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.

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

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

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

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.

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

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 nonnegative integers $\mathcal{Z}_+ = \{0, 1, \ldots\}$,

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

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).

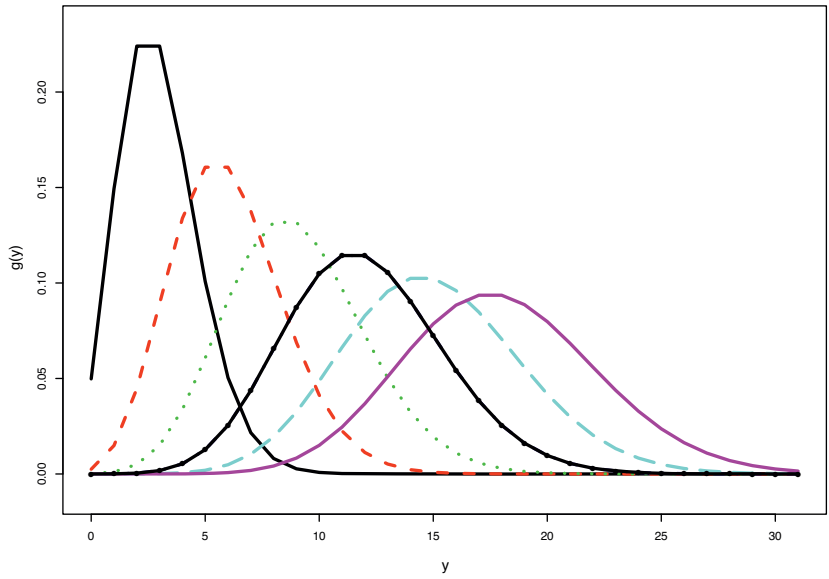


Figure 1.1 Poisson densities for $\mu = 3, 6, 9, 12, 15, 18$; heavy curve with dots for $\mu = 12$.

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

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

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

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

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

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 $j = 10$.

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

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

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

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

(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

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

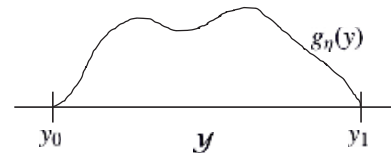
Note The classical exponential families – binomial, Poisson, normal, etc. – are those with closed-form CGFs ψ, yielding neat expressions for means, variances, skewnesses, and kurtoses.

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

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

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

So, if $g_\eta(y) = 0$ (or $\to 0$) 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?

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

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

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

for an unknown value of the parameter $\eta \in A$. The density of $\boldsymbol{y} = (y_1, \ldots, y_n)$ 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 $\eta^{(n)} = n\eta$ (so $\eta = \eta^{(n)}/n$);
- sufficient statistic $\bar{y} = \sum_1^n y_i/n$ ($\bar{\mu} = E_{\eta(n)}\{\bar{y}\} = \mu$);
- normalizing function $\psi^{(n)}(\eta^{(n)}) = n\psi(\eta^{(n)}/n)$;
- carrier density $\prod_{i=1}^n g_0(y_i)$ (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.

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

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 $\mathbf{1} = (1, \ldots, 1)$.

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 $y_1, \ldots, y_n \overset{\text{iid}}{\sim} \text{Poi}(\mu)$. 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 $\boldsymbol{y} = (y_1, \ldots, y_n)$ from a member $g_\eta(y)$ of an exponential family $\mathcal{G}$,

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

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

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

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

where $\bar{y} = \sum_1^n y_i/n$. The log likelihood function $l_\eta(\boldsymbol{y}) = \log g_\eta^{(n)}(\boldsymbol{y})$, with $\boldsymbol{y}$ fixed and η varying, is

$$l_\eta(\boldsymbol{y}) = n\left[\eta\bar{y} - \psi(\eta)\right],$$

giving *score function* $\dot{l}_\eta(\boldsymbol{y}) = \partial/\partial\eta\, l_\eta(y)$ equaling

$$\dot{l}_\eta(\boldsymbol{y}) = n(\bar{y} - \mu) \tag{1.24}$$

(remembering that $\dot{\psi}(\eta) = \partial/\partial\eta\,\psi(\eta)$ equals μ, the expectation parameter).

The maximum likelihood estimate (MLE) of η is the value $\hat{\eta}$ satisfying

$$\dot{l}_{\hat{\eta}}(\boldsymbol{y}) = 0.$$

Looking at (1.24), $\hat{\eta}$ is that η such that $\mu = \dot{\psi}(\eta)$ equals $\bar{y}$, that is,

$$\hat{\eta} : E_{\eta=\hat{\eta}}\left\{\overline{Y}\right\} = \bar{y}.$$

In other words, the MLE matches the theoretical expectation of $\overline{Y}$ to the observed mean $\bar{y}$.

We can also take the score function with respect to μ,

$$\frac{\partial}{\partial\mu}\, l_\eta(\boldsymbol{y}) = \dot{l}_\eta(\boldsymbol{y})\frac{\partial\eta}{\partial\mu} = \frac{\dot{l}_\eta(\boldsymbol{y})}{V} = \frac{n(\bar{y}-\mu)}{V}. \tag{1.25}$$

This gives

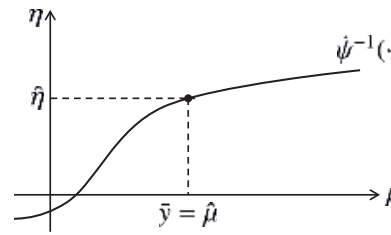
$$\frac{\partial}{\partial\mu}\, l_\eta(\boldsymbol{y})\Big|_{\mu=\bar{y}} = 0,$$

which shows that the MLE of μ is

$$\hat{\mu} = \bar{y}.$$

But $\mu = \dot{\psi}(\eta)$, a monotone one-to-one function; since MLEs map in the obvious way, we get

$$\hat{\eta} = \dot{\psi}^{-1}(\bar{y}).$$



For the Poisson $\hat{\eta} = \log\bar{y}$, and for the binomial, according to what we will see in Section 1.5,

$$\hat{\eta} = \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right), \qquad \text{where } \hat{\pi} = \frac{y}{N}.$$

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

Fisher information is the expected square of the score function – which, since the expected score is always zero, is also its variance – denoted

$$i_\eta^{(n)} = nV$$

for the information for η. We write simply i_η for the case $n = 1$. The information for μ is

$$i_\eta^{(n)}(\mu) = n/V,$$

using (1.25), the notation being understood as the information for μ in a sample of size n, evaluated for $g_\eta(\boldsymbol{y})$. As always, V stands for V_η, the variance of a single observation y from $g_\eta(\cdot)$.

Let $\zeta = h(\eta)$ be any smooth function of η, also expressed as, say,

$$\zeta = H(\mu) = h\left(\dot{\psi}^{-1}(\mu)\right).$$

Then ζ has MLE $\hat{\zeta} = h(\hat{\eta}) = H(\hat{\mu})$ and score

$$\frac{\partial}{\partial \zeta} l_\eta(\boldsymbol{y}) = \frac{\dot{l}_\eta(\boldsymbol{y})}{\dot{h}(\eta)}.$$

Figure 1.2 and Table 1.1 show the MLE and information relationships.

In general, the Fisher information i_θ for a one-parameter family $f_\theta(x)$ has two expressions in terms of the first and second derivatives of the log likelihood,

$$i_\theta = E\left\{\left(\frac{\partial l_\theta}{\partial \theta}\right)^2\right\} = -E\left\{\frac{\partial^2 l_\theta}{\partial \theta^2}\right\}. \tag{1.26}$$

For $i_\eta^{(n)}$, the Fisher information for η in $\boldsymbol{y} = (y_1, \ldots, y_n)$, we have

$$-\ddot{l}_\eta(\boldsymbol{y}) = -\frac{\partial^2}{\partial \eta^2}\, n(\eta \bar{y} - \psi) = -\frac{\partial}{\partial \eta}\, n(\bar{y} - \mu) = nV_\eta = i_\eta^{(n)}, \tag{1.27}$$

so in this case $-\ddot{l}_\eta(\boldsymbol{y})$ gives $i_\eta^{(n)}$ without requiring an expectation over $\boldsymbol{y}$.

Homework 1.9 (a) Does

$$i_\eta^{(n)}(\mu) = -\frac{\partial^2}{\partial \mu^2}\, l_\eta(\boldsymbol{y})\ ?$$

(b) Does

$$i_{\eta=\hat{\eta}}^{(n)}(\mu) = -\frac{\partial}{\partial \mu^2}\, l_\eta(\boldsymbol{y})\Big|_{\eta=\hat{\eta}} \qquad (\hat{\eta} \text{ the MLE})\ ?$$

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

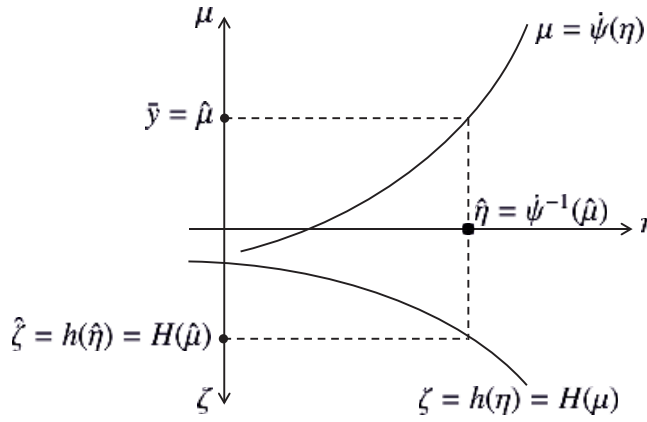


Figure 1.2 Maximum likelihood estimates.

Table 1.1 *Score functions and Fisher information.*

Score functions		Fisher information
$\eta:$	$\dot{l}_\eta(\mathbf{y}) = n(\bar{y}-\mu)$	$i_\eta^{(n)} = \mathrm{Var}_\eta\left[\dot{l}_\eta(\mathbf{y})\right] = nV = ni_\eta$
$\mu:$	$\dfrac{\partial l_\eta(\mathbf{y})}{\partial \mu} = \dfrac{n(\bar{y}-\mu)}{V}$	$i_\eta^{(n)}(\mu) = \dfrac{n}{V} = ni_\eta(\mu)$
$\zeta:$	$\dfrac{\partial l_\eta(\mathbf{y})}{\partial \zeta} = \dfrac{n(\bar{y}-\mu)}{\dot{h}(\eta)}$	$i_\eta^{(n)}(\zeta) = \dfrac{nV}{\dot{h}(\eta)^2} = ni_\eta(\zeta)$

Cramér–Rao Lower Bound

The Cramér–Rao lower bound (CRLB) for an unbiased estimator $\bar{\zeta}$ of a general parameter $\zeta = h(\eta)$ is

$$\mathrm{Var}_\eta(\bar{\zeta}) \geq \frac{1}{i_\eta^{(n)}(\zeta)} = \frac{\dot{h}(\eta)^2}{nV_\eta}. \tag{1.28}$$

For $\zeta \equiv$ the expectation parameter μ we get

$$\mathrm{Var}(\bar{\mu}) \geq \frac{V_\eta^2}{nV_\eta} = \frac{V_\eta}{n}. \tag{1.29}$$

In this case the MLE $\hat{\mu} = \bar{y}$ is unbiased and achieves the CRLB. This happens only for μ or linear functions of μ, and not for η, for instance. The regularity conditions necessary for the CRLB are almost always satisfied

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

in exponential families, exceptions occuring at boundary points η in those unusual cases where A is a finite or partially finite closed set.

