation parameter*, as with η. Functions like $\psi(\eta)$, E_η, and V_η can just as well be thought of as functions of μ. We will sometimes write ψ, V, etc. when it's not necessary to specify the argument. Notations such as V_μ formally mean $V_{\eta(\mu)}$.

Note Suppose that ζ is a parameter that can be defined as a function of either η or μ,

$$\zeta = h(\eta) = H(\mu).$$

Let $\dot{h} = dh/d\eta$ and $H' = dH/d\mu$. Then

$$H' = \dot{h}\frac{d\eta}{d\mu} = \frac{\dot{h}}{V}. \tag{1.16}$$

Skewness and Kurtosis

The first two moments of a random variable Y describe its expectation and variance. The third and fourth moments give its *skewness* and *kurtosis*, valuable for higher-order asymptotic approximations. For instance, a first-order Edgeworth expansion says that

$$\Pr\{Y \leq \text{ median }(Y)\} \doteq 0.5 + \frac{1}{6\sqrt{2\pi}} \text{ SKEWNESS }(Y),$$

while the second-order approximation also involves Y's kurtosis.

A pre-computer technology, *cumulants*[2] are certain linear combinations of moments that are easy to deal with in repeated sampling situations (Section 1.3). $\psi(\eta)$ is the *cumulant generating function* for g_0 and $\psi(\eta) - \psi(\eta_0)$ is the CGF for $g_{\eta_0}(y)$, that is,

$$e^{\psi(\eta)-\psi(\eta_0)} = \int_{\mathcal{Y}} e^{(\eta-\eta_0)y} g_{\eta_0}(y)m(dy).$$

By definition, the Taylor series for $\psi(\eta) - \psi(\eta_0)$ has the cumulants k_j of $g_{\eta_0}(y)$ as its coefficients,

$$\psi(\eta) - \psi(\eta_0) = k_1(\eta - \eta_0) + \frac{k_2}{2}(\eta - \eta_0)^2 + \frac{k_3}{6}(\eta - \eta_0)^3 + \cdots .$$

[2] Cumulants add correctly under independent sampling: if X and Y are independent then the jth cumulant of $X + Y$ is the sum of their jth cumulants, this holding for all j. This isn't true for central jth moments $E_0\{Y - \mu_0\}^j$ for $j > 3$. Cumulants are an algebraic computational tool for simplifying higher-order moment relationships, but here we will never go beyond $j = 4$. Older texts, such as Kendall and Stuart (1958), tabulate the relations of cumulants and moments up to $j = 10$.

Equivalently, letting dots indicate derivatives,

$$\dot\psi(\eta_0) = k_1 \quad (= \mu_0), \qquad \ddot\psi(\eta_0) = k_2 \quad (= V_0),$$
$$\dddot\psi(\eta_0) = k_3 \quad \left(= E_0\{Y-\mu_0\}^3\right),$$
$$\ddddot\psi(\eta_0) = k_4 \quad \left(= E_0\{Y-\mu_0\}^4 - 3V_0^2\right),$$

etc., where $k_1, k_2, k_3, k_4, \ldots$ are the cumulants of g_{η_0}.

A real-valued random variable Y has skewness and kurtosis defined by

$$\text{SKEWNESS}(Y) = \frac{E(Y-EY)^3}{(\text{Var}(Y))^{3/2}} \equiv \text{“}\gamma\text{”} = \frac{k_3}{k_2^{3/2}}$$

and

$$\text{KURTOSIS}(Y) = \frac{E(Y-EY)^4}{(\text{Var}(Y))^2} - 3 \equiv \text{“}\delta\text{”} = \frac{k_4}{k_2^2}.$$

Putting this together, if $Y \sim g_\eta(\cdot)$ is an exponential family, then

$$Y \sim \Big[\underset{\substack{\uparrow\\ \text{expectation}}}{\dot\psi,} \quad \underset{\substack{\uparrow\\ \text{standard}\\ \text{deviation}}}{\ddot\psi^{1/2},} \quad \underset{\substack{\uparrow\\ \text{skewness}}}{\dddot\psi/\ddot\psi^{3/2},} \quad \underset{\substack{\uparrow\\ \text{kurtosis}}}{\ddddot\psi/\ddot\psi^{2}}\Big], \tag{1.17}$$

where the derivatives are taken at η.

