5

Parametric Models and Exponential Families

We have been reviewing classic approaches to statistical inference—frequentist, Bayesian, and Fisherian—with an eye toward examining their strengths and limitations in modern applications. Putting philosophical differences aside, there is a common methodological theme in classical statistics: a strong preference for low-dimensional parametric models; that is, for modeling data-analysis problems using parametric families of probability densities (3.1),

$$\mathcal{F} = \{f_\mu(x); x \in \mathcal{X}, \mu \in \Omega\}, \tag{5.1}$$

where the dimension of parameter μ is small, perhaps no greater than 5 or 10 or 20. The inverted nomenclature "nonparametric" suggests the predominance of classical parametric methods.

Two words explain the classic preference for parametric models: mathematical tractability. In a world of sliderules and slow mechanical arithmetic, mathematical formulation, by necessity, becomes the computational tool of choice. Our new computation-rich environment has unplugged the mathematical bottleneck, giving us a more realistic, flexible, and far-reaching body of statistical techniques. But the classic parametric families still play an important role in computer-age statistics, often assembled as small parts of larger methodologies (as with the generalized linear models of Chapter 8). This chapter[1] presents a brief review of the most widely used parametric models, ending with an overview of exponential families, the great connecting thread of classical theory and a player of continuing importance in computer-age applications.

[1] This chapter covers a large amount of technical material for use later, and may be reviewed lightly at first reading.

5.1 Univariate Families

Univariate parametric families, in which the sample space $\mathcal{X}$ of observation x is a subset of the real line $\mathcal{R}^1$, are the building blocks of most statistical analyses. Table 5.1 names and describes the five most familiar univariate families: normal, Poisson, binomial, gamma, and beta. (The chi-squared distribution with n degrees of freedom χ^2_n is also included since it is distributed as $2 \cdot \text{Gam}(n/2, 1)$.) The normal distribution $\mathcal{N}(\mu, \sigma^2)$ is a shifted and scaled version of the $\mathcal{N}(0, 1)$ distribution[2] used in (3.27),

$$\mathcal{N}(\mu, \sigma^2) \sim \mu + \sigma\mathcal{N}(0, 1). \tag{5.2}$$

Table 5.1 *Five familiar univariate densities, and their sample spaces $\mathcal{X}$, parameter spaces Ω, and expectations and variances; chi-squared distribution with n degrees of freedom is* $2\,\text{Gam}(n/2, 1)$.

Name, Notation	Density	$\mathcal{X}$	Ω	Expectation, Variance
Normal $\mathcal{N}(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$	$\mathcal{R}^1$	$\mu \in \mathcal{R}^1$ $\sigma^2 > 0$	μ σ^2
Poisson $\text{Poi}(\mu)$	$\frac{e^{-\mu}\mu^x}{x!}$	$\{0, 1, \dots\}$	$\mu > 0$	μ μ
Binomial $\text{Bi}(n, \pi)$	$\frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}$	$\{0, 1, \dots, n\}$	$0 < \pi < 1$	$n\pi$ $n\pi(1-\pi)$
Gamma $\text{Gam}(\nu, \sigma)$	$\frac{x^{\nu-1}e^{-x/\sigma}}{\sigma^\nu\Gamma(\nu)}$	$x \geq 0$	$\nu > 0$ $\sigma > 0$	$\sigma\nu$ $\sigma^2\nu$
Beta $\text{Be}(\nu_1, \nu_2)$	$\frac{\Gamma(\nu_1+\nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} x^{\nu_1-1}(1-x)^{\nu_2-1}$	$0 \leq x \leq 1$	$\nu_1 > 0$ $\nu_2 > 0$	$\nu_1/(\nu_1+\nu_2)$ $\frac{\nu_1\nu_2}{(\nu_1+\nu_2)^2(\nu_1+\nu_2+1)}$

Relationships abound among the table's families. For instance, independent gamma variables $\text{Gam}(\nu_1, \sigma)$ and $\text{Gam}(\nu_2, \sigma)$ yield a beta variate according to

$$\text{Be}(\nu_1, \nu_2) \sim \frac{\text{Gam}(\nu_1, \sigma)}{\text{Gam}(\nu_1, \sigma) + \text{Gam}(\nu_2, \sigma)}. \tag{5.3}$$

The binomial and Poisson are particularly close cousins. A $\text{Bi}(n, \pi)$ distribution (the number of heads in n independent flips of a coin with probabil-

[2] The notation in (5.2) indicates that if $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y \sim \mathcal{N}(0, 1)$ then X and $\mu + \sigma Y$ have the same distribution.