In general, the MLE $\hat{\zeta}$ is *not* unbiased for $\zeta = h(\eta)$, but the bias is of order $1/n$,

$$E_\eta\{\hat{\zeta}\} = \zeta + B(\eta)/n$$

(see Section 10 of Efron, 1975). A more general form of the CRLB gives

$$\mathrm{Var}_\eta(\hat{\zeta}) \geq \frac{\left(\dot{h}(\eta) + \dot{B}(\eta)/n\right)^2}{nV_\eta} = \frac{\dot{h}(\eta)^2}{nV_\eta} + O\left(n^{-2}\right).$$

Usually $\dot{h}(\eta)^2/(nV_\eta)$ is a reasonable approximation for $\mathrm{Var}_\eta(\hat{\zeta})$.

Delta Method

The *delta method* uses a first-order Taylor series expansion to calculate approximate variances. If X has mean μ and variance σ^2, then $Y = H(X) \doteq H(\mu) + H'(\mu)(X - \mu)$ has approximate mean and variance

$$Y \dot{\sim} \left[H(\mu), \sigma^2 \left(H'(\mu)\right)^2\right].$$

Homework 1.10 Show that if $\zeta = h(\eta)$, then the MLE $\hat{\zeta}$ has delta method approximate variance

$$\mathrm{Var}_\eta(\hat{\zeta}) \doteq \frac{\dot{h}(\eta)^2}{nV_\eta},$$

in accordance with the CRLB $(i_\eta^{(n)}(\zeta))^{-1}$. (In practice one must substitute $\hat{\eta}$ for η in order to estimate $\mathrm{Var}_\eta(\hat{\zeta})$.)

The simple exponential form of exponential family densities has happy consequences for hypothesis testing as well as estimation. Suppose we wish to test the null hypothesis

$$H_0\colon\ \eta = \eta_0 \quad \text{versus} \quad H_1\colon\ \eta = \eta_1$$

for values $\eta_1 > \eta_0$. From (1.23) we get

$$\log \frac{g_{\eta_1}^{(n)}(\bar{y})}{g_{\eta_0}^{(n)}(\bar{y})} = n\left[(\eta_1 - \eta_0)\bar{y} - (\psi(\eta_1) - \psi(\eta_0))\right],$$

which is an increasing function of $\bar{y}$. By the Neyman–Pearson lemma, the most powerful level α test of H_0 ("MP_α") rejects for $\bar{y} \geq \overline{Y}_0^{(1-\alpha)}$, where $\overline{Y}_0^{(1-\alpha)}$ is the $(1-\alpha)$th quantile of $\overline{Y}$ under H_0. But this doesn't depend on η_1, so the test is uniformly most powerful level α ("UMP_α").

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

For nonexponential familes such as the Cauchy translation family

$$g_\eta(y) = \frac{1}{\pi}\,\frac{1}{1+(y-\eta)^2},$$

the MP_α test depends on η_1. Efron (1975) shows that in a certain geometric sense a one-parameter exponential family is a straight line through the space of probability distributions, and this accounts for the UMP property.

1.5 Some Important One-parameter Exponential Families

A good first course in statistics will include various distribution families – normal, Poisson, binomial, gamma – all of which turn out to be one-parameter exponential families. We introduced the Poisson in Section 1.1. This section examines other well-known families, and some that are not so well known. All of these have one important thing in common: their normalizing function $\psi(\eta)$, the CGF, has a closed-form expression. Modern computing ability lets us construct useful exponential families where $\psi(\eta)$ is *not* closed-form, a first example appearing in Section 1.7.

Normal with Variance 1

Normal distributions have played a central role in the evolution of inferential statistics. We begin with the simplest case, where the observed variable Y is normal with unknown expectation μ and fixed variance 1, indicated $Y \sim \mathcal{N}(\mu, 1)$. The family $\mathcal{G}$ has densities

$$g_\mu(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu)^2} \qquad (\mu, y \in \mathcal{R}^1) \tag{1.30}$$

with respect to Lebesque measure $m(dy) = dy$. This can be written in exponential family form as

$$g_\mu(y) = e^{\mu y - \mu^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-y^2/2}.$$

In terms of the definitions following (1.1), μ is the expectation parameter $E_\eta\{Y\}$, and

$$\eta = \mu, \quad y = y, \quad \psi = \frac{1}{2}\eta^2, \quad g_0(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$$

($g_0(y)$, the *standard normal density*, is often denoted as $\phi(y)$). The variance function is $V_\eta = 1$.

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

Homework 1.11 Suppose $Y \sim \mathcal{N}(\mu, \sigma^2)$ with σ^2 fixed and known. Derive η, y, ψ, and g_0.

Binomial

$Y \sim \text{Bi}(N, \pi)$, N known, indicates the number of successes in N independent flips of a coin with probability π of success. The density function is

$$g(y) = \binom{N}{y} \pi^y (1-\pi)^{N-y} \qquad (y = 0, 1, \ldots, N) \tag{1.31}$$

with respect to counting measure on $\{0, 1, \ldots, N\}$. This can be written as

$$\binom{N}{y} e^{\left(\log \frac{\pi}{1-\pi}\right) y + N \log(1-\pi)},$$

a one-parameter exponential family, with:

- $\eta = \log[\pi/(1-\pi)]$ $\quad$ (so $\pi = (1+e^{-\eta})^{-1}$, $1-\pi = (1+e^{\eta})^{-1}$);
- $A = (-\infty, \infty)$, $\mathcal{Y} = \{0, 1, \ldots, N\}$;
- $y = y$;
- expectation parameter $\mu = N\pi = N(1+e^{-\eta})^{-1}$;
- $\psi(\eta) = N \log(1+e^{\eta})$;
- variance function $V = N\pi(1-\pi)$ $\quad (= \mu(1-\mu/N))$;
- $g_0(y) = \binom{N}{y}$.

Homework 1.12 Show that the binomial has skewness and kurtosis

$$\gamma = \frac{1-2\pi}{\sqrt{N\pi(1-\pi)}} \quad \text{and} \quad \delta = \frac{1-6\pi(1-\pi)}{N\pi(1-\pi)}.$$

Homework 1.13 Notice that $A = (-\infty, \infty)$ does *not* include the cases $\pi = 0$ or $\pi = 1$. Why not?

Gamma and Chi-squared

The notation

$$Y \sim \lambda G_N \tag{1.32}$$

indicates that Y is a scaled gamma variable having positive scale factor λ, density

$$g(y) = \frac{y^{N-1} e^{-y/\lambda}}{\lambda^N \Gamma(N)}, \qquad \text{for } \mathcal{Y} = (0, \infty), \tag{1.33}$$

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

and N positive, fixed, and known. With λ the unknown parameter, this is a one-parameter exponential family,

$$\eta = -\frac{1}{\lambda}; \quad \mu = N\lambda = -\frac{N}{\eta};$$
$$V = \frac{N}{\eta^2} = \frac{\mu^2}{N} = N\lambda^2; \tag{1.34}$$
$$\psi = -N\log(-\eta); \quad \gamma = \frac{2}{\sqrt{N}}; \quad \delta = \frac{6}{N}.$$

Situation (1.32) is denoted $Y \sim \text{Gamma}(N, \lambda)$, or sometimes $Y \sim \text{Gamma}(N, 1/\lambda)$, where $r = 1/\lambda$ is the *rate* parameter. *Additivity*: if $Y_i \overset{\text{ind}}{\sim} \lambda G_{N_i}$ for $i = 1, \ldots, K$, then $\sum_1^K Y_i \sim \lambda G_{\sum N_i}$.

Homework 1.14 Derive the skewness and kurtosis γ and δ.

By definition, a chi-squared random variable with m degrees of freedom is twice a gamma with $N = m/2$, written as

$$Y \sim \chi^2_m = 2G_{m/2}.$$

Chi-squared distributions apply to estimates of variance from normal observations. If

$$x_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2), \qquad \text{for } i = 1, \ldots, m,$$

then

$$\hat{\sigma}^2 = \sum_1^m \frac{x_i^2}{m} \sim \frac{\sigma^2 \chi^2_m}{m}. \tag{1.35}$$

In this case $\hat{\sigma}^2$ is an unbiased estimate of σ^2, having mean, standard deviation, skewness, and kurtosis, in the notation of (1.17),

$$\hat{\sigma}^2 \sim \left[\sigma^2, \frac{\sigma^2}{\sqrt{m/2}}, \frac{2}{\sqrt{m/2}}, \frac{12}{m}\right]. \tag{1.36}$$

Chi-squared distributions entered statistics from the world of 19th century physics. The *Maxwell–Boltzmann* distribution describes the velocity of gas molecules. Individual molecules are assumed to have independent, normal signed speeds in three dimensions,

$$v_1, v_2, v_3 \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2),$$

in which case their velocity $v = (v_1^2 + v_2^2 + v_3^2)^{1/2}$ is distributed as

$$v \sim \sigma\sqrt{\chi^2_3},$$

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

a scaled "chi" distribution with three degrees of freedom. Boltzmann's theory of kinetic gases says that $\sigma = k_B T/\text{mass}$, where T is temperature, mass is the molecule's mass, and k_B is Boltzmann's constant. A surprising fact says that v is typically in the range of a thousand miles per hour at room temperature, something to contemplate as you relax on a nice summer's day.

Homework 1.15 Derive the Maxwell–Boltzmann density $f(v)$.

The Negative Binomial Distribution

A coin with probability of heads θ is flipped until exactly k heads are observed. Let $Y = \{\#\,\text{tails observed}\}$. It has density

$$\begin{aligned} g(y) &= \binom{y+k-1}{k-1}(1-\theta)^y\theta^k \\ &= \binom{y+k-1}{k-1}e^{[\log(1-\theta)]y+k\log\theta}, \end{aligned} \tag{1.37}$$

and sample space $\mathcal{Y} = \{0, 1, 2, \ldots\}$. This is a one-parameter exponential family with

$$\begin{gathered} \eta = \log(1-\theta), \quad \psi(\eta) = -k\log(1-e^\eta), \\ \mu = k\frac{1-\theta}{\theta}, \quad V = \frac{\mu}{\theta}. \end{gathered} \tag{1.38}$$

The negative binomial can be thought of as an overdispersed version of the Poisson. For a given value of μ, its variance V is always greater than the Poisson variance μ – as illustrated in Figure 1.3 – making the negative binomial useful for situations with count data where there are reasons to suspect overdispersion.[4] If $k \to \infty$ with θ changing to keep μ fixed, then $Y \to \text{Poi}(\mu)$.

For a given value of μ, the ratio $V/\mu = 1 + \mu/k$ decreases to 1 as

$$\phi \equiv 1/k,$$

with ϕ often called the *dispersion parameter*; distribution (1.37) is denoted by $\text{NB}(\mu, \phi)$. A minor disadvantage of the negative binomial family is that it is not exponential in the dispersion parameter ϕ (or in k). Section 3.9 discusses the "double Poisson" distribution, which is a full two-parameter exponential family of overdispersed Poisson distributions.

[4] As used in the popular algorithm DESeq2 for the analysis of genetic sequence data (Love et al., 2014).

