Introduction

The inner circle in Figure 1 represents normal theory, the preferred venue of classical applied statistics. Exact inference — t tests, F tests, chi-squared statistics, ANOVA, multivariate analysis — were feasible within the circle. Outside the circle was a general theory based on large-sample asymptotic approximation involving Taylor series and the central limit theorem.

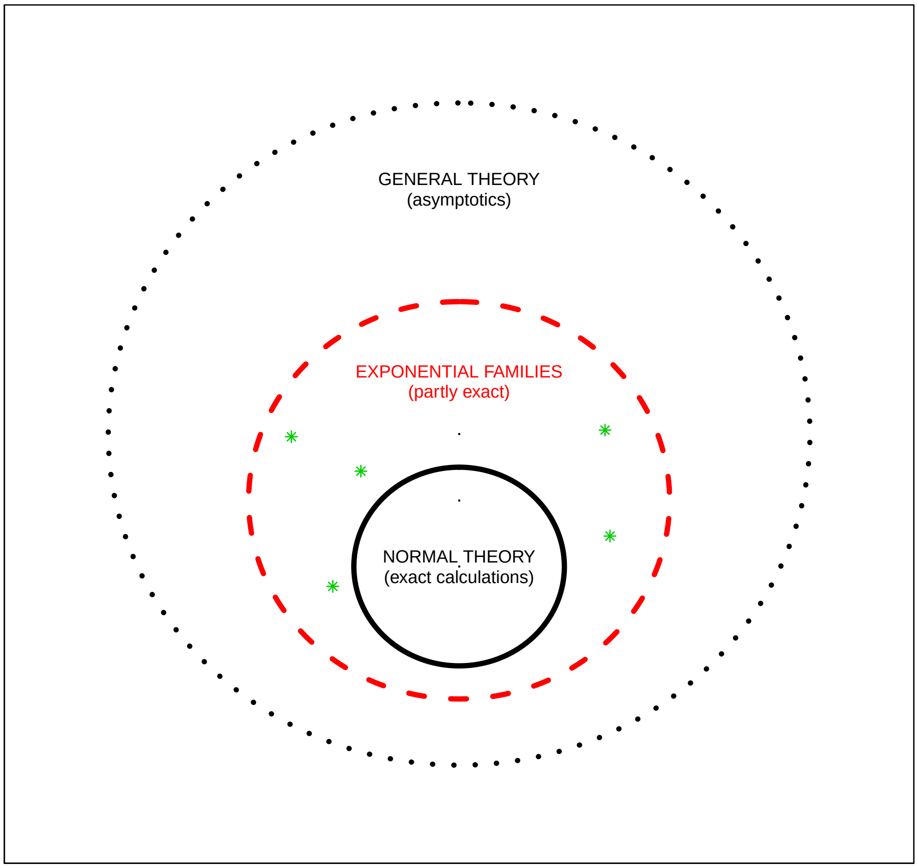


Figure 1: Three levels of statistical modeling

A few special exact results lay outside the normal circle, relating to specially tractable distributions such as the binomial, Poisson, gamma and beta families. These are the figure's green stars.

A happy surprise, though a slowly emerging one beginning in the 1930s, was that all the special cases were examples of a powerful general construction: *exponential families.* Within this superfamily, the intermediate circle in Figure 1, “almost exact” inferential theories such as generalized linear models (GLMs) are possible. This course will examine the theory of exponential families in a relaxed manner with an eye toward applications. A much more rigorous approach to the theory is found in Larry Brown’s 1986 monograph, *Fundamentals of Statistical Exponential Families*, IMS series volume 9.

A salient fact is that no one name is credited with the development of exponential families, though it will be clear from these notes that R.A. Fisher’s work was seminal. The name “exponential families” is relatively new. Until the late 1950s they were often referred to as “ Koopman–Darmois–Pitman” families (the names of three prominent statisticians working separately in three different countries), the long name suggesting little influence attached to the ideas.

Our title, “Exponential families in theory and practice,” might well be amended to “. . . *between* theory and practice.” These notes collect a large amount of material useful in statistical applications, but also of value to the theoretician trying to frame a new situation without immediate recourse to asymptotics. My own experience has been that when I can put a problem, applied or theoretical, into an exponential family framework, a solution is imminent. There are almost no proofs in what follows, but hopefully enough motivation and heuristics to make the results believable if not obvious. References are given when this doesn’t seem to be the case.

Part 1

One-parameter Exponential Families

One-parameter exponential families are the building blocks for the multiparameter theory developed in succeeding parts of this course. Useful and interesting in their own right, they unify a vast collection of special results from classical methodology. Part I develops their basic properties and relationships, with an eye toward their role in the general data-analytic methodology to follow.

1.1 Definitions and notation

This section reviews the basic definitions and properties of one-parameter exponential families. It also describes the most familiar examples — normal, Poisson, binomial, gamma — as well as some less familiar ones.

Basic definitions and notation

The fundamental unit of statistical inference is a family of probability densities $\mathcal{G}$, "density" here including the possibility of discrete atoms. A one-parameter exponential family has densities $g_\eta(y)$ of the form

$$\mathcal{G} = \{g_\eta(y) = e^{\eta y - \psi(\eta)} g_0(y) m(dy),\ \eta \in A,\ y \in \mathcal{Y}\}, \tag{1.1}$$

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

Terminology

- 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 statistic* 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).

- 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 (1.1) is the *normalizing function* or *cumulant generating function*; it scales the densities $g_\eta(y)$ to integrate to 1 over the sample space $\mathcal{Y}$,

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

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

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

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 $\tilde{g}_{\tilde{\eta}}(\tilde{y})$.

We can construct an exponential family $\mathcal{G}$ through any given density $g_0(y)$ by "tilting" it exponentially,

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

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

$$g_\eta(y) = e^{\eta y - \psi(\eta)} g_0(y) \qquad \left(e^{\psi(\eta)} = \int_{\mathcal{Y}} e^{\eta y} g_0(y) m(dy)\right).$$

It seems like we might employ other tilting functions, say

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

but only exponential tilting gives convenient properties under independent sampling.

If η_0 is any point on A we can write

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

This is the same exponential family, now represented with

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

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}. Notice that the sample space $\mathcal{Y}$ is the *same* for all members of $\mathcal{G}$, and that all put positive probability on every point in $\mathcal{Y}$.

The Poisson family

As an important first example we consider the Poisson family. A Poisson random variable Y having mean μ takes values on the non-negative integers $\mathcal{Z}_+ = \{0, 1, 2, \dots\}$,

$$\Pr_\mu\{Y = y\} = e^{-\mu}\mu^y/y! \qquad (y \in \mathcal{Z}_+).$$

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

$$g_\eta(y) = e^{\eta y - \psi(\eta)} g_0(y) \begin{cases} \eta = \log(\mu) & (\mu = e^\eta) \\ \psi(\eta) = e^\eta & (= \mu) \\ g_0(y) = {}^1\!/_{y!} & (\text{not a member of } \mathcal{G}). \end{cases}$$

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).