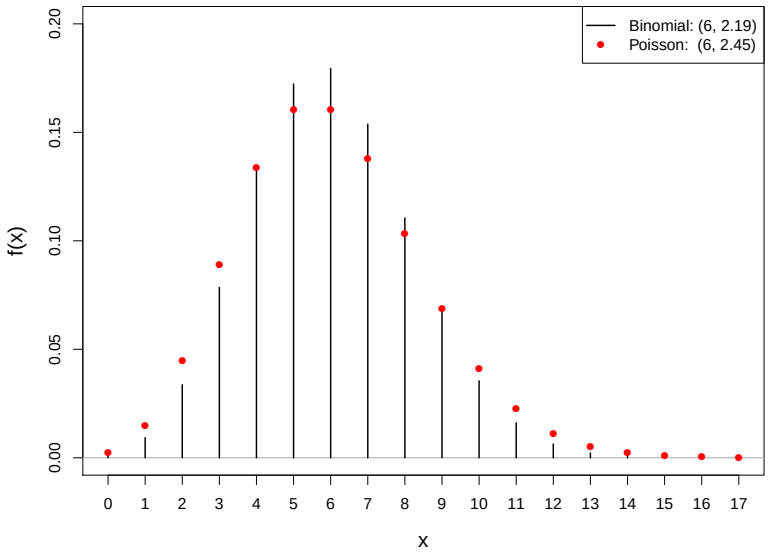

Figure 5.1 Comparison of the binomial distribution Bi(30, 0.2) (black lines) with the Poisson Poi(6) (red dots). In the legend we show the mean and standard deviation for each distribution.

ity of heads π) approaches a Poi($n\pi$) distribution,

$$\text{Bi}(n, \pi) \doteq \text{Poi}(n\pi) \tag{5.4}$$

as n grows large and π small, the notation $\doteq$ indicating approximate equality of the two distributions. Figure 5.1 shows the approximation already working quite effectively for $n = 30$ and $\pi = 0.2$.

The five families in Table 5.1 have five different sample spaces, making them appropriate in different situations. Beta distributions, for example, are natural candidates for modeling continuous data on the unit interval $[0, 1]$. Choices of the two parameters (ν_1, ν_2) provide a variety of possible shapes, as illustrated in Figure 5.2. Later we will discuss general exponential families, unavailable in classical theory, that greatly expand the catalog of possible shapes.

5.2 The Multivariate Normal Distribution

Classical statistics produced a less rich catalog of multivariate distributions, ones where the sample space $\mathcal{X}$ exists in $\mathcal{R}^p$, p-dimensional Eu-

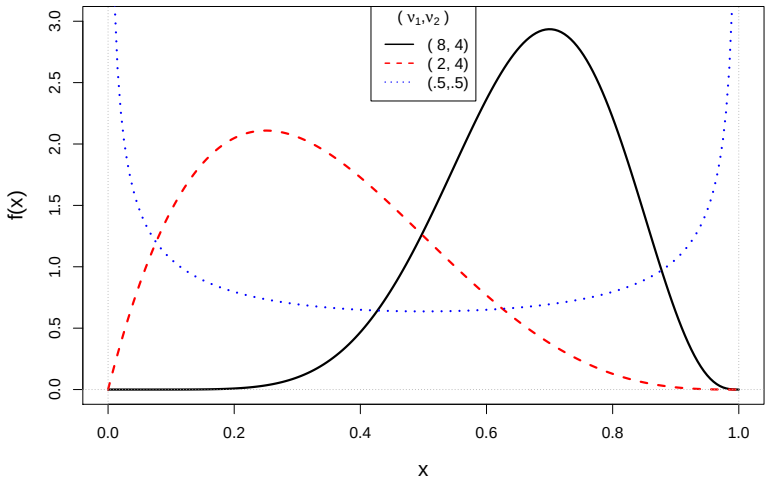

Figure 5.2 Three beta densities, with (ν_1, ν_2) indicated.

clidean space, $p > 1$. By far the greatest amount of attention focused on the multivariate normal distribution.