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

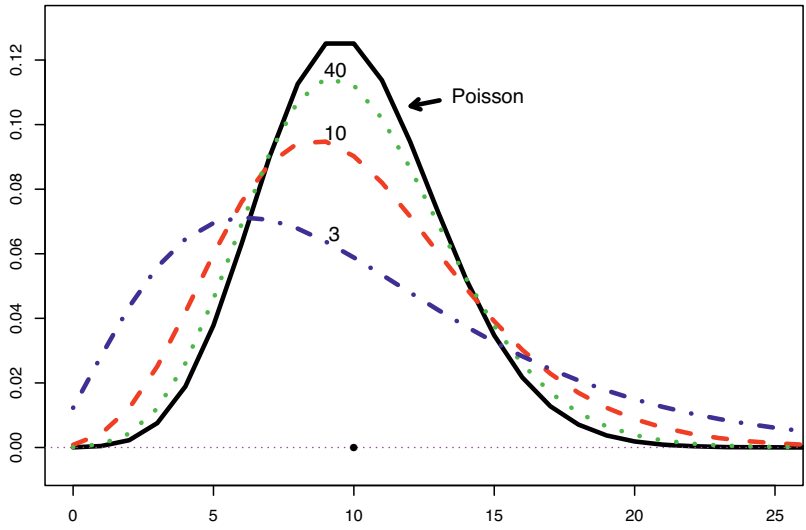


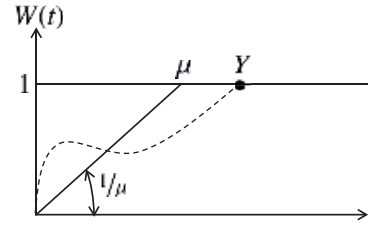
Figure 1.3 Poisson(10) and negative binomials with mean 10, $k = 40, 10$, and 3.

Homework 1.16 Notice that $\psi = k\psi_0$ where ψ_0 is the CGF for $k = 1$. Give a simple explanation for this. How does this affect the expressions for μ and V? What about the expression for the skewness γ?

Homework 1.17 Show that if μ is drawn from a gamma distribution and $y \sim \text{Poi}(\mu)$ is observed, then, marginally, y has a negative binomial distribution. (Another name for the negative binomial is "gamma-Poisson".)

Inverse Gaussian

Let $W(t)$ be a *Wiener process* with drift $1/\mu$, that is, $W(t) \sim \mathcal{N}(t/\mu, t)$ with $\text{Cov}[W(t), W(t+d)] = t$. Define Y as the first passage time to $W(t) = 1$. Then it turns out that Y has the "inverse Gaussian" or Wald density



$$g(y) = \frac{1}{\sqrt{2\pi y^3}}\, e^{-\frac{(y-\mu)^2}{2\mu^2 y}}$$

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

(y and μ in $\mathcal{R}'$). This is an exponential family with:

- $\eta = -(2\mu^2)^{-1}$;
- $\psi = -\sqrt{-2\eta}$;
- $V = \mu^3$.

We might have called the negative binomial the "inverse binomial" instead, since its definition is a discrete version of the first-passage time construction in the diagram above.

Reference Johnson and Kotz (1970a), *Continuous Univariate Distributions Vol. 1*, Chapter 15.

Homework 1.18 Show $Y \sim [\mu, \mu^{3/2}, 3\mu^{1/2}, 15\mu]$ as the mean, standard deviation, skewness, and kurtosis, respectively.

One way to characterize exponential families is by how the variance V behaves as a function of the expectation μ. Table 1.2 shows V_μ equaling various powers of μ. There is no family with $V_\mu \propto \mu^{1.5}$, say, but quasi-likelihood methods (Section 3.9) let us act as if there is. This was a tactic used in the early history of generalized linear models but will not be explored here.

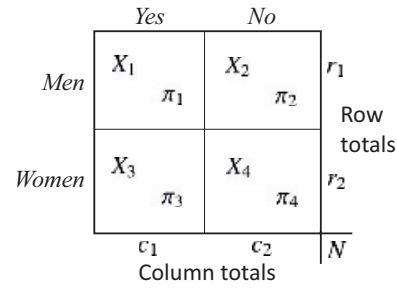
Table 1.2 *V as a function of μ in some one-parameter exponential families.*

	Normal	Poisson	Scaled Gamma	Inverse Normal
$V_\mu \propto$	μ^0	μ	μ^2	μ^3

2×2 *Tables (Tilted Hypergeometric Family)*

The diagram at right shows a hypothetical two-by-two table comparing men's and women's responses to a yes/no question, perhaps "Have you attended a ballet within the last five years?" The N respondents have provided counts

$$X = (X_1, X_2, X_3, X_4)$$



for the four possible categories – (man, yes); (man, no); (woman, yes);

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

(woman, no) – labeled 1, 2, 3, 4 as shown. As discussed in Section 2.9, X has a four-category multinomial distribution, with true probabilities

$$\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$$

for the four categories.

Perhaps we would like to answer the question, "Do men and women differ in their ballet attendance?" The question can be symmetrically stated in terms of the *log odds* parameter

$$\theta = \log\left(\frac{\pi_1/\pi_2}{\pi_3/\pi_4}\right), \tag{1.39}$$

as a test of the null hypothesis H_0: $\theta = 0$ (i.e., men and women have the same probability of answering yes). Karl Pearson suggested the chi-squared test of H_0 in the early 1900s. *Fisher's exact test* of H_0 (Fisher, 1925) leads to a one-parameter exponential family, the "tilted hypergeometric".

Testing H_0 in terms of X seems awkward since X is four-dimensional. Fisher suggested conditioning the 2×2 table on its marginal sums (r_1, r_2, c_1, c_2) or equivalently conditioning on (N, r_1, c_1), since $r_2 = N - r_1$ and $c_2 = N - c_1$. With the marginals fixed, we need only know x_1 to fill in the 2×2 table.

Fisher's suggestion was to base the test of H_0: $\theta = 0$ on the conditional distribution of x_1 given (r_1, r_2, c_1, c_2). Under H_0, x_1 has the *hypergeometric distribution*

$$g_0(x_1 \mid r_1, r_2, c_1, c_2) = \binom{r_1}{x_1}\binom{r_2}{c_1 - x_1}\Big/\binom{N}{c_1}, \tag{1.40}$$

for x_1 in the set of possible integer values consistent with the marginal constraints

$$\max(0, c_1 - r_2) \le x_1 \le \min(c_1, r_1); \tag{1.41}$$

x_1 has

$$\text{expectation} = \frac{r_1 c_1}{N} \quad \text{and} \quad \text{variance} = \frac{r_1 r_2 c_1 c_2}{N^2(N-1)}.$$

Conditioning has the effect of reducing a four-dimensional testing problem to one dimension, at the expense of losing whatever information the marginal totals have on H_0, not much in most situations.

What happens if H_0 is *not* true? Beginning with the probabilities $(\pi_1, \pi_2, \pi_3, \pi_4)$ for the four-category multinomial pictured above, Section 2.8 and

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

Section 2.9 show that the conditional distribution of x_1 given (r_1, r_2, c_1, c_2) forms a one-parameter exponential family

$$g_\theta(x_1) = \frac{g_0(x_1)e^{\theta x_1}\binom{N}{c_1}}{C(\theta)}, \tag{1.42}$$

where

$$C(\theta) = \sum_{x_1} \binom{r_1}{x_1}\binom{r_2}{c_1 - x_1} e^{\theta x_1},$$

x_1 as in (1.41). That is, we tilt the hypergeometric distribution (1.40) according to $e^{\theta x_1}$, θ the log odds parameter (1.39), and then renormalize to make g_θ sum to 1.

Reference Lehmann and Romano (2005), *Testing Statistical Hypotheses*, Section 4.5.

The Ulcer Data

A clinical trial was held in 41 cities comparing a new ulcer surgery, the Treatment, with the standard surgery, or Control. The 2×2 table at right shows the outcomes for city 14.[5] The obvious estimate of θ is

$$\hat\theta = \log\left(\frac{9/12}{7/17}\right) = 0.600.$$

	Success	*Failure*	
Treatment	**9**	**12**	**21**
Control	**7**	**17**	**24**
	16	**29**	**45**

Figure 1.4 graphs the likelihood, i.e., expression (1.42) as a function of θ, with the data held fixed (normalized so that $\max\{L(\theta)\} = 1$).

Homework 1.19 (a) Compute the likelihood $L(\theta) = g_\theta(\hat\theta)$ numerically and verify that it is maximized at $\hat\theta = 0.600$.

(b) Verify numerically that

$$-\frac{d^2 \log L(\theta)}{d\theta^2}\bigg|_{\hat\theta} = 2.56.$$

(Note that the quantity on the left is sometimes called "the observed Fisher information", as discussed in Part 4.)

(c) Using this result, guess the variance of $\hat\theta$. *Hint*: Think of the same calculation if $\hat\theta \sim \mathcal{N}(\theta, \sigma^2)$.

[5] The results for all 41 cities are in the file named `ulcdata`.

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

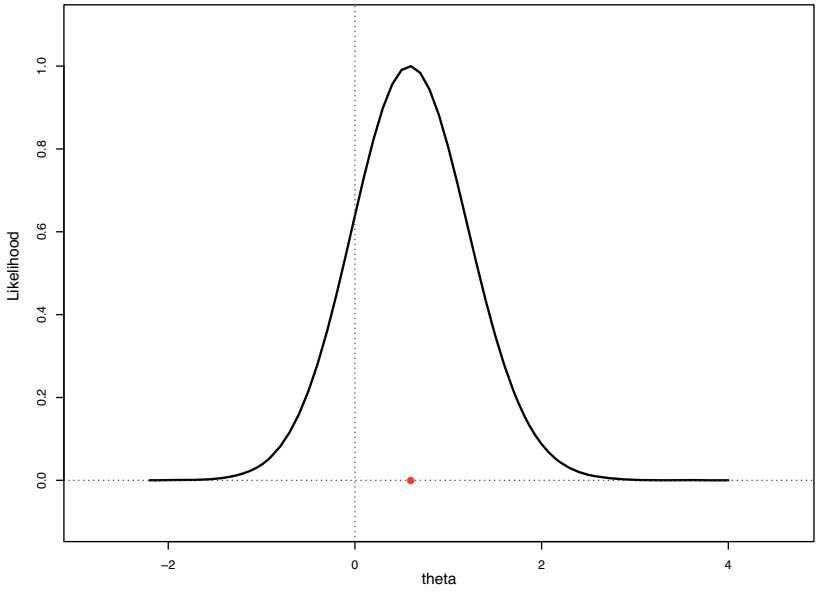


Figure 1.4 `ulcdata` #14; likelihood function for log odds ratio θ; max at $\theta = 0.600$; $-\ddot{l} = 2.56$.

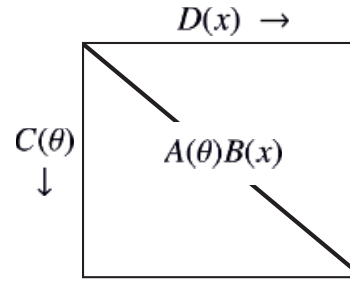
The Structure of One-parameter Exponential Families

Suppose $f_\theta(x)$, θ and x possibly vectors, is a family of densities satisfying

$$\log f_\theta(x) = A(\theta)B(x) + C(\theta) + D(x),$$

with A, B, C, D real. Then $\{f_\theta(x)\}$ is a one-parameter exponential family with:

- $\eta = A(\theta)$;
- $y = B(x)$;
- $\psi = -C(\theta)$;
- $\log g_0 = D(x)$.



A two-way table of $\log f_\theta(x)$ would have additive components $C(\theta) + D(x)$, and an interaction term $A(\theta)B(x)$.

Homework 1.20 I constructed a 14×9 matrix P with ijth element

$$p_{ij} = \text{Bi}(x_i, \theta_j, 13),$$

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

the binomial probability of x_i for probability θ_j, sample size $n = 13$, where

$$x_i = i, \qquad \text{for } i = 0, 1, \ldots, 13,$$
$$\theta_j = 0.1, \ldots, 0.9.$$

Then I calculated the singular value decomposition (R function `svd`) of $\log P$. How many non-zero singular values did I see? (Equivalently, what was the rank of $\log P$?)