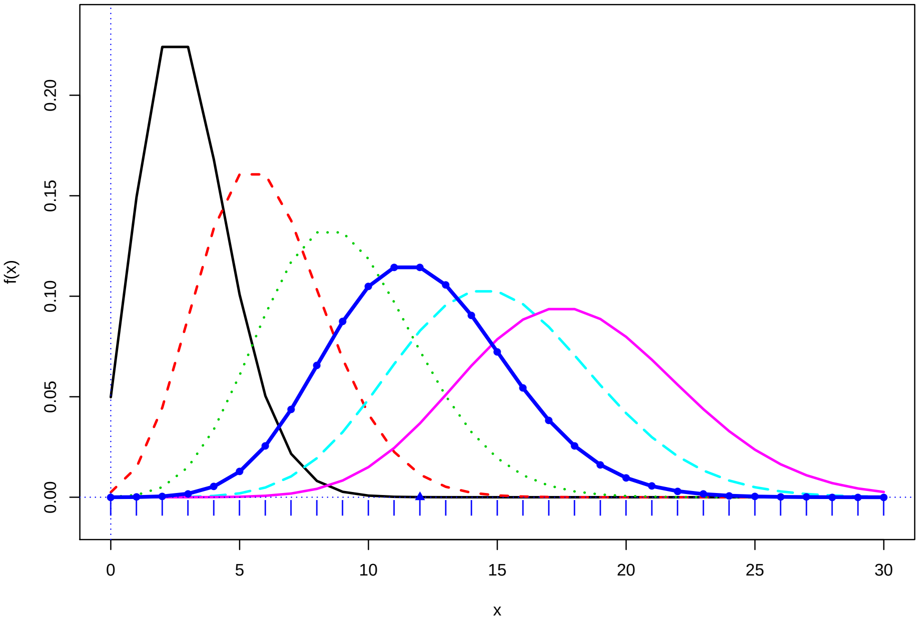


Figure 1.1: Poisson densities for $\mu = 3, 6, 9, 12, 15, 18$; heavy curve with dots for $\mu = 12$.

1.2 Moment relationships

Expectation and variance

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

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

and

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

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

$$\dot{\psi}(\eta) = E_\eta(Y) \equiv \text{“}\mu_\eta\text{”}$$

and

$$\ddot{\psi}(\eta) = \text{Var}_\eta\{Y\} \equiv \text{“}V_\eta\text{”};$$

V_η is greater than 0, implying that $\psi(\eta)$ is a convex function of η.

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

$$\zeta = h(\eta) = h\left(\eta(\mu)\right) = \text{“}H(\mu)\text{”}.$$

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

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

Skewness and kurtosis

$\psi(\eta)$ is the *cumulant generating function* for g_0 and $\psi(\eta) - \psi(\eta_0)$ is the CGF for $g_{\eta_0}(y)$, i.e.,

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

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

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

Equivalently,

$$\begin{array}{llll} \dot{\psi}(\eta_0) = k_1, & \ddot{\psi}(\eta_0) = k_2, & \dddot{\psi}(\eta_0) = k_3, & \ddddot{\psi}(\eta_0) = k_4 \\ \Big[\quad = \mu_0 & = V_0 & = E_0\{y_0 - \mu_0\}^3 & = E_0\{y_0 - \mu_0\}^4 - 3V_0^2 \Big] \end{array}$$

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

By definition, for a real-valued random variable Y,

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

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 all together, if $Y \sim g_\eta(\cdot)$ in an exponential family,

$$Y \sim \left[\begin{array}{cccc} \dot{\psi}, & \ddot{\psi}^{1/2}, & \dddot{\psi}/\ddot{\psi}^{3/2}, & \ddddot{\psi}/\ddot{\psi}^{2} \\ \uparrow & \uparrow & \uparrow & \uparrow \\ \text{expectation} & \text{standard} & \text{skewness} & \text{kurtosis} \\ & \text{deviation} & & \end{array} \right]$$

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

$$Y \sim \left[\begin{array}{cccc} \mu, & \sqrt{\mu}, & 1/\sqrt{\mu}, & 1/\mu \\ \uparrow & \uparrow & \uparrow & \uparrow \\ \text{exp} & \text{st dev} & \text{skew} & \text{kurt} \end{array} \right]$$

A useful result

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

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

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

$$\gamma = 2(V^{1/2})' = 2\frac{d}{d\mu}\,\mathrm{sd}_\mu$$

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.

Homework 1.4. Show that

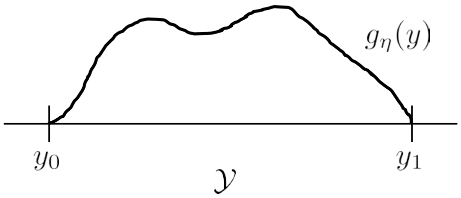
$$\text{(a)} \quad \delta = V'' + \gamma^2 \quad \text{and} \quad \text{(b)} \quad \gamma' = \left(\delta - \frac{3}{2}\gamma^2\right) \Big/ \mathrm{sd}\,.$$

Note. All of the classical exponential families — binomial, Poisson, normal, etc. — are those with closed form CGFs ψ. This yields neat expressions for means, variances, skewnesses, and kurtoses.

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 $l_0'(y) = \frac{dl_0(y)}{dy}$.
- Suppose $\mathcal{Y} = [y_0, y_1]$ (both possibly infinite) and that $m(y) = 1$.



Lemma 1.

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

Homework 1.5. Prove the lemma. (*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 investigate how well $E_\eta\{-l_0'(y)\}$ approximates η in the Poisson family.

1.3 Repeated sampling

Suppose that $y_1, y_2, \ldots, y_n$ is an independent and identically distributed (i.i.d.) sample from an exponential family $\mathcal{G}$:

$$y_1, y_2, \ldots, y_n \overset{\text{iid}}{\sim} g_\eta(\cdot),$$

for an unknown value of the parameter $\eta \in A$. The density of $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ 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 $g_\eta^{\boldsymbol{Y}}(\boldsymbol{y})$ indicate the density of $\boldsymbol{y}$ with respect to $\prod_{i=1}^n m(dy_i)$,

$$g_\eta^{\boldsymbol{Y}}(\boldsymbol{y}) = e^{n[\eta\bar{y} - \psi(\eta)]} \prod_{i=1}^n g_0(y_i). \tag{1.2}$$

This is 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 $\bar{y} \sim [\mu, \sqrt{V/n}, \gamma/\sqrt{n}, \delta/n]$.

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

Notice that $\boldsymbol{y}$ is now a vector, and that the tilting factor $e^{\eta^{(n)}\bar{y}}$ is tilting the *multivariate* 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 $\mathbf{1} = (1, 1, \ldots, 1)$.

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

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

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

The density (1.2) can also be written 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. $y_1, y_2, \ldots, y_n \overset{\text{iid}}{\sim} \text{Poi}(\mu)$. Describe the distributions of $\overline{Y}$ and S, and say what are the exponential family quantities $(\eta, y, \psi, g_0, m, \mu, V)$ in both cases.

1.4 Some well-known one-parameter families

We've already examined the Poisson family. This section examines some other well-known (and not so well-known) examples.

Normal with variance 1

$\mathcal{G}$ is the normal family $Y \sim \mathcal{N}(\mu, 1)$, μ in $\mathcal{R}^1$. The densities, taken with respect to $m(dy) = dy$, Lebesque measure,

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

can be written in exponential family form (1.1) with

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

Homework 1.9. Suppose $Y \sim \mathcal{N}(\mu, \sigma^2)$ with σ^2 known. Give η, y, ψ, and g_0.

Binomial

$Y \sim \text{Bi}(N, \pi)$, N known, so

$$g(y) = \binom{N}{y} \pi^y (1-\pi)^{N-y}$$

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:

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

Homework 1.10. Show that for the binomial

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

Homework 1.11. Notice that $A = (-\infty, \infty)$ does *not* include the cases $\pi = 0$ or $\pi = 1$. Why not?