For the Poisson family

$$\psi = e^\eta = \mu,$$

so all the cumulants equal μ

$$\dot\psi = \ddot\psi = \dddot\psi = \ddddot\psi = \mu,$$

giving

$$Y \sim \Big[\underset{\substack{\uparrow\\ \text{exp}}}{\mu,} \quad \underset{\substack{\uparrow\\ \text{st dev}}}{\sqrt{\mu},} \quad \underset{\substack{\uparrow\\ \text{skew}}}{1/\sqrt{\mu},} \quad \underset{\substack{\uparrow\\ \text{kurt}}}{1/\mu}\Big]. \tag{1.18}$$

A Useful Result

Continuing to use dots for derivatives with respect to η and primes for derivatives with μ, notice that

$$\gamma = \frac{\dddot\psi}{\ddot\psi^{3/2}} = \frac{\dot V}{V^{3/2}} = \frac{V'}{V^{1/2}} \tag{1.19}$$

(using $H' = \dot{h}/V$). Therefore

$$\gamma = 2\left(\sqrt{V}\right)' = 2\frac{d}{d\mu}\,\mathrm{sd}_\mu, \tag{1.20}$$

where $\mathrm{sd}_\mu = V_\mu^{1/2}$ is the standard deviation of y. In other words, $\gamma/2$ is the rate of change of sd_μ with respect to μ; this plays a role in the theory of bootstrap confidence intervals (Part 5).

Homework 1.4 Show that

$$\text{(a)}\ \ \delta = V'' + \gamma^2 \quad \text{and} \quad \text{(b)}\ \ \gamma' = \frac{\delta - {}^3\!/\!{}_2\gamma^2}{\mathrm{sd}}.$$

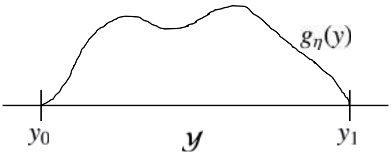
Note The classical exponential families – binomial, Poisson, normal, etc. – are those with closed-form CGFs ψ, yielding neat expressions for means, variances, skewnesses, and kurtoses.

Modern computing power lets us work with general exponential families where results like (1.17) can be exploited numerically, no matter what the form of $\psi(\eta)$.

Unbiased Estimate of η

By definition y is an unbiased estimate of μ (and, in fact, by completeness the only unbiased estimate of form $t(y)$). What about η?

- Let $l_0(y) = \log g_0(y)$ and $l_0'(y) = dl_0(y)/dy$.
- Suppose $\mathcal{Y} = [y_0, y_1]$ (a possibly infinite interval) and that $m(y) = 1$ for all $y \in \mathcal{Y}$.



Lemma 1.1

$$E_\eta\left\{-l_0'(y)\right\} = \eta - \left(g_\eta(y_1) - g_\eta(y_0)\right).$$

Homework 1.5 Prove Lemma 1.1. (*Hint*: Integration by parts.)

So, if $g_\eta(y) = 0$ (or → 0) at the extremes of $\mathcal{Y}$, then $-l_0'(y)$ is a unbiased estimate of η.

Homework 1.6 Numerically calculate values of $-l_0'(y)$ to estimate η using Lemma 1.1 for $y \sim \mathrm{Poi}(\mu)$. Does it work?

1.3 Repeated Sampling

One-parameter exponential families have a crucial property that makes them simple to deal with, both in theory and practice: in repeated sampling situations, they retain one-parameter exponential family structure.[3]