1.6 Bayes Families

Exponential families play a major role in Bayesian calculations. Suppose we observe $Y = y$ from a one-parameter exponential family

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

where η itself has a prior density

$$\eta \sim \pi(\eta)$$

with respect to Lebesgue measure on a set $\mathcal{A}$. Bayes rule

$$\pi(\eta \mid y) = \pi(\eta) g_\eta(y)/g(y), \tag{1.43}$$

yields the posterior density of η given y, $\pi(\eta \mid y)$, where $g(y)$ is the *marginal density*

$$g(y) = \int_{\mathcal{A}} \pi(\eta) g_\eta(y)\, d\eta.$$

(Note that here $g_\eta(y)$ is the *likelihood function*, with y fixed and η varying.) Putting all this together gives

$$\pi(\eta \mid y) = e^{y\eta - \log(g(y)/g_0(y))} \left(\pi(\eta) e^{-\psi(\eta)}\right). \tag{1.44}$$

We recognize this as a one-parameter exponential family with:

- natural parameter "η" = y;
- sufficient statistic "y" = η;
- CGF $\psi = \log(g(y)/g_0(y))$;
- carrier $g_0 = \pi(\eta) e^{-\psi(\eta)}$.

Homework 1.21 (a) Show that prior $\pi(\eta)$ for η corresponds to prior $\pi(\eta)/V_\eta$ for μ. (b) What is the posterior density $\pi(\mu \mid y)$ for μ?

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

Conjugate Priors

Certain choices of $\pi(\eta)$ yield particularly simple forms for $\pi(\eta \mid y)$ or $\pi(\mu \mid y)$, and these are called *conjugate priors*. They play an important role in modern Bayesian applications. As an example, the conjugate prior for Poisson is the gamma.

Homework 1.22 (a) Suppose $y \sim \text{Poi}(\mu)$ and $\mu \sim mG_\nu$, a scale multiple of a gamma with ν degrees of freedom. Show that

$$\mu \mid y \sim \frac{m}{m+1} G_{y+\nu}.$$

(b) Also show that

$$E\{\mu \mid y\} = \frac{m}{m+1}\, y + \frac{1}{m+1}(m\nu)$$

(compared to $E\{\mu\} = m\nu$ *a priori*, so $E\{\mu \mid y\}$ is a linear combination of y and $E\{\mu\}$).

(c) What is the posterior distribution of μ having observed $y_1, y_2, \ldots, y_n \overset{\text{iid}}{\sim} \text{Poi}(\mu)$?

Diaconis and Ylvisaker (1979) provide a general formulation of conjugacy: if we observe

$$y_1, \ldots, y_n \overset{\text{iid}}{\sim} g_\eta(y) = e^{\eta y - \psi(\eta)} g_0(y),$$

the conjugate prior for μ with respect to Lebesgue measure is

$$\pi_{n_0, y_0}(\mu) = c_0 e^{n_0[\eta y_0 - \psi(\eta)]} \Big/ V_\eta, \tag{1.45}$$

where y_0 is notionally the average of n_0 hypothetical prior observations of y (c_0 is the constant making $\pi_{n_0,y_0}(\mu)$ integrate to 1). Prior (1.45) yields a particularly convenient posterior density for μ:

Theorem 1.2

$$\pi(\mu \mid y_1, \ldots, y_n) = \pi_{n_+, y_+}(\mu),$$

where

$$n_+ = n_0 + n \quad \textit{and} \quad y_+ = \frac{(n_0 y_0 + \sum_1^n y_i)}{n_+}.$$

Moreover,

$$E\{\mu \mid y_1, \ldots, y_n\} = y_+.$$

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

The first result, which justifies the notional interpretation of n_0 and y_0, is almost immediate, but the second is more involved and won't be verified here.

Homework 1.23 Make the explicit connections between Theorem 1.2 and Homework 1.22.

Binomial Case

Suppose $y_1, \ldots, y_n$ are independent *Bernoulli* observations,

$$y_i = \begin{cases} 0 & \text{with probability } 1-\pi \\ 1 & \text{with probability } \pi, \end{cases}$$

so $y = \sum_{i=1}^{n} y_i$ is binomial, $y \sim \text{Bi}(n, \pi)$. As in Section 1.5, y is the sufficient statistic of a one-parameter exponential family having $\eta = \log[\pi/(1-\pi)]$ and $\mu = n\pi$.

Homework 1.24 Remembering that μ equals π in the binomial case, show that the conjugate family (1.45) is

$$\pi_{n_0, y_0}(\pi) = c_0 \pi^{s_1 - 1}(1-\pi)^{s_0 - 1}, \tag{1.46}$$

where (s_1, s_0) are the number of 1s and 0s in the hypothetical prior sample (a "beta" distribution; see Part 2). Theorem 1.2 gives posterior expectation

$$E\{\pi \mid y_1, \ldots, y_n\} = \frac{s_1 + y}{n_0 + n}. \tag{1.47}$$

The interpretation of the prior (1.46) is as a hypothetical binomial sample of size n_0, with observed number $s_1 = n_0 y_0$ of successes. Current Bayes practice favors using small amounts of hypothetical prior information, in the binomial case maybe $s_1 = 1$ and $n_0 = 2$ (so $y_0 = 1/2$), giving

$$\hat{\theta} = \frac{1+y}{2+n},$$

pulling $\hat{\theta}$ a little toward $1/2$, compared to the MLE y/n.

Tweedie's Formula

Equation (1.44) gave

$$\pi(\eta \mid y) = e^{y\eta - \lambda(y)} \pi_0(y),$$

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

where

$$\pi_0(y) = \pi(\eta)e^{-\psi(\eta)} \quad \text{and} \quad \lambda(y) = \log \frac{g(y)}{g_0(y)},$$

with $g(y)$ the marginal density of y. Define

$$l(y) = \log g(y) \quad \text{and} \quad l_0(y) = \log g_0(y).$$

We can now differentiate $\lambda(y)$ with respect to y to get the posterior moments (and cumulants) of η given y,

$$E\{\eta \mid y\} = \lambda'(y) = l'(y) - l_0'(y),$$
$$\text{Var}\{\eta \mid y\} = \lambda''(y) = l''(y) - l_0''(y).$$

Homework 1.25 Suppose $y \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 known, where μ has prior density $\pi(\mu)$. Show that the posterior mean and variance of μ given y is

$$\mu \mid y \sim \left[y + \sigma^2 l'(y), \sigma^2\left(1 + \sigma^2 l''(y)\right)\right]. \tag{1.48}$$

Reference Efron (2011), "Tweedie's formula and selection bias", *JASA* 1602–1614.

1.7 Empirical Bayes Inference

Bayes rule, when applicable, provides a wonderfully satisfying path for statistical inference. The catch is that the prior density, $\pi(\eta)$ in (1.43), is most often unknown in typical applications. A surprising development, post-World War II, was that when simultaneously dealing with many similar inference problems, the data itself may provide an estimate of the prior. This is the *empirical Bayes* concept, an approach where exponential families have played a central part.

Table 1.3 displays the *insurance data*, a summary of one year's record of claims from a European auto insurance company: 7840 of the 9461 policyholders made no claims during the year, 1317 made a single claim, 239 made two claims each, going on to the one person, possibly a very bad driver, who made seven claims. In the notation of Table 1.3,

$$y_x = \#\{\text{policyholders who made } x \text{ claims}\}, \tag{1.49}$$

for $x = 0, 1, \ldots, 7$.

Suppose the company wants to know how many accident claims it can expect next year from a driver with x claims this year. A commonly used

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

Table 1.3 *Insurance data counts and claims, and two empirical Bayes estimates of future claims per driver.*

Claims x	0	1	2	3	4	5	6	7
Counts y_x	7840	1317	239	42	14	4	4	1
Robbins' formula	0.168	0.363	0.527	1.33	1.43	6.00	1.25	
Gamma MLE	0.164	0.398	0.632	0.87	1.10	1.34	1.57	

actuarial model assumes that each driver has a Poisson distribution of annual accidents, $g_\mu(x) = e^{-\mu}\mu^x/x!$, μ varying from driver to driver, and with μ having some prior density $\pi(\mu)$,

$$\pi(\mu) \longrightarrow \mu \longrightarrow x \sim \text{Poi}(\mu). \tag{1.50}$$

The insurance company would like to know the Bayes posterior expectation of μ given x,

$$E\{\mu \mid x\} = \int_0^\infty \mu\pi(\mu \mid x)\, d\mu = \frac{\int_0^\infty [e^{-\mu}\mu^{x+1}/x!]\pi(\mu)\, d\mu}{\int_0^\infty [e^{-\mu}\mu^x/x!]\pi(\mu)\, d\mu}, \tag{1.51}$$

but unless they know the prior $\pi(\mu)$ this is out of reach.

Here is where empirical Bayes makes its appearance. Notice that we can rewrite (1.51) as

$$\begin{aligned} E\{\mu \mid x\} &= \frac{(x+1)\int_0^\infty \left[e^{-\mu}\mu^{x+1}/(x+1)!\right]\pi(\mu)\, d\mu}{\int_0^\infty [e^{-\mu}\mu^x/x!]\pi(\mu)\, d\mu} \\ &= \frac{(x+1)g(x+1)}{g(x)}, \end{aligned} \tag{1.52}$$

where $g(x)$ is the marginal density of x,

$$g(x) = \int_0^\infty \frac{e^{-\mu}\mu^x}{x!}\pi(\mu)\, d\mu.$$

Homework 1.26 Give a careful derivation of (1.51)–(1.52).

We don't know $g(\cdot)$ either but, as the marginal distribution of x, it has an obvious estimate in terms of the counts y_x,

$$\hat{g}(x) = \frac{y_x}{N} \qquad \left(N = \sum y_x\right),$$

$N = 9461$ here; (1.52) leads to *Robbins' formula* (Robbins, 1956)

$$\widehat{E}\{\mu \mid x\} = (x+1)\frac{y_{x+1}}{y_x}. \tag{1.53}$$

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

The third line of Table 1.3 shows $\widehat{E} = \{\mu \mid x = 0\} = 0.168$, so last year's perfect driver can expect about one-sixth of a claim this year, and so on up the table.

Small values of y_x make $\widehat{E}\{\mu \mid x\}$ erratic near the right end of Table 1.3. We can get a less variable estimate of $E\{\mu \mid x\}$ by assuming a parametric model for $\pi(\mu)$ in (1.52). A natural choice is the conjugate prior, the scaled Gamma,

$$\pi(\mu) = \frac{\mu^{a-1}e^{-\mu/b}}{b^a\Gamma(a)} \qquad (\mu > 0), \tag{1.54}$$

that is, $\mu \sim bG_a$ (1.32) (making the marginal $g(x)$ negative binomial, as in Homework 1.17). Choosing (a, b) to be the maximum likelihood estimates based on the counts $y_0, y_1, \ldots, y_n$, and substituting $\hat{\pi}(\mu)$ for $\pi(\mu)$ in (1.52), gave the estimates $\widehat{E}\{\mu \mid x\}$ in the last row of Table 1.3.

Homework 1.27 What is Robbins' formula if x is binomial rather than Poisson in (1.50)?