Gamma

$Y \sim \lambda G_N$ where G_N is a standard gamma variable, N known, and λ an unknown scale parameter,

$$g(y) = \frac{y^{N-1} e^{-y/\lambda}}{\lambda^N \Gamma(N)} \qquad [\mathcal{Y} = (0, \infty)].$$

This is a one-parameter exponential family with:

- $\eta = -1/\lambda$
- $\mu = N\lambda = -N/\eta$
- $V = N/\eta^2 = N\lambda^2 = \mu^2/N$

- $\psi = -N \log(-\eta)$
- $\gamma = 2/\sqrt{N}$
- $\delta = 6/N$

Negative binomial

A coin with probability of heads θ is flipped until exactly $k+1$ heads are observed. Let $Y = \#$ of tails observed. Then

$$\begin{aligned} g(y) &= \binom{y+k}{k}(1-\theta)^y \theta^{k+1} \\ &= \binom{y+k}{k} e^{[\log(1-\theta)]y + (k+1)\log\theta} \qquad [\mathcal{Y} = (0, 1, 2, \dots)] \,. \end{aligned}$$

This is a one-parameter exponential family with:

- $\eta = \log(1-\theta)$
- $\psi = -(k+1)\log(1-e^\eta)$

Homework 1.12. (a) Find μ, V, and γ as a function of θ. (b) Notice that $\psi = (k+1)\psi_0$ where ψ_0 is the normalizing function for $k = 0$. Give a simple explanation for this. (c) How does it affect the formula for μ, V, and γ?

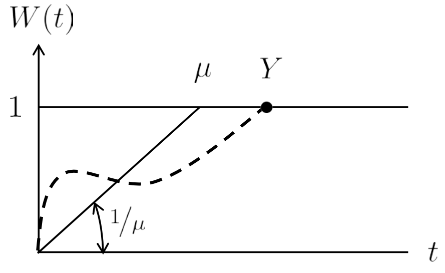
Inverse Gaussian

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

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

This is an exponential family with:

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



Reference Johnson and Kotz, *Continuous Univariate Densities Vol. 1*, Chapter 15

Homework 1.13. Show $Y \sim [\mu, \mu^{3/2}, 3\sqrt{\mu}, 15\mu]$ as the mean, standard deviation, skewness, and kurtosis, respectively.

Note. The early Generalized Linear Model literature was interested in the construction of non-standard exponential families with relations such as $V = \mu^{1.5}$.

	Normal	Poisson	Gamma	Inverse normal
$V \propto$	constant	μ	μ^2	μ^3

2×2 table

Let $X = (x_1, x_2, x_3, x_4)$ be a multinomial sample of size N from a 4-category multinomial layout, where the categories form a double dichotomy as shown.

$$(x_1, x_2, x_3, x_4) \sim \text{Mult}_4\left[N, (\pi_1, \pi_2, \pi_3, \pi_4)\right]$$

	Yes	*No*	Row totals
Men	X_1 π_1	X_2 π_2	r_1
Women	X_3 π_3	X_4 π_4	r_2
	c_1	c_2	N
	Column totals		

with $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)$ the true probabilities, $\sum_1^4 \pi_i = 1$. Given the table's marginal totals (N, r_1, c_1), we need only know x_1 to fill in (x_1, x_2, x_3, x_4). (Fisher suggested analyzing the table with marginals thought of as fixed ancillaries, for reasons discussed next.)

The conditional density of x_1 given (N, r_1, c_1) depends only on the *log odds* parameter

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

so conditioning has reduced our four-parameter inferential problem to a simpler, one-parameter situation. Notice that $\theta = 0$ corresponds to $\pi_1/\pi_2 = \pi_3/\pi_4$, which is equivalent to independence between the two dichotomies.

The conditional density of $x_1 \mid (N, r_1, c_1)$, with respect to counting measure, is

$$\begin{aligned} g_\theta(x_1 \mid N, r_1, c_1) &= \binom{r_1}{x_1}\binom{r_2}{c_1 - x_1} e^{\theta x_1} \big/ C(\theta), \\ C(\theta) &= \sum_{x_1} \binom{r_1}{x_1}\binom{r_2}{c_1 - x_1} e^{\theta x_1}, \end{aligned} \tag{1.3}$$

the sum being over the sample space of possible x_1 values,

$$\max(0, c_1 - r_2) \leq x_1 \leq \min(c_1, r_1).$$

Reference Lehmann, "Testing statistical hypotheses", Section 4.5

This is a one-parameter exponential family with:

- $\eta = \theta$
- $y = x$
- $\psi = \log(C)$ $\quad$ ($\theta = 0$ corresponds to the *hypergeometric distribution.*)

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

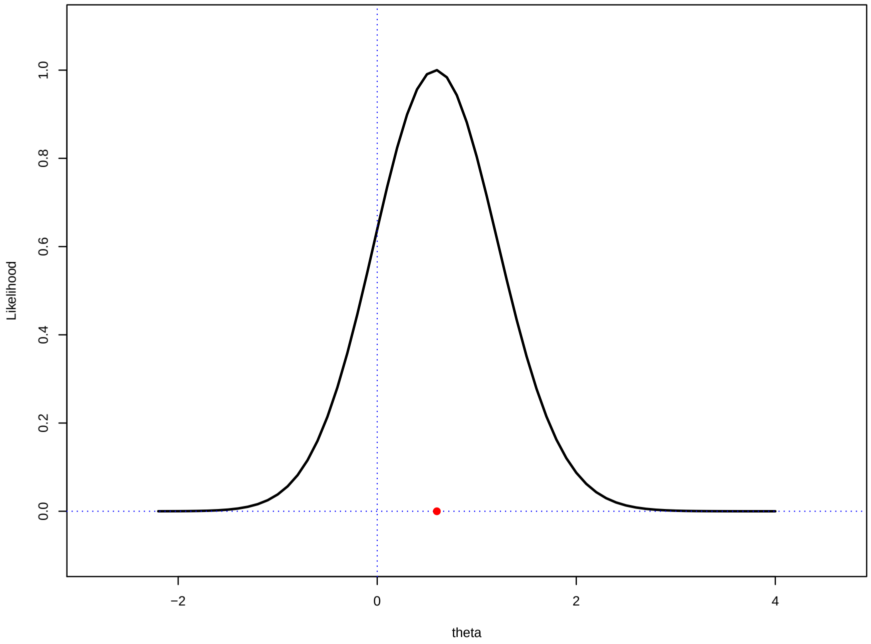


Figure 1.2: `ulcdata` #14; likelihood function for crossproduct ratio θ; max at $\theta = 0.600$; $-\ddot{l} = 2.56$

Example. The 14th experiment on `ulcdata` involved 45 patients in a clinical trial comparing a new experimental surgery for stomach ulcers with the standard control procedure. The obvious estimate of θ is

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

Figure 1.2 graphs the likelihood, i.e., expression (1.3) as a function of θ, with the data held fixed as observed (normalized so that $\max\{L(\theta)\} = 1$).

Homework 1.14. (a) Compute the likelihood numerically and verify that it is maximized at $\hat{\theta} = 0.600$. (b) Verify numerically that

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

(c) Using this result, guess the variance of $\hat{\theta}$.

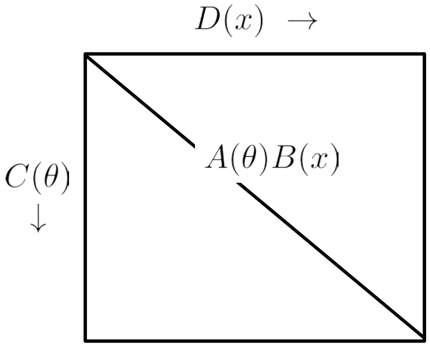
The structure of one-parameter exponential families

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

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

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 1.15. I constructed a 14×9 matrix P with ijth element

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

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

$$\begin{aligned} x_i &= i \qquad \text{for } i = 0, 1, 2, \dots, 13 \\ \theta_j &= 0.1, 0.2, \dots, 0.9. \end{aligned}$$

Then I calculated the singular value decomposition (svd) of $\log P$. How many non-zero singular values did I see?