A random vector $x = (x_1, x_2, \ldots, x_p)'$, normally distributed or not, has *mean vector*

$$\mu = E\{x\} = \big(E\{x_1\}, E\{x_2\}, \ldots, E\{x_p\}\big)' \tag{5.5}$$

and $p \times p$ *covariance matrix*[3]

$$\boldsymbol{\Sigma} = E\left\{(x-\mu)(x-\mu)'\right\} = \left(E\left\{(x_i-\mu_i)(x_j-\mu_j)\right\}\right). \tag{5.6}$$

(The outer product uv' of vectors u and v is the matrix having elements $u_i v_j$.) We will use the convenient notation

$$x \sim (\mu, \boldsymbol{\Sigma}) \tag{5.7}$$

for (5.5) and (5.6), reducing to the familiar form $x \sim (\mu, \sigma^2)$ in the univariate case.

Denoting the entries of $\boldsymbol{\Sigma}$ by σ_{ij}, for i and j equaling $1, 2, \ldots, p$, the diagonal elements are variances,

$$\sigma_{ii} = \text{var}(x_i). \tag{5.8}$$

[3] The notation $\boldsymbol{\Sigma} = (\sigma_{ij})$ defines the ijth element of a matrix.

The off-diagonal elements relate to the correlations between the coordinates of x,

$$\operatorname{cor}(x_i, x_j) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}. \tag{5.9}$$

The multivariate normal distribution extends the univariate definition $\mathcal{N}(\mu, \sigma^2)$ in Table 5.1. To begin with, let $z = (z_1, z_2, \ldots, z_p)'$ be a vector of p independent $\mathcal{N}(0, 1)$ variates, with probability density function

$$f(z) = (2\pi)^{-\frac{p}{2}} e^{-\frac{1}{2}\sum_1^p z_i^2} = (2\pi)^{-\frac{p}{2}} e^{-\frac{1}{2}z'z} \tag{5.10}$$

according to line 1 of Table 5.1.

The multivariate normal family is obtained by linear transformations of z: let μ be a p-dimensional vector and $\boldsymbol{T}$ a $p \times p$ nonsingular matrix, and define the random vector

$$x = \mu + \boldsymbol{T}z. \tag{5.11}$$

Following the usual rules of probability transformations yields the density of x,

$$f_{\mu,\boldsymbol{\Sigma}}(x) = \frac{(2\pi)^{-p/2}}{|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu)'\boldsymbol{\Sigma}^{-1}(x-\mu)}, \tag{5.12}$$

where $\boldsymbol{\Sigma}$ is the $p \times p$ symmetric positive definite matrix

$$\boldsymbol{\Sigma} = \boldsymbol{T}\boldsymbol{T}' \tag{5.13}$$

and $|\boldsymbol{\Sigma}|$ its determinant;[†] $f_{\mu,\boldsymbol{\Sigma}}(x)$, the p-dimensional multivariate normal distribution with mean μ and covariance $\boldsymbol{\Sigma}$, is denoted [†1]

$$x \sim \mathcal{N}_p(\mu, \boldsymbol{\Sigma}). \tag{5.14}$$

Figure 5.3 illustrates the bivariate normal distribution with $\mu = (0, 0)'$ and $\boldsymbol{\Sigma}$ having $\sigma_{11} = \sigma_{22} = 1$ and $\sigma_{12} = 0.5$ (so $\operatorname{cor}(x_1, x_2) = 0.5$). The bell-shaped mountain on the left is a plot of density (5.12). The right panel shows a scatterplot of 2000 points drawn from this distribution. Concentric ellipses illustrate curves of constant density,

$$(x - \mu)'\boldsymbol{\Sigma}^{-1}(x - \mu) = \text{constant}. \tag{5.15}$$

Classical multivariate analysis was the study of the multivariate normal distribution, both of its probabilistic and statistical properties. The notes reference some important (and lengthy) multivariate texts. Here we will just recall a couple of results useful in the chapters to follow.