Robbins' formula applies to Poisson sampling models (1.50). Suppose instead we have a normal model,

$$\pi(\mu) \longrightarrow \mu \longrightarrow y \sim \mathcal{N}(\mu, \sigma^2), \tag{1.55}$$

σ^2 known. Tweedie's formula (1.48) says that

$$E\{\mu \mid y\} = y + \sigma^2 l'(y), \tag{1.56}$$

with $l'(y)$ the derivative of the log marginal density $g(y)$; y is the MLE of μ, the usual frequentist (non-Bayesian) estimate of μ, so (1.56) amounts to

$$E\{\mu \mid y\} = \text{MLE} + \text{Bayes correction}. \tag{1.57}$$

Empirical Bayes methods come into play when we have many realizations of (1.55) to deal with at once, say $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ for $i = 1, \ldots, N$, where we can use all the data to estimate the Bayes correction for each case. In other words, we can enjoy the advantages of Bayesian estimation without the requisite prior knowledge. A microarray example follows next.

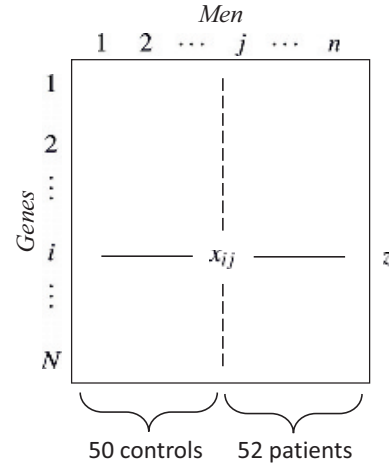
https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

In a study of prostate cancer, $n = 102$ men each had his genetic expression level x_{ij} measured on $N = 6033$ genes,

$$x_{ij} = \begin{cases} i = 1, \ldots, N & \text{genes} \\ j = 1, \ldots, n & \text{men.} \end{cases}$$

There were:

- $n_1 = 50$ healthy controls;
- $n_2 = 52$ prostate cancer patients.



For gene$_i$, let t_i equal a two-sample t statistic comparing patients with controls and

$$z_i = \Phi^{-1}[F_{100}(t_i)] \qquad (F_{100} \text{ CDF of a } t_{100} \text{ distribution});$$

z_i is a *z-value*, i.e., a statistic having a $\mathcal{N}(0, 1)$ distribution under the null hypothesis that there is no difference in gene$_i$ expression between patients and controls.[6]

A reasonable model for the z_is is

$$z_i \sim \mathcal{N}(\delta_i, 1),$$

where δ_i is the *effect size* for gene i. (In terms of our previous notation, z and δ are playing the roles of y and μ.) The investigators were looking for genes with large values of δ_i, either positive or negative. Figure 1.5 shows the histogram of the 6033 z_i values. It is a little wider than a $\mathcal{N}(0, 1)$ density, suggesting some non-null ($\delta_i \neq 0$) genes. Which ones and how much?

An empirical Bayes analysis proceeds in four steps:

1. Compute $z_1, \ldots, z_N$; $N = 6033$.
2. Fit a smooth parametric estimate $\hat{g}(z)$ to histogram (details in Part 2).
3. Numerically differentiate $\hat{l}(z) = \log \hat{g}(z)$ to get $\hat{l}'(x)$.
4. Estimate $E\{\delta_i \mid z_i\}$ by

$$E\{\delta_i \mid z_i\} = z_i + \hat{l}'(z_i), \tag{1.58}$$

for $i = 1, \ldots, N$. Notice that *all* of the z-values play a role in the estimation of any one δ_i, through their part in estimating $\hat{g}(\cdot)$.

[6] Since $t_i \sim F_{100}$ under the null hypothesis that $\delta_i = 0$, the "probability integral transformation" $F_{100}(t_i)$ has a uniform distribution over $[0, 1]$; then the inverse transformation $Z = \Phi^{-1}(F_{100}(t_i)) \sim \mathcal{N}(0, 1)$.

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

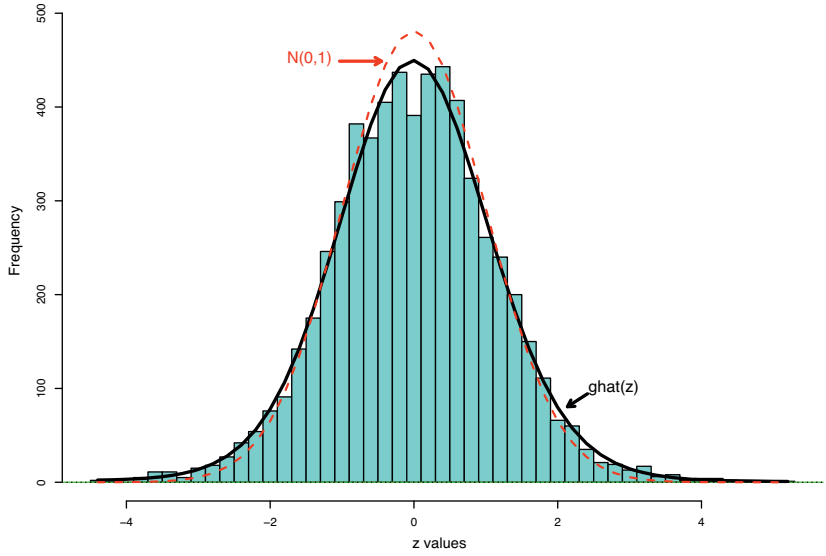


Figure 1.5 *Prostate data* microarray study. 6033 z-values; heavy curve is $\hat{g}(z)$ from GLM fit; dashed line is $\mathcal{N}(0, 1)$.

Figure 1.6 shows $\widehat{E}\{\delta \mid z\}$. It is near zero ("nullness") for $|z| \leq 2$. At $z = 3$, $\widehat{E}\{\delta \mid z\} = 1.31$. At $z = 5.29$, the largest observed z_i value (gene #610), $E\{\delta \mid z\} = 3.94$. Even though each z_i is unbiased for its δ_i it isn't true that $z_{i_{\max}}$, $i_{\max} = 610$, is unbiased for $\delta_{i_{\max}}$. The Bayes correction in (1.57) is quite negative, an example of *selection bias* or *the winner's curse*: being largest in a group of unbiased estimates involves luck as well as a genuinely large value of δ, and that's what empirical Bayes is accounting for in Figure 1.6.

The purpose of a large-scale study like that for the prostate data is to weed out the great proportion, say π_0, of null genes, those having $\delta_i = 0$, in order to focus attention on those with large effect sizes. The *local false discovery rate* "fdr(z_i)" is the posterior probability of nullness,

$$\text{fdr}(z_i) = \Pr\{\delta_i = 0 \mid z_i\}.$$

Homework 1.28 (a) Show that

$$\text{fdr}(z) = \pi_0 g_0(z)/g(z), \tag{1.59}$$

with π_0 the prior probability of nullness, $g(z)$ the marginal density of z,

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

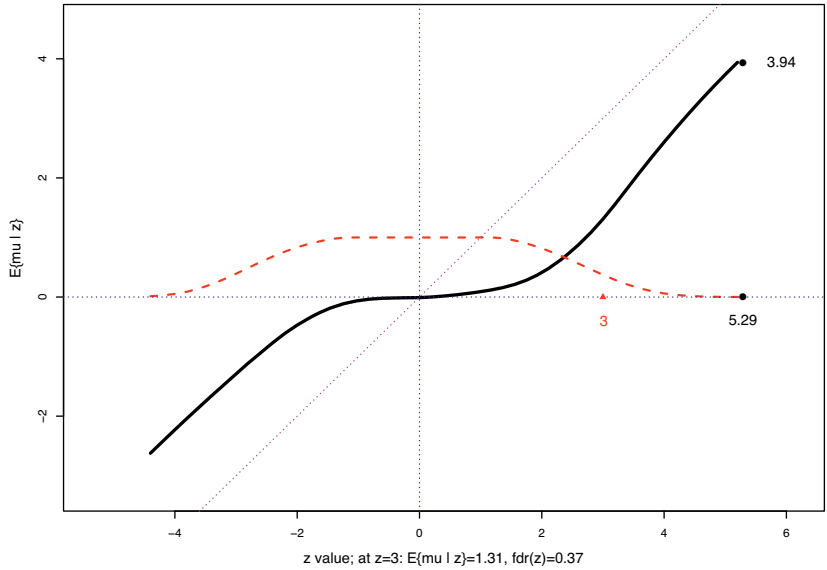


Figure 1.6 Tweedie estimate of $E\{\mu \mid z\}$, prostate study. Dashed curve is estimated local false discovery rate fdr(z).

and

$$g_0(z) = \frac{e^{-z^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}.$$

(b) If $\pi(\delta)$ is the prior density of δ, as in (1.45), show that

$$E\{\delta \mid z\} = \frac{d}{dz} \log \text{fdr}(z).$$

The red dashed curve in Figure 1.6 is $\widehat{\text{fdr}}(z) = g_0(z)/\hat{g}(z)$, taking $\pi_0 = 1$ as an upper bound, usually a close one in practice; $\widehat{\text{fdr}}(3) = 0.37$ for the prostate data, so even $z = 3$ standard deviations away from 0, there is still substantial probability of $\delta = 0$.

1.8 Deviance and Hoeffding's Formula

Traditional normal theory methods depend on notions of Euclidean distance. *Deviance* is an analogue of Euclidean distance applying to exponen-

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

tial families. By definition the deviance $D(g_1, g_2)$ between g_1 and g_2 is

$$\begin{aligned} D(g_1, g_2) &= 2E_{g_1}\left\{\log\frac{g_1(y)}{g_2(y)}\right\} \\ &= 2\int_{\mathcal{Y}} g_1(y)\left(\log\frac{g_1(y)}{g_2(y)}\right) m(dy). \end{aligned} \tag{1.60}$$

We will also write $D(\eta_1, \eta_2)$, $D(\mu_1, \mu_2)$, or just $D(1, 2)$ as convenient. If $g_1(y)$ and $g_2(y)$ are densities $g_{\eta_1}(y)$ and $g_{\eta_2}(y)$ in an exponential family, then it is easy to verify that

$$D(1, 2) = 2\left[(\eta_1 - \eta_2)\mu_1 - (\psi(\eta_1) - \psi(\eta_2))\right].$$

Homework 1.29 Show that $D(1, 2) \geq 0$, with strict inequality unless the two densities are identical.

Note In general, $D(1, 2) \neq D(2, 1)$.

The "Kullback–Leibler distance", using an older name for the same idea, equals $D(\eta_1, \eta_2)/2$. Information theory uses "mutual information" for $D(f(x, y), f(x)f(y))/2$, where $f(x, y)$ is a bivariate density and $f(x)$ and $f(y)$ its marginals.