1.5 Bayes families

Suppose we observe $Y = y$ from

$$g_\eta(y) = e^{\eta y - \psi(\eta)} g_0(y), \tag{1.4}$$

where η itself has a prior density

$$\eta \sim \pi(\eta) \qquad \text{(with respect to Lebesgue measure on } A).$$

Bayes rule gives posterior density for η

$$\pi(\eta \mid y) = \pi(\eta) g_\eta(y) / g(y),$$

where $g(y)$ is the *marginal density*

$$g(y) = \int_A \pi(\eta) g_\eta(y) \; d\eta.$$

(Note that $g_\eta(y)$ is the *likelihood function*, with y fixed and η varying.) Plugging in (1.4) gives

$$\pi(\eta \mid y) = e^{y\eta - \log[g(y)/g_0(y)]} \left[\pi(\eta) e^{-\psi(\eta)}\right]. \tag{1.5}$$

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

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

Homework 1.16. (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?

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 1.17. (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) Then

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

(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 $y_1, y_2 \dots, y_n \overset{\text{iid}}{\sim} \text{Poi}(\mu)$?

Diaconis and Ylvisaker (1979, *Ann. Statist.* 269–281) provide a general formulation of conjugacy:

$$y_1, y_2, \dots, y_n \overset{\text{iid}}{\sim} g_\eta(y) = e^{\eta y - \psi(\eta)} g_0(y);$$

the prior for μ wrt Lebesgue measure is

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

where y_0 is notionally the average of n_0 hypothetical prior observations of y (c_0 the constant making $\pi_{n_0,y_0}(\mu)$ integrate to 1).

Theorem 1.

$$\pi(\mu \mid y_1, y_2, \ldots, y_n) = \pi_{n_+,y_+}(\mu),$$

where

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

Moreover,

$$E\{\mu \mid y_1, y_2, \ldots, y_n\} = y_+.$$

Binomial case

$y \sim \mathrm{Bi}(n, \pi)$, with hypothetical prior observations y_0 successes out of n_0 tries. Assuming a "beta" prior (Part 2) yields Bayes posterior expectation

$$\hat{\theta} = E\{\pi \mid y\} = \frac{y_0 + y}{n_0 + n}.$$

Current Bayes practice favors small amounts of hypothetical prior information, in the binomial case maybe $y_0 = 1$ and $n_0 = 2$, giving

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

pulling the MLE y/n a little toward $1/2$.

Tweedie's formula

Equation (1.5) gave

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

where

$$\pi_0(y) = \pi(\eta) e^{-\psi(\eta)} \quad \text{and} \quad \lambda(y) = \log\left[g(y)/g_0(y)\right],$$

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

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

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

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

and

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

Homework 1.18. 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

$$\mu \mid y \sim \left\{y + \sigma^2 l'(y), \sigma^2 \left[1 + \sigma^2 l''(y)\right]\right\}. \tag{1.6}$$

Reference Efron (2012), "Tweedie's formula and selection bias", *JASA*

1.6 Empirical Bayes

With $y \sim \mathcal{N}(\mu, \sigma^2)$ and $\mu \sim \pi(\mu)$, Tweedie's formula gives posterior expectation

$$\hat{\theta} = E\{\mu \mid y\} = y + \sigma^2 l'(y);$$

y is the MLE of μ so we can think of this as

$$\hat{\theta} = \text{MLE} + \text{Bayes correction}.$$

That's fine if we know the prior $\pi(\mu)$, but what if not? In some situations, where we have many parallel experiments observed at the same time, we can effectively learn $\pi(\mu)$ from the data. This is the *empirical Bayes* approach, as illustrated next.

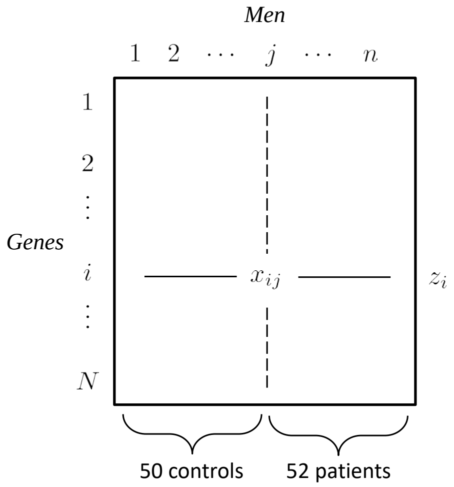
A microarray analysis

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

$$x_{ij} = \begin{cases} i = 1, 2, \ldots, N & \text{genes,} \\ j = 1, 2, \ldots, n & \text{men.} \end{cases}$$

There were:

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



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

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

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. (Note: in terms of our previous notation, $y = z_i$ and $\mu = \delta_i$.)

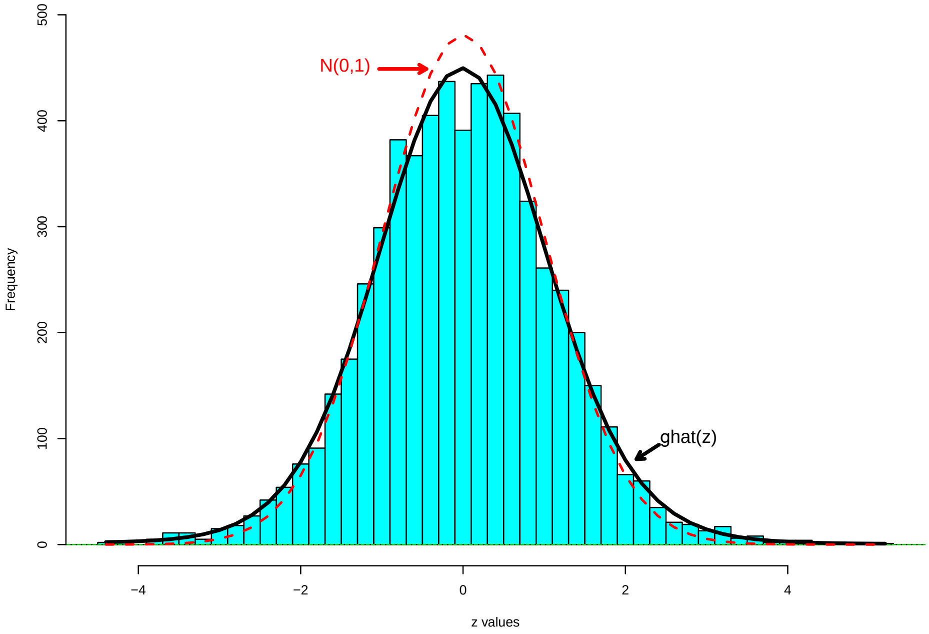


Figure 1.3: Prostate data microarray study. 6033 z-values; heavy curve is $\hat{g}(z)$ from GLM fit; dashed line is $\mathcal{N}(0,1)$.

A reasonable model is

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

where δ_i is the *effect size* for gene i. The investigators were looking for genes with large values of δ_i, either positive or negative. Figure 1.3 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 ($\delta_i = 0$) genes. Which ones and how much?