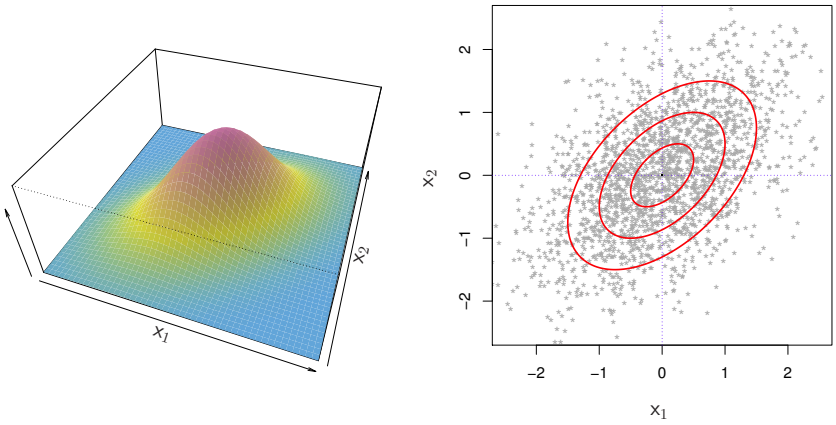

Figure 5.3 *Left:* bivariate normal density, with $\text{var}(x_1) = \text{var}(x_2) = 1$ and $\text{cor}(x_1, x_2) = 0.5$. *Right:* sample of 2000 (x_1, x_2) pairs from this bivariate normal density.

Suppose that $x = (x_1, x_2, \ldots, x_p)'$ is partitioned into

$$x_{(1)} = (x_1, x_2, \ldots, x_{p_1})' \quad \text{and} \quad x_{(2)} = (x_{p_1+1}, x_{p_1+2}, \ldots, x_{p_1+p_2})', \tag{5.16}$$

$p_1 + p_2 = p$, with μ and $\boldsymbol{\Sigma}$ similarly partitioned,

$$\begin{pmatrix} x_{(1)} \\ x_{(2)} \end{pmatrix} \sim \mathcal{N}_p \left(\begin{pmatrix} \mu_{(1)} \\ \mu_{(2)} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right) \tag{5.17}$$

(so $\boldsymbol{\Sigma}_{11}$ is $p_1 \times p_1$, $\boldsymbol{\Sigma}_{12}$ is $p_1 \times p_2$, etc.). Then the conditional distribution of $x_{(2)}$ given $x_{(1)}$ is itself normal,† †2

$$x_{(2)}|x_{(1)} \sim \mathcal{N}_{p_2} \left(\mu_{(2)} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(x_{(1)} - \mu_{(1)}), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \right). \tag{5.18}$$

If $p_1 = p_2 = 1$, then (5.18) reduces to

$$x_2|x_1 \sim \mathcal{N}\left(\mu_2 + \frac{\sigma_{12}}{\sigma_{11}}(x_1 - \mu_1),\ \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}} \right); \tag{5.19}$$

here σ_{12}/σ_{11} is familiar as the linear regression coefficient of x_2 as a function of x_1, while $\sigma_{12}^2/\sigma_{11}\sigma_{22}$ equals $\text{cor}(x_1, x_2)^2$, the squared proportion R^2 of the variance of x_2 explained by x_1. Hence we can write the (unexplained) variance term in (5.19) as $\sigma_{22}(1 - R^2)$.