Suppose $y_1, \dots, y_n$ is an independent and identically distributed (i.i.d.) sample from an exponential family $\mathcal{G}$:

$$y_1, \dots, y_n \overset{\text{iid}}{\sim} g_\eta(\cdot), \tag{1.21}$$

for an unknown value of the parameter $\eta \in A$. The density of $\boldsymbol{y} = (y_1, \dots, y_n)$ is

$$\begin{aligned}\prod_{i=1}^n g_\eta(y_i) &= e^{\sum_1^n(\eta y_i - \psi)} \prod_{i=1}^n g_0(y_i)\\ &= e^{n(\eta\bar{y}-\psi)} \prod_{i=1}^n g_0(y_i),\end{aligned}$$

where $\bar{y} = \sum_{i=1}^n y_i/n$. Letting $g_\eta^{(n)}(\boldsymbol{y})$ indicate the density of $\boldsymbol{y}$ with respect to $\prod_{i=1}^n m(dy_i)$,

$$g_\eta^{(n)}(\boldsymbol{y}) = e^{n(\eta\bar{y}-\psi(\eta))} \prod_{i=1}^n g_0(y_i). \tag{1.22}$$

This is a one-parameter exponential family, with:

- natural parameter $\eta^{(n)} = n\eta \quad (\text{so } \eta = \eta^{(n)}/n)$;
- sufficient statistic $\bar{y} = \sum_1^n y_i/n \quad (\bar{\mu} = E_{\eta^{(n)}}\{\bar{y}\} = \mu)$;
- normalizing function $\psi^{(n)}(\eta^{(n)}) = n\psi(\eta^{(n)}/n)$;
- carrier density $\prod_{i=1}^n g_0(y_i) \quad$ (with respect to $\prod m(dy_i)$).

Homework 1.7 Show that, in the bracket notation of (1.17),

$$\bar{y} \sim \left[\mu, \sqrt{\frac{V}{n}}, \frac{\gamma}{\sqrt{n}}, \frac{\delta}{n}\right].$$

Note In the following, we usually index the parameter space by η rather than $\eta^{(n)}$.

[3] The older name, “Koopman–Darmois–Pitman” families, came from the separate efforts of the three authors to show that, under mild conditions, *only* definition (1.2) allowed this kind of sufficiency property.

Notice that $\boldsymbol{y}$ is now a vector, and that the tilting factor $e^{\eta^{(n)}\bar{y}}$ is tilting the *multivariate* carrier density $\prod_1^n g_0(y_i)$. This is still a one-parameter exponential family because the tilting is in a single direction, along $\mathbf{1} = (1, \dots, 1)$.

The sufficient statistic $\bar{y}$ also has a one-parameter exponential family of densities,

$$g_\eta^{(n)}(\bar{y}) = e^{n(\eta\bar{y}-\psi)} g_0^{(n)}(\bar{y}),$$

where $g_0^{(n)}(\bar{y})$ is the g_0 density of $\bar{y}$ with respect to $m^{(n)}(d\bar{y})$, the induced carrying measure.

The density (1.22) can also be written (ignoring the carrier) as

$$e^{\eta S - n\psi}, \qquad \text{where } S = \sum_{i=1}^n y_i.$$

This moves a factor of n from the definition of the natural parameter to the definition of the sufficient statistic. For any constant c we can re-express an exponential family $\{g_\eta(y) = \exp(\eta y - \psi) g_0(y)\}$ by mapping η to η/c and y to cy. This tactic will be useful when we consider multiparameter exponential families.

Homework 1.8 $y_1, \dots, y_n \overset{\text{iid}}{\sim} \text{Poi}(\mu)$. Describe the distributions of $\bar{y}$ and S, and say what are the exponential family quantities $(\eta, y, \psi, g_0, m, \mu, V)$ in both cases.

1.4 Maximum Likelihood Estimation in Exponential Familes

This section briefly reviews some basic results on maximum likelihood estimation (also with a few words about testing). The methodology is particularly simple in exponential families, as we will see. A good reference is Lehmann and Casella (1998), *Theory of Point Estimation*.

Suppose we observe a random sample $\boldsymbol{y} = (y_1, \dots, y_n)$ from a member $g_\eta(y)$ of an exponential family $\mathcal{G}$,

$$y_i \overset{\text{iid}}{\sim} g_\eta(y), \qquad i = 1, \dots, n,$$

and wish to estimate η. According to (1.22) in Section 1.3, the density of $\boldsymbol{y}$ is

$$g_\eta^{(n)}(\boldsymbol{y}) = e^{n[\eta\bar{y} - \psi(\eta)]} \prod_{i=1}^n g_0(y_i), \tag{1.23}$$