Empirical Bayes analysis

1.1 Compute $z_1, z_2, \dots, z_N$; $N = 6033$.

1.2 Fit a smooth parametric estimate $\hat{g}(z)$ to histogram (details in Part 2).

1.3 Compute

$$\hat{\lambda}(z) = \log\left[\hat{g}(z)/g_0(z)\right] \qquad \left(g_0(z) = \frac{1}{\sqrt{2\pi}}\, e^{-1/2 z^2}\right).$$

1.4 Differentiate $\hat{\lambda}(z)$ to give Tweedie estimates

$$\hat{E}\{\delta \mid z\} = \hat{\lambda}'(z) \quad \text{and} \quad \widehat{\mathrm{Var}}\{\delta \mid z\} = \hat{\lambda}''(z).$$

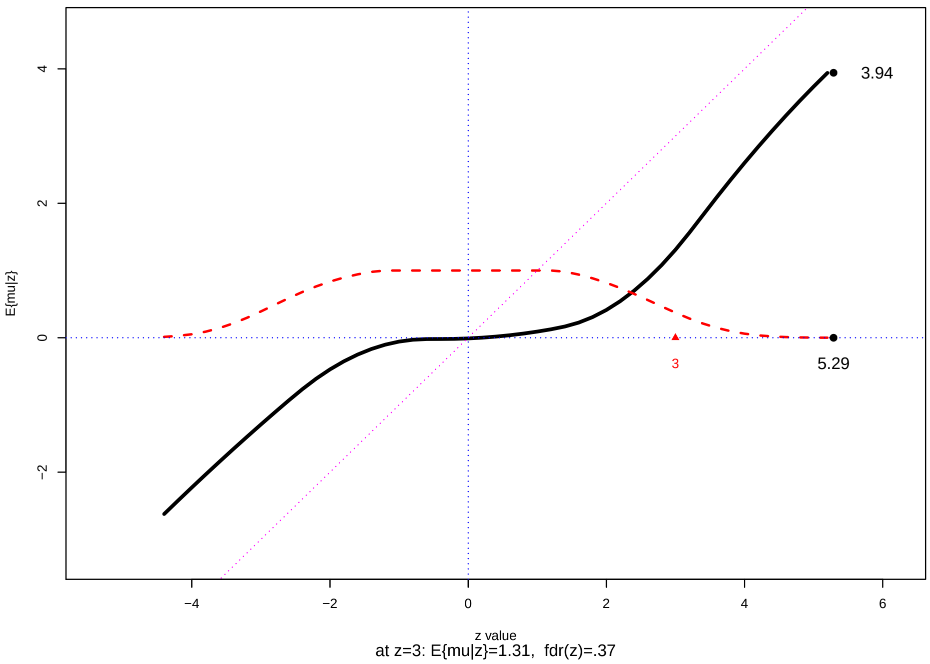


Figure 1.4: Tweedie estimate of $E\{\mu \mid z\}$, prostate study. Dashed curve is estimated local false discovery rate $\mathrm{fdr}(z)$.

Figure 1.4 shows $\hat{E}\{\delta \mid z\}$. It is near zero ("nullness") for $|z| \le 2$. At $z = 3$, $\hat{E}\{\delta \mid z\} = 1.31$. At $z = 5.29$, the largest observed z_i value (gene #610), $E\{\delta \mid z\} = 3.94$.

The "winner's curse" (regression to the mean)

Even though each z_i is unbiased for its δ_i, it isn't true that $z_{i_{\max}}$ is unbiased for $\delta_{i_{\max}}$ ($i_{\max} = 610$ here). The empirical Bayes estimates $\hat{\delta}_i = \hat{E}\{\delta_i \mid z_i\}$ help correct for the winner's curse ("selection bias"), moving the estimates for the extreme z_i values closer to zero.

False discovery rates

Let π_0 be the *prior* probability of a null gene, i.e., $\delta = 0$. The "local false discovery rate" is the *posterior* null probability,

$$\mathrm{fdr}(z) = \Pr\{\delta = 0 \mid z\}.$$

Homework 1.19. (a) Show that

$$\mathrm{fdr}(z_i) = \pi_0 g_0(z_i)/g(z_i),$$

where $g(\cdot)$ is the marginal density. (b) In the normal case $z \sim \mathcal{N}(\delta, 1)$, what is the relationship between $\text{fdr}(z)$ and $E\{\delta \mid z\}$?

In practice, $\text{fdr}(z)$ is often estimated by

$$\widehat{\text{fdr}}(z) = g_0(z)/\hat{g}(z),$$

setting $\pi_0 = 1$, an upper bound. This is the dashed curve in Figure 1.4.

1.7 Some basic statistical results

This section briefly reviews some basic statistical results on estimation and testing as they apply to exponential families. A good reference is Lehmann and Romano's *Theory of Point Estimation* (2008), 3rd edition, from Springer.

Maximum likelihood and Fisher information

We observe a random sample $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ from a member $g_\eta(y)$ of an exponential family $\mathcal{G}$,

$$y_i \overset{\text{iid}}{\sim} g_\eta(y), \qquad i = 1, 2, \dots, n.$$

According to (1.2) in Section 1.3, the density of $\boldsymbol{y}$ is

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

where $\bar{y} = \sum_1^n y_i/n$. The log likelihood function $l_\eta(\boldsymbol{y}) = \log g_\eta^{\boldsymbol{Y}}(\boldsymbol{y})$, $\boldsymbol{y}$ 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

$$\dot{l}_\eta(y) = n(\bar{y} - \mu) \tag{1.7}$$

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

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

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,

$$\begin{aligned} \frac{\partial}{\partial\mu}\, l_\eta(\boldsymbol{y}) &= \dot{l}_\eta(\boldsymbol{y})\frac{\partial\eta}{\partial\mu} = \dot{l}_\eta(\boldsymbol{y})/V \\ &= n(\bar{y}-\mu)/V. \end{aligned} \tag{1.8}$$

This gives

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

which shows that the MLE of μ is

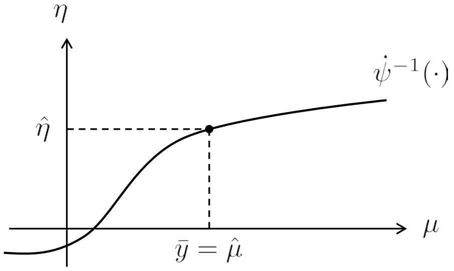
$$\hat{\mu} = \bar{y}.$$

But $\mu = \dot{\psi}(\eta)$, a monotone one-to-one function, so, 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 Section 1.4,

$$\hat{\eta} = \log\frac{\hat{\pi}}{1-\hat{\pi}} \qquad \text{where } \hat{\pi} = y/N.$$



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, and writing simply i_η for the case $n = 1$. The information for μ is

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

using (1.8), the notation being understood as the information for μ in a sample of size n, evaluated at $g_\eta(\boldsymbol{y})$. As always, V stands for V_η, the variance of a single observation y from $g_\eta(\cdot)$.

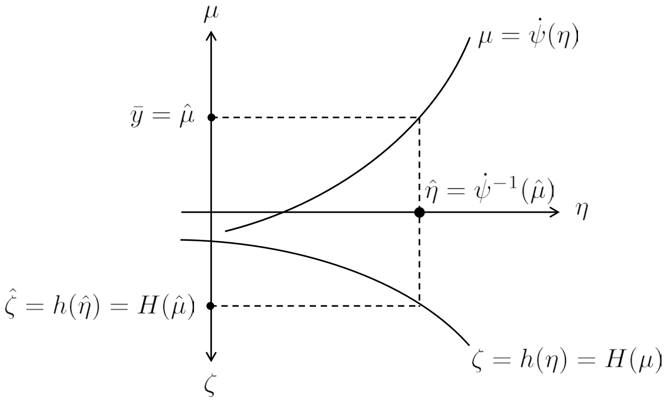
Let $\zeta = h(\eta)$ be any function of η, also expressed as, say,

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

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

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

The figure and table which follow show the MLE and information relationships.