Bayesian statistics also makes good use of the normal family. It helps to begin with the univariate case $x \sim \mathcal{N}(\mu, \sigma^2)$, where now we assume that

the expectation vector itself has a normal prior distribution $\mathcal{N}(M, A)$:

$$\mu \sim \mathcal{N}(M, A) \quad \text{and} \quad x|\mu \sim \mathcal{N}(\mu, \sigma^2). \tag{5.20}$$

Bayes' theorem and some algebra show that the posterior distribution of μ having observed x is normal,[†] [†3]

$$\mu|x \sim \mathcal{N}\left(M + \frac{A}{A+\sigma^2}(x-M), \frac{A\sigma^2}{A+\sigma^2}\right). \tag{5.21}$$

The posterior expectation $\hat{\mu}_{\text{Bayes}} = M + (A/(A+\sigma^2))(x-M)$ is a *shrinkage estimator* of μ: if, say, A equals σ^2, then $\hat{\mu}_{\text{Bayes}} = M + (x-M)/2$ is shrunk half the way back from the unbiased estimate $\hat{\mu} = x$ toward the prior mean M, while the posterior variance $\sigma^2/2$ of $\hat{\mu}_{\text{Bayes}}$ is only one-half that of $\hat{\mu}$.

The multivariate version of the Bayesian setup (5.20) is

$$\mu \sim \mathcal{N}_p(M, \boldsymbol{A}) \quad \text{and} \quad x|\mu \sim \mathcal{N}_p(\mu, \boldsymbol{\Sigma}), \tag{5.22}$$

now with M and μ p-vectors, and $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ positive definite $p \times p$ matrices. As indicated in the notes, the posterior distribution of μ given x is then

$$\mu|x \sim \mathcal{N}_p\left(M + \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{\Sigma})^{-1}(x-M),\ \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}\right), \tag{5.23}$$

which reduces to (5.21) when $p = 1$.

5.3 Fisher's Information Bound for Multiparameter Families

The multivariate normal distribution plays its biggest role in applications as a large-sample approximation for maximum likelihood estimates. We suppose that the parametric family of densities $\{f_\mu(x)\}$, normal or not, is smoothly defined in terms of its p-dimensional parameter vector μ. (In terms of (5.1), Ω is a subset of $\mathcal{R}^p$.)

The MLE definitions and results are direct analogues of the single-parameter calculations beginning at (4.14) in Chapter 4. The *score function* $\dot{l}_x(\mu)$ is now defined as the gradient of $\log\{f_\mu(x)\}$,

$$\dot{l}_x(\mu) = \nabla_\mu \{\log f_\mu(x)\} = \left(\ldots, \frac{\partial \log f_\mu(x)}{\partial \mu_i}, \ldots\right)', \tag{5.24}$$

the p-vector of partial derivatives of $\log f_\mu(x)$ with respect to the coordinates of μ. It has mean zero,

$$E_\mu\left\{\dot{l}_x(\mu)\right\} = 0 = (0, 0, 0, \ldots, 0)'. \tag{5.25}$$

By definition, the Fisher information matrix $\boldsymbol{\mathcal{I}}_\mu$ for μ is the $p \times p$ covariance matrix of $\dot{l}_x(\mu)$; using outer product notation,

$$\boldsymbol{\mathcal{I}}_\mu = E_\mu\left\{\dot{l}_x(\mu)\dot{l}_x(\mu)'\right\} = \left(E_\mu\left\{\frac{\partial \log f_\mu(x)}{\partial \mu_i}\frac{\partial \log f_\mu(x)}{\partial \mu_j}\right\}\right). \quad (5.26)$$

The key result is that the MLE $\hat{\mu} = \arg\max_\mu\{f_\mu(x)\}$ has an approximately normal distribution with covariance matrix $\boldsymbol{\mathcal{I}}_\mu^{-1}$,

$$\hat{\mu} \doteq \mathcal{N}_p(\mu, \boldsymbol{\mathcal{I}}_\mu^{-1}). \quad (5.27)$$

Approximation (5.27) is justified by large-sample arguments, say with x an iid sample in $\mathcal{R}^p$, $(x_1, x_2, \ldots, x_n)$, n going to infinity.