Homework 1.30 Verify these formulas for the deviance.

$$\text{Normal } Y \sim \mathcal{N}(\mu, 1)\text{: } D(\mu_1, \mu_2) = (\mu_1 - \mu_2)^2$$

(This motivates the factor 2 in (1.60).)

$$\text{Poisson } Y \sim \text{Poi}(\mu)\text{: } D(\mu_1, \mu_2) = 2\mu_1\left[\log\left(\frac{\mu_1}{\mu_2}\right) - \left(1 - \frac{\mu_2}{\mu_1}\right)\right]$$

$$\begin{aligned} \text{Binomial } Y \sim \text{Bi}(N, \pi)\text{: } D(\pi_1, \pi_2) &= 2N\left[\pi_1\log\left(\frac{\pi_1}{\pi_2}\right)\right. \\ &\qquad \left. +(1 - \pi_1)\log\left(\frac{1 - \pi_1}{1 - \pi_2}\right)\right] \end{aligned}$$

$$\begin{aligned} \text{Gamma } Y \sim \lambda G_N\text{: } D(\lambda_1, \lambda_2) &= 2N\left[\log\left(\frac{\lambda_2}{\lambda_1}\right) + \left(\frac{\lambda_1}{\lambda_2} - 1\right)\right] \\ &= 2N\left[\log\left(\frac{\mu_2}{\mu_1}\right) + \left(\frac{\mu_1}{\mu_2} - 1\right)\right] \end{aligned}$$

$$\text{Negative binomial (1.37)}\text{: } D(\theta_1, \theta_2) = k\left[\left(\frac{1 - \theta_1}{\theta_1}\right)\log\left(\frac{1 - \theta_1}{1 - \theta_2}\right) + \log\left(\frac{\theta_1}{\theta_2}\right)\right]$$

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

Hoeffding's Formula

An exponential family of densities $\mathcal{G} = \{g_\eta(y) = \exp(\eta y - \psi(\eta))\}$ can be rewritten in a form that is particularly helpful in discussing maximum likelihood estimation:

Lemma 1.3 (Hoeffding, 1965) *Let $\hat\eta$ be the MLE of η and $\hat\mu = y$ the MLE of μ. Then*

$$g_\eta(y) = g_{\hat\eta}(y)e^{-D(\hat\eta,\eta)/2} \tag{1.61}$$

or, reparameterizing $\mathcal{G}$ in terms of μ (and recalling that $\hat\mu = y$),

$$g_\mu(y) = g_y(y)e^{-D(y,\mu)/2}. \tag{1.62}$$

This says that a plot of the log likelihood $\log g_\mu(y)$ declines from its maximum at $\mu = y$ according to the deviance,

$$\log g_\mu(y) = \log g_y(y) - \frac{D(y,\mu)}{2}.$$

In our applications of the deviance, the first argument will always be the data, the second a proposed value of the unknown parameter.

Proof The deviance in an exponential family is

$$\begin{aligned}\frac{D(\eta_1,\eta_2)}{2} = E_{\eta_1}\left\{\log\frac{g_{\eta_1}(y)}{g_{\eta_2}(y)}\right\} &= E_{\eta_1}\{(\eta_1-\eta_2)y - \psi(\eta_1) + \psi(\eta_2)\}\\ &= (\eta_1-\eta_2)\mu_1 - \psi(\eta_1) + \psi(\eta_2),\end{aligned}$$

and

$$\frac{g_\eta(y)}{g_{\hat\eta}(y)} = \frac{e^{\eta y-\psi(\eta)}}{e^{\hat\eta y-\psi(\hat\eta)}} = e^{(\eta-\hat\eta)y-\psi(\eta)+\psi(\hat\eta)} = e^{(\eta-\hat\eta)\hat\mu-\psi(\eta)+\psi(\hat\eta)}.$$

Taking $\eta_1 = \hat\eta$ and $\eta_2 = \eta$ above, this last is $e^{-D(\hat\eta,\eta)/2}$. ■

Repeated Sampling

If $\boldsymbol{y} = (y_1, \ldots, y_n)$ is an i.i.d. sample from $g_\eta(\cdot)$ then the deviance based on $\boldsymbol{y}$, say $D^{(n)}(\eta_1,\eta_2)$, is

$$\begin{aligned}D^{(n)}(\eta_1,\eta_2) &= 2E_{\eta_1}\left\{\log\frac{g^{(n)}_{\eta_1}(\boldsymbol{y})}{g^{(n)}_{\eta_2}(\boldsymbol{y})}\right\} = 2E_{\eta_1}\left\{\log\prod_{i=1}^n\frac{g_{\eta_1}(y_i)}{g_{\eta_2}(y_i)}\right\}\\ &= 2\sum_{i=1}^n E_{\eta_1}\left\{\log\frac{g_{\eta_1}(y_i)}{g_{\eta_2}(y_i)}\right\} = nD(\eta_1,\eta_2).\end{aligned}$$

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

(This shows up in the binomial, Poisson, gamma, and negative binomial cases of Homework 1.30.) Hoeffding's formula (1.61) applied to $\boldsymbol{y}$ is

$$g_\eta^{(n)}(\boldsymbol{y}) = g_{\hat\eta}^{(n)}(\boldsymbol{y}) e^{-nD(\hat\eta,\eta)/2}. \tag{1.63}$$

For η_2 near η_1, the deviance is related to the Fisher information $i_{\eta_1} = V_{\eta_1}$ (in a single observation y, for η_1 and at η_1):

$$D(\eta_1, \eta_2) = i_{\eta_1}(\eta_2 - \eta_1)^2 + O(\eta_2 - \eta_1)^3.$$

Proof

$$\begin{aligned}\frac{\partial}{\partial\eta_2} D(\eta_1,\eta_2) &= \frac{\partial}{\partial\eta_2} 2\left[(\eta_1-\eta_2)\mu_1 - (\psi(\eta_1)-\psi(\eta_2))\right] \\ &= 2(-\mu_1+\mu_2) \\ &= 2(\mu_2-\mu_1).\end{aligned}$$

Also

$$\frac{\partial^2}{\partial\eta_2^2} D(\eta_1,\eta_2) = 2\frac{\partial\mu_2}{\partial\eta_2} = 2V_{\eta_2}.$$

Therefore

$$\frac{\partial}{\partial\eta_2} D(\eta_1,\eta_2)\Big|_{\eta_2=\eta_1} = 0 \quad \text{and} \quad \frac{\partial^2}{\partial\eta_2^2} D(\eta_1,\eta_2)\Big|_{\eta_2=\eta_1} = 2V_{\eta_1},$$

so a Taylor expansion gives

$$D(\eta_1,\eta_2) = 2V_{\eta_1}\frac{(\eta_2-\eta_1)^2}{2} + O(\eta_2-\eta_1)^3. \qquad \blacksquare$$

Homework 1.31 What is $\partial^3 D(\eta_1,\eta_2)/\partial\eta_2^3$? Give an improved version of the relationship above.

Reference Efron (1978), "The geometry of exponential families", *Ann. Stat.* 362–376.

An Informative Picture

We know that $\psi(\eta)$ is a convex function of η since $\ddot\psi(\eta) = V_\eta > 0$. Figure 1.7 shows $\psi(\eta)$ passing through $(\eta_1, \psi(\eta_1))$ at slope $\mu_1 = \dot\psi(\eta_1)$. The difference between $\psi(\eta_2)$ and the linear bounding line $\psi(\eta_1) + (\eta_2 - \eta_1)\mu_1$ is $\psi(\eta_2) - \psi(\eta_1) + (\eta_1 - \eta_2)\mu_1 = D(\eta_1,\eta_2)/2$.

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

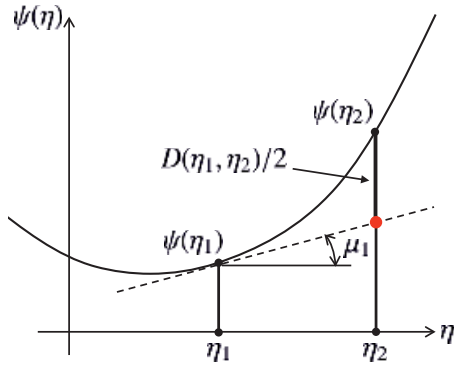


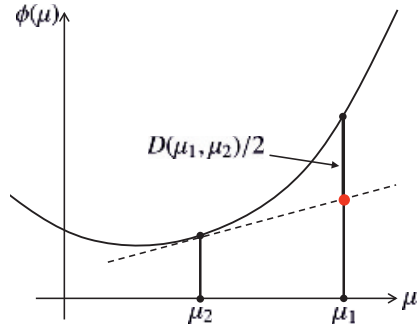
Figure 1.7 Convex function $\psi(\eta)$ passing through $(\eta_1, \psi(\eta_1))$ at slope $\mu_1 = \dot{\psi}(\eta_1)$.

Unlike our other results, Figure 1.7 depends on parameterizing the deviance as $D(\eta_1, \eta_2)$. A version that uses $D(\mu_1, \mu_2)$ depends on the *dual function* $\phi(y)$ to $\psi(y)$,

$$\phi(y) = \max_{\eta} \{\eta y - \psi(\eta)\}.$$

Homework 1.32 Show that:

(a) $\phi(\mu) = \eta\mu - \psi(\eta)$, where $\mu = \dot{\psi}(\eta)$;
(b) $\phi(\mu)$ is convex as a function of μ;
(c) $d\phi(\mu)/d\mu = \eta$.



Now verify the diagram here.

Homework 1.33 *Parametric bootstrap*: We resample y^* from $g_{\hat{\eta}}(\cdot)$, $\hat{\eta}$ the

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

MLE based on y. Show that, given y and $\hat{\eta}$,

$$g_\eta(y^*) = g_{\hat{\eta}}(y^*)e^{(\eta-\hat{\eta})(y^*-y)-D(\hat{\eta},\eta)/2}.$$

Deviance Residuals

The idea here is that if $D(y,\mu)$ is the exponential family analogue of $(y-\mu)^2$ in a normal model, then

$$\text{sign}(y-\mu)\sqrt{D(y,\mu)}$$

should be the exponential family analogue of the normal residual $y-\mu$.

We will work in the repeated sampling framework

$$y_i \overset{\text{iid}}{\sim} g_\mu(\cdot), \qquad i = 1,\ldots,n,$$

with MLE $\hat{\mu} = \bar{y}$ and total deviance $D^{(n)}(\hat{\mu},\mu) = nD(\bar{y},\mu)$. The *deviance residual*, of $\hat{\mu} = \bar{y}$ from true mean μ, is defined to be

$$R = \text{sign}(\bar{y}-\mu)\sqrt{D^{(n)}(\bar{y},\mu)}. \tag{1.64}$$

The hope is that R will be nearly $\mathcal{N}(0,1)$, at least closer to normal than the more obvious "Pearson residual"

$$R_P = \frac{\bar{y}-\mu}{\sqrt{V_\mu/n}}$$

(called "z_i" later). Our hope is bolstered by the following theorem, verified in Appendix C of McCullagh and Nelder (1983).

Theorem 1.4 *The asymptotic distribution of R as $n \to \infty$ is*

$$R \doteq \mathcal{N}\left[-a_n, (1+b_n)^2\right], \tag{1.65}$$

where a_n and b_n are defined in terms of the skewness γ_μ and kurtosis δ_μ of the original $(n=1)$ exponential family,

$$a_n = \frac{\gamma_\mu/6}{\sqrt{n}} \quad \textit{and} \quad b_n = \frac{(7/36)\,\gamma_\mu^2 - \delta_\mu}{n}.$$

The normal approximation in (1.65) *is accurate through $O_p(n^{-1})$, with errors of order $O_p(n^{-3/2})$, for instance,*

$$\Pr\left\{\frac{R+a_n}{1+b_n} > 1.96\right\} = 0.025 + O\left(n^{-3/2}\right)$$

(so-called "third-order accuracy").

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

Corollary

$$D^{(n)}(\bar{y}, \mu) = R^2 \doteq \left(1 + \frac{5\gamma_\mu^2 - 3\delta_\mu}{12n}\right) \cdot \chi_1^2,$$

where χ_1^2 is a chi-squared random variable with degrees of freedom 1. Since, according to Hoeffding's formula,

$$D^{(n)}(\bar{y}, \mu) = 2 \log \frac{g_{\hat{\mu}}^{(n)}(\mathbf{y})}{g_\mu^{(n)}(\mathbf{y})},$$

this is an improved version of Wilks' theorem: $2\log(g_{\hat{\mu}}/g_\mu) \to \chi_1^2$ in one-parameter situations.