μ
$\mu = \dot{\psi}(\eta)$
$\bar{y} = \hat{\mu}$
$\hat{\eta} = \dot{\psi}^{-1}(\hat{\mu})$
η
$\hat{\zeta} = h(\hat{\eta}) = H(\hat{\mu})$
$\zeta = h(\eta) = H(\mu)$
ζ

Score Functions

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

$$\mu: \quad \frac{\partial l_\eta(\boldsymbol{y})}{\partial \mu} = \frac{n(\bar{y} - \mu)}{V}$$

$$\zeta: \quad \frac{\partial l_\eta(\boldsymbol{y})}{\partial \zeta} = \frac{n(\bar{y} - \mu)}{\dot{h}(\eta)}$$

Fisher Information

$$i_\eta^{(n)} = \text{Var}_\eta \left[\dot{l}_\eta(\boldsymbol{y})\right] = nV = ni_\eta$$

$$i_\eta^{(n)}(\mu) = \frac{n}{V} = ni_\eta(\mu)$$

$$i_\eta^{(n)}(\zeta) = \frac{nV}{\dot{h}(\eta)^2} = ni_\eta(\zeta)$$

In general the Fisher information i_θ for a one-parameter family $f_\theta(x)$ has two expressions, in terms of the 1st and 2nd derivatives of the log likelihood,

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

For $i_\eta^{(n)}$, the Fisher information for η in $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$, we have

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

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

Homework 1.20. (a) Does

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

(b) Does

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

Cramér–Rao lower bound

The CRLB for an unbiased estimator $\bar{\zeta}$ for a general parameter ζ is

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

For $\zeta \equiv \mu$,

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

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.

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

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

A more general form of the CRLB gives

$$\mathrm{Var}_\eta(\hat{\zeta}) \geq \frac{\left[\dot{h}(\eta) + \dot{B}(\eta)/n\right]^2}{nV_\eta} = \frac{\dot{h}(\eta)^2}{nV_\eta} + O\left(\frac{1}{n^2}\right).$$

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

Delta method

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

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

Homework 1.21. Show that if $\zeta = h(\eta) = H(\mu)$, 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 $1/i_\eta^{(n)}(\zeta)$. (In practice we must substitute $\hat{\eta}$ for η in order to estimate $\mathrm{Var}_\eta(\hat{\zeta})$.)

Hypothesis testing (Lehmann)

Suppose we wish to test

$$H_0: \quad \eta = \eta_0 \quad \text{versus} \quad H_A: \eta = \eta_1 \qquad (\eta_1 > \eta_0).$$

- $\log\{g_{\eta_1}(y)/g_{\eta_0}(y)\} = (\eta_1 - \eta_0)y - [\psi(\eta_1) - \psi(\eta_0)] \uparrow y$

- By the Neyman–Pearson lemma, MP_α test rejects for $y \geq Y_0^{(1-\alpha)}$ where $Y_0^{(1-\alpha)}$ is $(1-\alpha)$th quantile of Y under H_0.
- This doesn't depend on η_1, so the test is UMP_α.
- For non-exponential families, such as Cauchy translation family, the MP_α test depends on η_1: "A one-parameter exponential family is a straight line through the space of probability distributions." (Efron 1975, *Ann. Statist.* pp. 1189-1281)

1.8 Deviance and Hoeffding's formula

Deviance is an analogue of Euclidean distance applied to exponential families $g_\eta(y) = e^{\eta y - \psi(\eta)} g_0(y)$. By definition the deviance $D(\eta_1, \eta_2)$ between g_{η_1} and g_{η_2} in family $\mathcal{G}$ is

$$\begin{aligned} D(\eta_1,\eta_2) &= 2E_{\eta_1}\left\{\log\left(\frac{g_{\eta_1}(y)}{g_{\eta_2}(y)}\right)\right\} \\ &= 2\int_{\mathcal{Y}} g_{\eta_1}(y)\log\left[g_{\eta_1}(y)/g_{\eta_2}(y)\right]\ m(dy). \end{aligned}$$

We will also write $D(\mu_1, \mu_2)$ or just $D(1,2)$; the deviance is the distance between the two densities, not their indices.

Homework 1.22. Show that $D(\eta_1, \eta_2) \geq 0$, with strict inequality unless the two densities are identical.

Note. In general, $D(\eta_1, \eta_2) \neq D(\eta_2, \eta_1)$.

Older name

The "Kullback–Leibler distance" 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 1.23. Verify these formulas for the deviance:

$$\begin{aligned} \text{Poisson } Y \sim \mathrm{Poi}(\mu): \quad & D(\mu_1,\mu_2) = 2\mu_1\left[\log\left(\frac{\mu_1}{\mu_2}\right) - \left(1 - \frac{\mu_2}{\mu_1}\right)\right] \\ \text{Binomial } Y \sim \mathrm{Bi}(N,\pi): \quad & D(\pi_1,\pi_2) = 2N\left[\pi_1\log\left(\frac{\pi_1}{\pi_2}\right) + (1-\pi_1)\log\left(\frac{1-\pi_1}{1-\pi_2}\right)\right] \\ \text{Normal } Y \sim \mathcal{N}(\mu,1): \quad & D(\mu_1,\mu_2) = (\mu_1-\mu_2)^2 \\ \text{Gamma } Y \sim \lambda G_N: \quad & D(\lambda_1,\lambda_2) = 2N\left[\log\left(\frac{\lambda_2}{\lambda_1}\right) + \left(\frac{\lambda_1}{\lambda_2} - 1\right)\right] \\ & \qquad\qquad\quad = 2N\left[\log\left(\frac{\mu_2}{\mu_1}\right) + \left(\frac{\mu_1}{\mu_2} - 1\right)\right] \end{aligned}$$

Hoeffding's formula

Let $\hat{\eta}$ be the MLE of η having observed y. Then

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

Indexing the family with the expectation parameter μ rather than η, and remembering that $\hat{\mu} = y$, we get a more memorable version of Hoeffding's formula,

$$\begin{aligned} g_\mu(y) &= g_{\hat{\mu}}(y)e^{-D(\hat{\mu},\mu)/2} \\ &= g_y(y)e^{-D(y,\mu)/2}. \end{aligned} \tag{1.9}$$

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

$$\log\left[g_\mu(y)\right] = \log\left[g_y(y)\right] - D(y,\mu)/2.$$

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

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

Therefore

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

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

Repeated sampling

If $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ is an iid sample from $g_\eta(\cdot)$ then the deviance based on $\boldsymbol{y}$, say $D_n(\eta_1,\eta_2)$, is

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

(This fact shows up in the binomial, Poisson, and gamma cases of Homework 1.16.)

Note. We are indexing the possible distributions of $\boldsymbol{Y}$ with η, not $\eta^{(n)} = n\eta$.

Homework 1.24. What is the deviance formula for the negative binomial family?

Relationship with Fisher information

For η_2 near η, 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):

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

Proof.

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

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

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

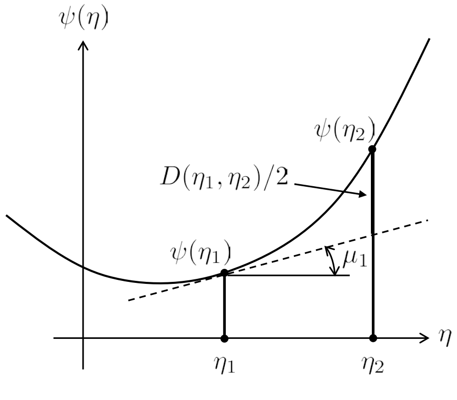
so

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

Homework 1.25. What is $\partial^3 D(\eta_1, \eta_2)/\partial \eta_2^3$?

An informative picture

$\psi(\eta)$ is a convex function of η since $\ddot{\psi}(\eta) = V_\eta > 0$. The picture 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 $\psi(\eta_2) - \psi(\eta_1) + (\eta_1 - \eta_2)\mu_1 = D(\eta_1, \eta_2)/2$.

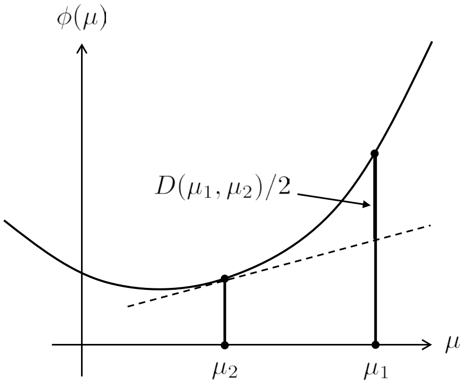


The previous picture, unlike our other results, depends on parameterizing the deviance as $D(\eta_1, \eta_2)$. A version that uses $D(\mu_1, \mu_2)$ depends on the *dual function* $\phi(y)$ to $\psi(y)$,

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

Reference Efron (1978), "Geometry of exponential families", *Ann. Statist.*

Homework 1.26. Show that (a) $\phi(\mu) = \eta\mu - \psi(\eta)$, where $\mu = \dot{\psi}(\eta)$; (b) $\phi(\mu)$ is convex as a function of μ; and (c) $d\phi(\mu)/d\mu = \eta$. (d) Verify the picture at right.



Homework 1.27. Parametric bootstrap: we resample y^* from $g_{\hat{\eta}}(\cdot)$, $\hat{\eta}$ = MLE based on y. Show that

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

Deviance residuals

The idea: if $D(y,\mu)$ is the 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 a normal residual $y-\mu$.

We will work in the repeated sampling framework

$$y_i \overset{\text{iid}}{\sim} g_\mu(\cdot), \qquad i = 1,2,\dots,n,$$

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

$$R = \text{sign}(\bar{y}-\mu)\sqrt{D_n(\bar{y},\mu)}. \tag{1.10}$$

The hope is that R will be nearly $\mathcal{N}(0,1)$, closer to normal than the 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, *Generalized Linear Models*.

Theorem 2. *The asymptotic distribution of R as $n \to \infty$ is*

$$R \dot{\sim} \mathcal{N}\left[-a_n, (1+b_n)^2\right], \tag{1.11}$$

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

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

The normal approximation in (1.11) *is accurate through $O_p(1/n)$, with errors of order $O_p(1/n^{3/2})$,* *e.g.,*

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

(so-called "third order accuracy").

Corollary 1.

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

χ_1^2 a chi-squared random variable with degrees of freedom 1. Since

$$D_n(\bar{y},\mu) = 2\log\left[g_{\hat{\mu}}^{\boldsymbol{Y}}(\boldsymbol{y})/g_{\mu}^{\boldsymbol{Y}}(\boldsymbol{y})\right]$$

according to Hoeffding's formula, the corollary is an improved version of Wilks' theorem, i.e., $2\log(g_{\hat{\mu}}/g_\mu) \to \chi_1^2$ in one-parameter situations.

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

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

Since $a_n = O({}^{1}\!/\!{}_{\sqrt{n}})$ and $b_n = O({}^{1}\!/\!{}_{n})$, the expectation correction in (1.11) is more important than the variance correction.

Homework 1.28. Gamma case, $y \sim \lambda G_N$ with N fixed (N can be thought of as n). (a) Show that the deviance residual $\operatorname{sign}(y-\lambda N)\sqrt{D(y,\lambda N)}$ has the same distribution for all choices of λ. (b) What is the skewness of the Pearson residual $(y-\lambda N)/\lambda\sqrt{N}$?

Homework 1.29. Use our previous results to show that

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

An example

Figure 1.5 shows the results of 2000 replications of $y \sim G_5$ (or equivalently,

$$\bar{y} = \sum_1^5 y_i/5,$$

where y_i are independent G_1 variates, that is, standard one-sided exponentials). The qq-plot shows the deviance residuals (black) much closer to $\mathcal{N}(0,1)$ than the Pearson residuals (red).

Homework 1.30. Compute a version of Figure 1.5 applying to $y \sim \text{Poi}(16)$.

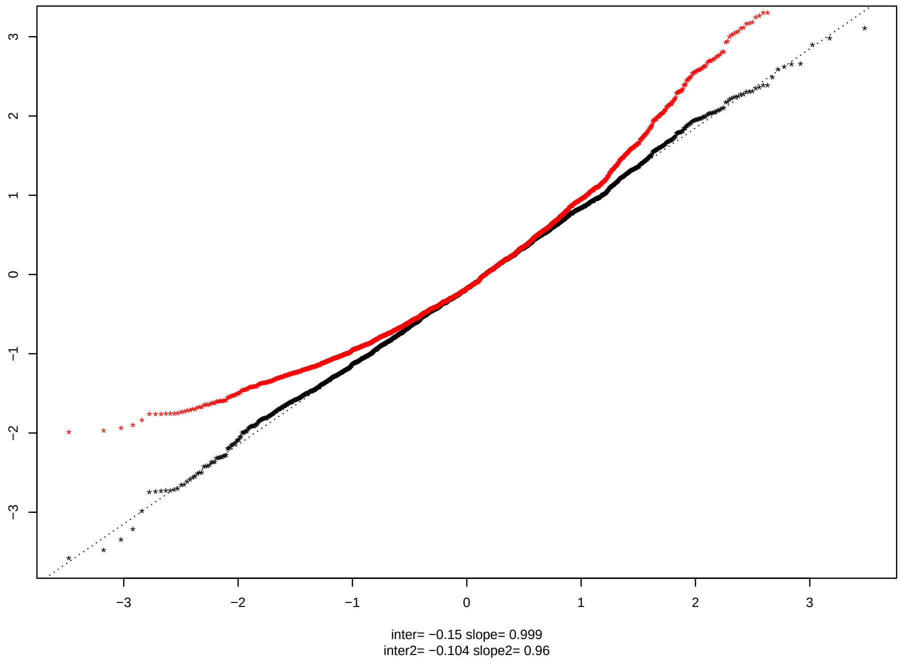


Figure 1.5: qq comparison of deviance residuals (black) with Pearson residuals (red); gamma $N = 1$, $\lambda = 1$, $n = 5$; $B = 2000$ simulations.

An example of Poisson deviance analysis

Reference Thisted and Efron (1987), "Did Shakespeare write a newly discovered poem?", *Biometrika*

- A newly discovered poem is of total length 429 words, comprising 258 different words. An analysis is done to test the hypothesis that Shakespeare wrote the poem.

- 9 of the 258 words never appeared in the 884,647 total words of known Shakespeare; 7 of the 258 words appeared once each in the known Shakespeare, etc., as presented in column "y" of the table.

- A simple theory predicts 6.97 for the expected number of "new" words, given Shakespearean authorship, 4.21 "once before" words, etc., presented in column "ν" of the table. The theory also predicts independent Poisson distributions for the y values,

$$y_i \overset{\text{ind}}{\sim} \text{Poi}(\nu_i) \qquad \text{for } i = 1, 2, \ldots, 11.$$

- "Dev" shows the Poisson deviances; the total deviance 19.98 is moderately large compared to a chi-squared distribution with 11 degrees of freedom, $P\{\chi^2_{11} > 19.98\} = 0.046$. This casts some moderate doubt on Shakespearean authorship.

- "R" is the signed square root of the deviance; "a_n" is the correction $1/6 \times \nu^{1/2}$ suggested by the theorem (1.11); "RR" is the corrected residual $R + a_n$. These should be approximately

$\mathcal{N}(0,1)$ under the hypothesis of Shakespearean authorship. The residual for 20–29 looks suspiciously large.