The constants a_n and b_n are called "Bartlett corrections". The theorem says that

$$R \doteq \frac{Z + a_n}{1 + b_n}, \qquad \text{where } Z \sim \mathcal{N}(0, 1).$$

Since $a_n = O(n^{-1/2})$ and $b_n = O(n^{-1})$, the expectation correction in (1.65) is more important than the variance correction.

Homework 1.34 Consider the gamma case, $y \sim \lambda G_N$ with N fixed (N can be thought of as n).

(a) Show that the deviance residual $\text{sign}(y - \lambda N)\sqrt{D(y, \lambda N)}$ has the same distribution for all choices of λ.

(b) What is the skewness of the Pearson residual $(y - \lambda N)/\lambda(N^{1/2})$?

(c) Use our previous results to show that

$$D^{(n)}(\bar{y}, \mu) \doteq R_P^2 + \frac{\gamma}{6\sqrt{n}} R_P^3 + O_P\left(n^{-1}\right).$$

As an example, Figure 1.8 is a simulation showing 2000 replications of $\bar{y} = \sum_1^5 y_i$, where the y_i are independent G_1 variates; that is, standard one-sided exponentials, as in the Gamma case of Homework 1.30 with $\lambda = N = 1$. This makes $\bar{y} \sim G_5/5$, so that the deviance residual R in (1.64) is calculated as in Homework 1.30 again, now with $N = 5$. The qq-plot shows the deviance residuals (black) much closer to $\mathcal{N}(0, 1)$ than the Pearson residuals (red).

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

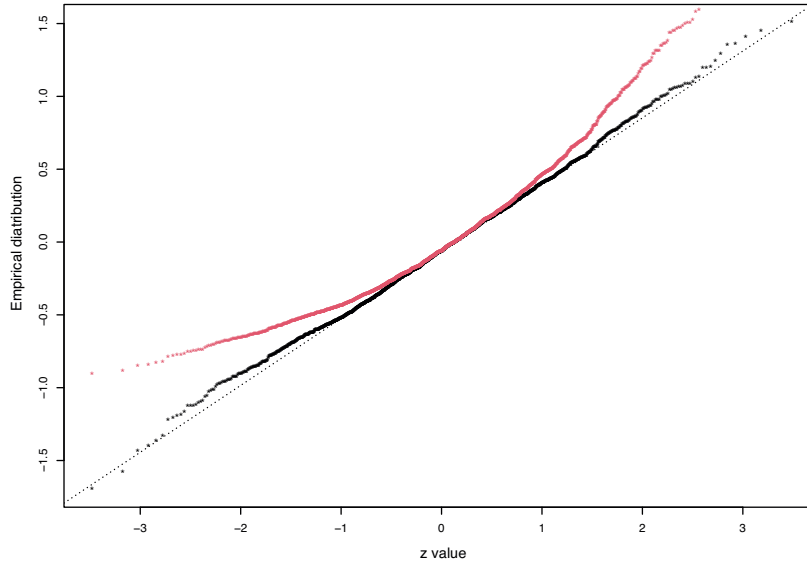


Figure 1.8 qq comparison of deviance residuals (black) with Pearson residuals (red); gamma $N = 1$, $\lambda = 1$, $n = 5$; $B = 2000$ simulations.

An Example of Deviance Analysis

REFERENCE Thisted and Efron (1987), "Did Shakespeare write a newly discovered poem?", *Biometrika* 445–455.

On November 14, 1985, Gary Taylor, a respected Shakespearean scholar, found a short poem of 429 words in the Bodleian Library that he attributed to Shakespeare. This was a controversial stance, as no "new" text by Shakespeare had been discovered in centuries. A word-count analysis was carried out comparing the poem with Shakespeare's attributed works (the "canon"). Table 1.4 shows a small proportion of the results that involved deviance residuals:

- The analysis focused on rare words, that didn't appear often in the canon. Column "y" of the table shows 9 distinct words in the poem that had *never* appeared in the canon, "Prev" = 0; 7 that had previously appeared once each; 5 twice each; and so on, up to 5 that had appeared 80 to 99 times each.
- Column "v" gives predictions for the y values assuming Shakespearean

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

authorship (based on an empirical Bayes Poisson theory relating to Robbins' formula).

- "Dev" and "R" show the Poisson deviance and deviance residuals (1.64) between the counts y and predictions ν.
- a_n is the leading Bartlett correction factor in (1.65), and "RR" the partially corrected residual $R + a_n$.

Table 1.4 *Word-count deviance analysis of newly discovered poem.*

# Prev	y	ν	Dev	R	a_n	RR
0	9	6.97	0.5410	0.736	0.0631	0.799
1	7	4.21	1.5383	1.240	0.0812	1.321
2	5	3.33	0.7247	0.851	0.0913	0.943
3–4	8	5.36	1.1276	1.062	0.0720	1.134
5–9	11	10.24	0.0551	0.235	0.0521	0.287
10–19	10	13.96	1.2478	−1.117	0.0446	−1.072
20–29	21	10.77	7.5858	2.754	0.0508	2.805
30–39	16	8.87	4.6172	2.149	0.0560	2.205
40–59	18	13.77	1.1837	1.088	0.0449	1.133
60–79	8	9.99	0.4257	−0.652	0.0527	−0.600
80–99	5	7.48	0.9321	−0.965	0.0609	−0.904

The RRs should be approximately $\mathcal{N}(0, 1)$ under a hypothesis of Shakespearean authorship. There are some suspicious discrepancies, for instance $RR = 2.805$ for the 20–29 category. The sum of the Dev's is 19.98, moderately large compared to a chi-squared distribution with 11 degrees of freedom,

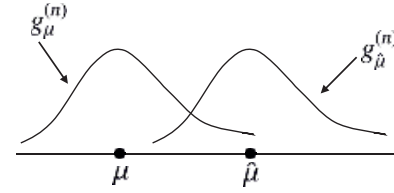
$$\Pr\{\chi^2_{11} > 19.98\} = 0.046.$$

Nevertheless, compared with the same analysis applied to known non-Shakespeare poems, the authors felt that Taylor's poem had at least some chance of being genuine. It remains controversial, and is not usually included in the canon.

1.9 The Saddlepoint Approximation

Suppose we observe a random sample of size n from some member of an exponential family $\mathcal{G}$,

$$y_1, \ldots, y_n \overset{\text{iid}}{\sim} g_\mu(\cdot)$$



https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

(now indexed by expectation parameter μ), and wish to approximate the density under $g_\mu^{(n)}$ of the sufficient statistic $\hat{\mu} = \bar{y}$ for a value of $\hat{\mu}$ perhaps far removed from μ. Let $g_\mu^{(n)}(\hat{\mu})$ denote this density.

The normal approximation

$$g_\mu^{(n)}(\hat{\mu}) \doteq \sqrt{\frac{n}{2\pi V_\mu}} e^{-\frac{1}{2}\frac{n}{V_\mu}(\hat{\mu}-\mu)^2} \tag{1.66}$$

is likely to be inaccurate if $\hat{\mu}$ is, say, several standard errors away from μ. Hoeffding's formula (1.62) provides a much better result, called the *saddlepoint approximation*. We write

$$g_\mu^{(n)}(\hat{\mu}) = g_{\hat{\mu}}^{(n)}(\hat{\mu}) e^{-nD(\hat{\mu},\mu)/2}. \tag{1.67}$$

For $\mu = \hat{\mu}$, $\bar{y}$ is approximately $\mathcal{N}(\hat{\mu}, \widehat{V}/n)$, where $\widehat{V} = \ddot{\psi}(\hat{\mu})$ is the variance of a single y under $g_{\hat{\mu}}$, giving $g_{\hat{\mu}}^{(n)}(\hat{\mu}) \doteq [n/(2\pi\widehat{V})]^{1/2}$ and, substituting in (1.67), the saddlepoint approximation

$$g_\mu^{(n)}(\hat{\mu}) \doteq \sqrt{\frac{n}{2\pi\widehat{V}}} e^{-nD(\hat{\mu},\mu)/2}. \tag{1.68}$$

Because (1.68) only involves applying the central limit theorem at the *center* of the $g_{\hat{\mu}}^{(n)}(\cdot)$ distribution, just where it is most accurate, the error in (1.68) is a factor of only $1 + O(n^{-1})$, compared to $1 + O(n^{-1/2})$ for (1.66). There is an enormous literature of extensions and improvements to the saddlepoint approximation, a good review article being Reid (1988).

Let

$$L_{\boldsymbol{y}}(\mu) = g_\mu^{(n)}(\boldsymbol{y})$$

be the likelihood function having observed data $\boldsymbol{y}$, expressed in terms of the expectation parameter μ. Hoeffding's formula (1.62), (1.63), gives

$$e^{-nD(\hat{\mu},\mu)/2} = \frac{L_{\boldsymbol{y}}(\mu)}{L_{\boldsymbol{y}}(\hat{\mu})},$$

so the saddlepoint approximation can be expressed as

$$g_\mu^{(n)}(\hat{\mu}) \doteq \sqrt{\frac{n}{2\pi\widehat{V}}} \frac{L_{\boldsymbol{y}}(\mu)}{L_{\boldsymbol{y}}(\hat{\mu})},$$

which provides an expression for the density of $\bar{y} = \hat{\mu}$ in terms of the likelihood function.

Since $d\hat{\mu}/d\hat{\eta} = \widehat{V}$, there is an equivalent expression for the density of $\hat{\eta}$,

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

say $g_\eta^{(n)}(\hat\eta)$ (abusing notation somewhat),

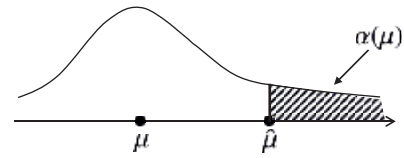
$$g_\eta^{(n)}(\hat\eta) \doteq \sqrt{\frac{n\widehat{V}}{2\pi}}\,\frac{L_y(\eta)}{L_y(\hat\eta)}. \tag{1.69}$$

Barndorff-Nielsen (1980) showed approximation (1.69) holding in a variety of situations, including *curved exponential families* (Part 4), and it is sometimes known as his "magic formula".

The Lugananni–Rice Formula

The saddlepoint formula can be integrated to give an approximation to $\alpha(\mu)$, the *attained significance level* or "p-value" of parameter value μ having observed $\bar y = \hat\mu$:

$$\alpha(\mu) = \int_{\hat\mu}^{\infty} g_\mu^{(n)}(t)m(dt).$$



Numerical integration is required to compute $\alpha(\mu)$ from the saddlepoint formula itself, but the *Lugananni–Rice formula* provides a highly accurate closed-form approximation:

$$\alpha(\mu) \doteq 1 - \Phi(R) - \varphi(R)\left(\frac{1}{R} - \frac{1}{Q}\right) + O\left(n^{-3/2}\right),$$

where Φ and φ are the standard normal CDF and density,

$$R = \text{sign}(\hat\mu - \mu)\sqrt{nD(\hat\mu, \mu)}$$

the deviance residual, and

$$Q = \sqrt{n\widehat{V}} \cdot (\hat\eta - \eta)$$

the crude form of the Pearson residual based on the canonical parameter η, not on μ. (Remember that $\widehat{\text{sd}}(\hat\eta) \doteq (n\widehat{V})^{-1/2}$, so $Q = (\hat\eta - \eta)/\widehat{\text{sd}}(\hat\eta)$.) Reid (1988) is also an excellent reference here, giving versions of the Lugananni–Rice formula that apply not only to exponential family situations but also to general distributions of $\bar y$. See also Section 6 of Daniels (1983).

Homework 1.35 Suppose we observe $y \sim \lambda G_N$, G_N gamma df = N, with $N = 10$ and $\lambda = 1$. Use the Lugananni–Rice formula to calculate $\alpha(\mu)$ for

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

$y = \hat{\mu} = 15, 20, 25, 30$, and compare the results with the exact values. (You can use any expression above for R.)

Homework 1.36 Another version of the Lugananni–Rice formula is

$$1 - \alpha(\mu) \doteq \Phi(R'),$$

where

$$R' = R + \frac{1}{R} \log \frac{Q}{R}.$$

How does this relate to the first form?

Large Deviations and the Chernoff Bound

In a generic "large deviations" problem, we observe an i.i.d. sample

$$y_1, \ldots, y_n \overset{\text{iid}}{\sim} g_0(\cdot)$$

from a *known* density g_0 having mean and standard deviation $y_i \sim (\mu_0, \sigma_0)$. We wish to compute

$$\alpha_n(\mu_1) = \Pr_{g_0}\{\bar{y} \geq \mu_1\}$$

for some fixed value $\mu_1 > \mu_0$. As $n \to \infty$, the number of standard errors $\sqrt{n}(\mu_1 - \mu_0)/\sigma_0$ gets big, rendering the central limit theorem useless.

Homework 1.37 ("Chernoff bound") Let $g_\eta(y) = e^{\eta y - \psi(\eta)} g_0(y)$ ("the exponential family through g_0").

(a) For any $\lambda > 0$ show that $\alpha_n(\mu_1) = \Pr_{g_0}\{\bar{y} \geq \mu_1\}$ satisfies

$$\alpha_n(\mu_1) \leq \beta_n(\mu_1) \equiv \int_{\mathcal{Y}} e^{n\lambda(\bar{y} - \mu_1)} g_0^{(n)}(\bar{y})\, d\bar{y}.$$

(b) Show that $\beta_n(\mu_1)$ is minimized at λ equal the value $\hat{\eta}$ such that

$$\dot{\psi}(\hat{\eta}) = \mu_1.$$

(c) Finally, verify Chernoff's large deviation bound

$$\Pr_{g_0}\{\bar{y} \geq \mu_1\} \leq e^{-nD(\mu_1, 0)/2}, \tag{1.70}$$

where $D(\mu_1, 0)$ is the deviance between $g_{\hat{\eta}}(y)$ and $g_0(y)$.

Notice that for fixed μ_1, neither $\hat{\eta}$ nor $D(\mu_1, 0)$ depends on n, so $\alpha_n(\mu_1) \to 0$ exponentially fast, which is typical for large deviation results.

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

1.10 Transformation Theory

Reference DiCiccio (1984), "On parameter transformations and interval estimation", *Biometrika* 477–485.

Reference Efron (1982), "Transformation theory: How normal is a family of distributions?", *Ann. Stat.* 323–339.

Reference Hougaard (1982), "Parametrizations of nonlinear models", *JR SS-B* 244–252.

Power transformations are used to make exponential families more like the standard normal translation family $Y \sim \mathcal{N}(\mu, 1)$. For example, $Y \sim \text{Poi}(\mu)$ has variance $V_\mu = \mu$ depending on the expectation μ, while the transformation

$$Z = H(Y) = 2\sqrt{Y}$$

yields, approximately, $\text{Var}(Z) = 1$ for all μ. In a regression situation with Poisson responses $y_1, \ldots, y_n$, we might first change to $z_i = 2y_i^{1/2}$ and then employ standard linear model methods. (That's *not* how we will proceed in Part 3, where generalized linear model techniques are discussed. The introduction of GLMs reduced, but did not eliminate, interest in transformation theory.)

Table 1.5, credited to unpublished work by R. Wedderburn, encompasses a considerable number of special transformations for one-parameter exponential families. Let ζ be a transformation of μ,

$$\zeta = H(\mu) \quad \text{and} \quad \hat{\zeta} = H(\hat{\mu}),$$

where $\hat{\mu}$ is the MLE of μ based on observing a single $y \sim g_\mu(\cdot)$. If we make the derivative $H'(\mu)$ satisfy

$$H'(\mu) = V_\mu^{\delta-1}, \tag{1.71}$$

then various choices of δ result in $\hat{\zeta} = H(\hat{\mu})$ satisfying the properties shown in Table 1.5, as explained next.

Table 1.5 *Wedderburn's transformations* (1.71) *and their results.*

δ	0	1/3	1/2	2/3	1
Result	Canonical parameter η	Normal likelihood	Stabilized variance	Normal density	Expectation parameter μ

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

The choice $\delta = 0$ has

$$\frac{d\zeta}{d\mu} = H'(\mu) = \frac{1}{\sqrt{V_\mu}}.$$

But $d\eta/d\mu = 1/V_\mu$, so in this case $\zeta = \eta$. At the other end of the scale, $\delta = 1$ has $H'(\mu) = 1$, that is, $\zeta = \mu$.

The stabilized variance result, $\delta = 1/2$, follows from the delta method:

$$\hat{\zeta} = H(\hat{\mu}), \qquad \text{with } H'(\mu) = \frac{1}{\sqrt{V_\mu}},$$

implies that

$$\mathrm{sd}_\mu(\hat{\zeta}) \doteq \frac{\mathrm{sd}_\mu(\hat{\mu})}{\sqrt{V_\mu}} = 1.$$

For the Poisson family, with $V_\mu = \mu$,

$$H'(\mu) = \frac{1}{\sqrt{\mu}}$$

gives

$$H(\mu) = 2\sqrt{\mu} + \text{any constant},$$

as above. For $Y \sim \mathrm{Poi}(\mu)$, the usual approximation for expectation and variance is

$$2\sqrt{Y} \dot{\sim} (2\sqrt{\mu}, 1). \tag{1.72}$$

Homework 1.38 Numerically calculate how well (1.72) works for $\mu = 5, 8, 12, 18, 25$.

Small adjustments to the $\delta = 1/2$ formula are known to improve variance stabilization. For the binomial case

$$p \sim \mathrm{Bi}(N, \pi)/N,$$

Anscombe's transformation

$$\hat{\zeta} = 2\sqrt{N}\sin^{-1}\left(\sqrt{\frac{Np + 3/8}{N + 3/4}}\right) \tag{1.73}$$

does a good job of making $\mathrm{Var}_\pi(\hat{\zeta}) \doteq 1$ for n say 15 or more.

Homework 1.39 Ignoring correction terms $3/8$ and $3/4$, show that (1.71) with $\delta = 1/2$ gives (1.73).

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

Normal density, $\delta = 2/3$, means that $\hat{\zeta} = H(\hat{\mu})$ is approximately $\mathcal{N}(H(\mu), 1)$. (This is *not* the same as $\delta = 1/2$ in Table 1.5, where the emphasis is on constant variance rather than normality.) Its rationale is based on asymptotic expansions. Working in a repeated sampling framework, $y_1, \ldots, y_n \overset{\text{iid}}{\sim} g_\mu(\cdot)$, where $\hat{\mu} = \bar{y}$, define

$$z_n = \frac{\bar{y} - \mu}{\sqrt{V/n}} \qquad (V = V_\mu),$$

so that z_n has expectation 0, variance 1, and skewness $\gamma \cdot n^{-1/2}$, where γ is the skewness for a single y_i, as in Homework 1.7. A two-term *Cornish–Fisher expansion* suggests that the distribution of z_n can be normalized by the transformation

$$Z = z_n - \frac{\gamma}{6\sqrt{n}}(z_n^2 - 1), \tag{1.74}$$

which makes the skewness of Z approximately 0.

We want to show that $\hat{\zeta} = H(\hat{\mu})$, with $\delta = 2/3$ in (1.71), asymptotically agrees with (1.74), at least through two terms. The Taylor expansion

$$\begin{aligned}\hat{\zeta} &\doteq H(\mu) + H'(\mu)(\hat{\mu} - \mu) + H''(\mu)\frac{(\hat{\mu} - \mu)^2}{2} \\ &= H(\mu) + H'(\mu)\sqrt{\frac{V}{n}}z_n + H''(\mu)\frac{V}{2n}z_n^2\end{aligned}$$

gives $\hat{\zeta}$ approximate expectation and standard deviation

$$\begin{aligned}E_\mu\{\hat{\zeta}\} &= H(\mu) + H''(\mu)\frac{V}{2}, \\ \mathrm{sd}_\mu\{\hat{\zeta}\} &= H'(\mu)\sqrt{\frac{V}{n}}\end{aligned}$$

(assuming $H'(\mu) > 0$). Therefore

$$\frac{\hat{\zeta} - E_\mu\{\hat{\zeta}\}}{\mathrm{sd}_\mu\{\hat{\zeta}\}} \doteq z_n + \frac{H''(\mu)}{2H'(\mu)}\sqrt{\frac{V}{n}}(z_n^2 - 1). \tag{1.75}$$

Expansion (1.75) agrees with (1.74) if

$$\frac{H''(\mu)}{H'(\mu)}\sqrt{V} = -\frac{\gamma}{3},$$

or equivalently if

$$\frac{d\log H'(\mu)}{d\mu} = -\frac{\gamma}{3\sqrt{V}}.$$

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press

However, $\gamma = V' \cdot V^{-1/2}$ is an exponential family (1.19), so in order to get agreement we need

$$\frac{d \log H'(\mu)}{d\mu} = -\frac{V'}{3V} = \frac{d}{d\mu} \log V^{-1/3}$$

or

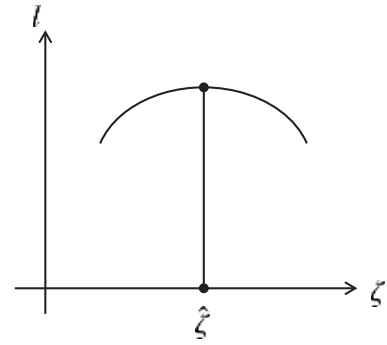
$$H'(\mu) = cV^{-1/3},$$

i.e, $\delta = 2/3$, as in Table 1.5.

Normal likelihood, $\delta = 1/3$, means that the transformation $\hat{\zeta} = H(\hat{\mu})$ results in

$$\left.\frac{\partial^3 l_\mu(y)}{\partial \zeta^3}\right|_{\hat{\zeta}} = 0, \qquad (1.76)$$

where $l_\mu(y) = \log g_\mu(y)$. This makes the log likelihood look parabolic near its maximum at $\zeta = \hat{\zeta}$. Efron (1982) gives an argument connecting $H'(\mu) = V_\mu^{-2/3}$ and (1.76).



For the Poisson case $Y \sim \text{Poi}(\mu)$, the three choices $\delta = 1/3, 1/2$, or $2/3$ correspond to the transformations $3/2Y^{2/3}, 2Y^{1/2}$, or $3Y^{1/3}$. All three have been referred to in the literature as "the" Poisson transformation.

Homework 1.40 We observe independent χ^2 variables

$$\hat{\sigma}_i^2 \sim \frac{\sigma_i^2 \chi^2_{\nu_i}}{\nu_i},$$

the ν_i being known degrees of freedom, and wish to regress $\hat{\sigma}_i^2$ versus some known covariates. Two frequently suggested transformations are $\log(\hat{\sigma}_i^2)$ and $(\hat{\sigma}_i^2)^{1/3}$, the latter being the "Wilson–Hilferty" transformation. Discuss the two transformations in terms of Table 1.5.

https://doi.org/10.1017/9781108773157.003 Published online by Cambridge University Press