- 8 out of 11 of the RR's are positive, suggesting that the y's may be systematically larger than the ν's. Adding up the 11 cases,

$$y^+ = 118, \qquad \nu^+ = 94.95.$$

This gives $D^+ = \text{Dev}(y^+, \nu^+) = 5.191$, $R^+ = 2.278$, and $RR^+ = 2.295$. The normal probability of exceeding 2.295 is 0.011, considerably stronger evidence (but see the paper). The actual probability is

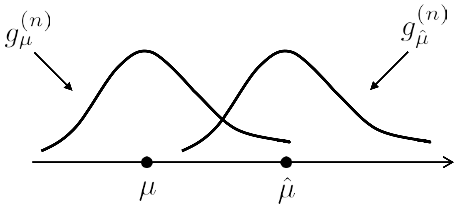
$$\Pr\{\text{Poi}(94.95) \geq 118\} = 0.011.$$

# Prev	y	ν	Dev	R	a_n	RR
0	9	6.97	.5410	.736	.0631	.799
1	7	4.21	1.5383	1.240	.0812	1.321
2	5	3.33	.7247	.851	.0913	.943
3–4	8	5.36	1.1276	1.062	.0720	1.134
5–9	11	10.24	.0551	.235	.0521	.287
10–19	10	13.96	1.2478	−1.117	.0446	−1.072
20–29	21	10.77	7.5858	2.754	.0508	2.805
30–39	16	8.87	4.6172	2.149	.0560	2.205
40–59	18	13.77	1.1837	1.088	.0449	1.133
60–79	8	9.99	.4257	−.652	.0527	−.600
80–99	5	7.48	.9321	−.965	.0609	−.904

1.9 The saddlepoint approximation

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

$$y_1, y_2, \ldots, y_n \overset{\text{iid}}{\sim} g_\mu(\cdot)$$



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

The normal approximation

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

is likely to be inaccurate if $\hat{\mu}$ is say several standard errors removed from μ. Hoeffding's formula gives a much better result, called the *saddlepoint approximation*:

$$\begin{aligned} g_\mu^{(n)}(\hat\mu) &= g_{\hat\mu}^{(n)}(\hat\mu)e^{-D_n(\hat\mu,\mu)/2} \qquad [D_n(\hat\mu,\mu) = nD(\hat\mu,\mu)] \\ &\boxed{\doteq \sqrt{\frac{n}{2\pi V_{\hat\mu}}}\, e^{-D_n(\hat\mu,\mu)/2}} \end{aligned} \tag{1.12}$$

Here $V_{\hat\mu} = \ddot\psi(\hat\eta)$, the variance of a single y_i if $\mu = \hat\mu$.

The approximation

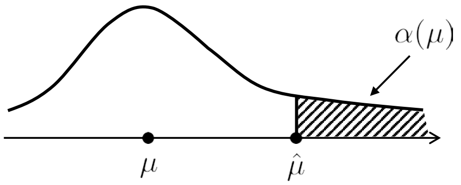
$$g_{\hat\mu}^{(n)}(\hat\mu) \doteq \sqrt{\frac{n}{2\pi V_{\hat\mu}}}$$

comes from applying the central limit theorem *at the center of the $g_{\hat\mu}^{(n)}(\cdot)$ distribution*, just where it is most accurate. There is an enormous literature of extensions and improvements to the saddlepoint approximation: a good review article is Reid (1988) in *Statistical Science*.

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:

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

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

$$Q = \sqrt{nV_{\hat\mu}} \cdot (\hat\eta - \eta)$$

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

Homework 1.31. Suppose we observe $y \sim \lambda G_N$, G_N gamma df $= N$, with $N = 10$ and $\lambda = 1$. Use the L-R formula to calculate $\alpha(\mu)$ for $y = \hat\mu = 15, 20, 25, 30$, and compare the results with the exact values. (You can use any function above for R.)

Homework 1.32. Another version of the L-R formula is

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

where

$$R' = R + \frac{1}{R} \log\left(\frac{Q}{R}\right).$$

How does this relate to the first form?

Large deviations and exponential tilting

In a generic "large deviations" problem, we observe an iid sample

$$y_1, y_2, \dots, y_n \overset{\text{iid}}{\sim} g_0(\cdot)$$

from a known density g_0 having mean and standard deviation

$$y_i \sim (\mu_0, \sigma_0).$$

We wish to compute

$$\alpha_n(\mu) = \Pr{}_0\{\bar{y} \geq \mu\}$$

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

Homework 1.33 ("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 $\alpha_n(\mu) = \Pr_0\{\bar{y} \geq \mu\}$ satisfies

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

(b) Show that $\beta_n(\mu)$ is minimized at $\lambda = \eta$, the value of λ corresponding to μ.

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

$$\Pr{}_0\{\bar{y} \geq \mu\} \leq e^{-nD(\mu,0)},$$

where $D(\mu, 0)$ is the deviance between $g_\eta(y)$ and $g_0(y)$.

Notice that for fixed μ, $\alpha_n(\mu) \to 0$ exponentially fast, which is typical for large deviation results.

Homework 1.34. Extra credit: Suppose $g_0(y) = 1$ for y in $[0, 1]$ and 0 otherwise. Calculate the Chernoff bound for $\Pr_0\{\bar{y} \geq 0.9\}$.

1.10 Transformation theory

Reference Hougaard (1982), *JRSS-B*; DiCiccio (1984) *Biometrika*; Efron (1982), *Ann. Statist.*

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 $y_1, y_2, \dots, y_n$, we might first change to $z_i = 2\sqrt{y_i}$ and then employ standard linear model methods. (That's *not* how we will proceed in Part 2, where generalized linear model techniques are introduced.)

The following display summarizes an enormous number of transformation results for one-parameter exponential families. Let

$$\zeta = H(\mu)$$

and likewise $Z = H(Y)$ and $\hat{\zeta} = H(\hat{\mu})$. The choice of transformation $H(\cdot)$ satisfying

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

then results in:

δ	1/3	1/2	2/3
result	normal likelihood	stabilized variance	normal density

The stabilized variance result follows from the delta method:

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

implies that

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

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

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

gives

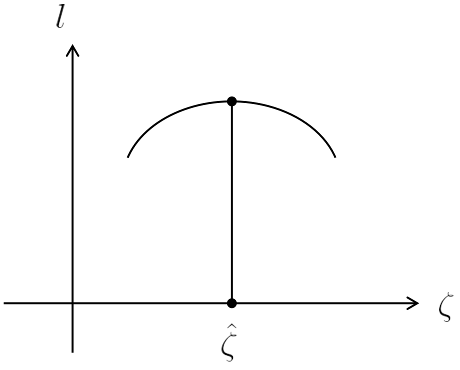
$$H(\mu) = 2\sqrt{\mu} + \text{any constant}$$

as above.

“Normal likelihood” means that the transformation $\hat\zeta = H(\hat\mu)$ results in

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

where $l_\mu(y) = \log g_\mu(y)$, the densities indexed by μ. This makes the log likelihood look parabolic near its maximum at $\zeta = \hat\zeta$. For the Poisson the transformation is $H' = V^{-2/3} = \mu^{-2/3}$, or



$$H(\mu) = 3\mu^{1/3} + \text{constant}.$$

“Normal density” means that $\hat\zeta = H(\hat\mu) \mathrel{\dot\sim} \mathcal{N}(0,1)$. For the Poisson $H' = \mu^{-1/3}$ or

$$H(\mu) = \frac{3}{2}\,\mu^{2/3} + \text{constant} \qquad \text{(makes skewness } \hat\zeta \doteq 0\text{)}.$$

One sees all three transformations $2\mu^{1/2}$, $3\mu^{1/3}$, and ${}^3\!/\!{}_2\mu^{2/3}$ referred to as “the” transformation for the Poisson.

Homework 1.35. Numerically compare the three transformations for the Poisson for $n = 5, 10, 15$, 20, and 25.

Our transformation results apply to any sample size n, with $V_\mu^{(n)} = V_\mu/n$. Verification of the normal density and normal likelihood cases appear in Efron (1982).

Homework 1.36. We observe independent χ^2 variables

$$\hat\sigma_i^2 \sim \sigma_i^2\chi^2_{\nu_i}/\nu_i,$$

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 the previous results table.