Suppose the statistician is particularly interested in μ_1, the first coordinate of μ. Let $\mu_{(2)} = (\mu_2, \mu_3, \ldots, \mu_p)$ denote the other $p-1$ coordinates of μ, which are now "nuisance parameters" as far as the estimation of μ_1 goes. According to (5.27), the MLE $\hat{\mu}_1$, which is the first coordinate of $\hat{\mu}$, has

$$\hat{\mu}_1 \doteq \mathcal{N}\left(\mu_1, (\boldsymbol{\mathcal{I}}_\mu^{-1})_{11}\right), \quad (5.28)$$

where the notation indicates the upper leftmost entry of $\boldsymbol{\mathcal{I}}_\mu^{-1}$.

We can partition the information matrix $\boldsymbol{\mathcal{I}}_\mu$ into the two parts corresponding to μ_1 and $\mu_{(2)}$,

$$\boldsymbol{\mathcal{I}}_\mu = \begin{pmatrix} \mathcal{I}_{\mu 11} & \mathcal{I}_{\mu 1(2)} \\ \mathcal{I}_{\mu(2)1} & \boldsymbol{\mathcal{I}}_{\mu(22)} \end{pmatrix} \quad (5.29)$$

(with $\mathcal{I}_{\mu 1(2)} = \mathcal{I}'_{\mu(2)1}$ of dimension $1\times(p-1)$ and $\boldsymbol{\mathcal{I}}_{\mu(22)}$ $(p-1)\times(p-1)$).

†4 The endnotes show that†

$$(\boldsymbol{\mathcal{I}}_\mu^{-1})_{11} = \left(\mathcal{I}_{\mu 11} - \mathcal{I}_{\mu 1(2)}\boldsymbol{\mathcal{I}}_{\mu(22)}^{-1}\mathcal{I}_{\mu(2)1}\right)^{-1}. \quad (5.30)$$

The subtracted term on the right side of (5.30) is nonnegative, implying that

$$(\boldsymbol{\mathcal{I}}_\mu^{-1})_{11} \geq \mathcal{I}_{\mu 11}^{-1}. \quad (5.31)$$

If $\mu_{(2)}$ were known to the statistician, rather than requiring estimation, then $f_{\mu_1\mu_{(2)}}(x)$ would be a one-parameter family, with Fisher information $\mathcal{I}_{\mu 11}$ for estimating μ_1, giving

$$\hat{\mu}_1 \doteq \mathcal{N}(\mu_1, \mathcal{I}_{\mu 11}^{-1}). \quad (5.32)$$

Comparing (5.28) with (5.32), (5.31) shows that the variance of the MLE $\hat{\mu}_1$ must always increase[4] in the presence of nuisance parameters. † †5

Maximum likelihood, and in fact any form of unbiased or nearly unbiased estimation, pays a nuisance tax for the presence of "other" parameters. Modern applications often involve thousands of *others*; think of regression fits with too many predictors. In some circumstances, biased estimation methods can reverse the situation, using the others to actually improve estimation of a target parameter; see Chapter 6 on empirical Bayes techniques, and Chapter 16 on ℓ_1 regularized regression models.

5.4 The Multinomial Distribution

Second in the small catalog of well-known classic multivariate distributions is the multinomial. The multinomial applies to situations in which the observations take on only a finite number of discrete values, say L of them. The 2×2 ulcer surgery of Table 4.1 is repeated in Table 5.2, now with the cells labeled $1, 2, 3,$ and 4. Here there are $L = 4$ possible outcomes for each patient: (`new`, `success`), (`new`, `failure`), (`old`, `success`), (`old`, `failure`).

Table 5.2 *The ulcer study of Table 4.1, now with the cells numbered 1 through 4 as shown.*

	`success`	`failure`
`new`	1 **9**	2 **12**
`old`	3 **7**	4 **17**

A number n of cases has been observed, $n = 45$ in Table 5.2. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_L)$ be the vector of counts for the L possible outcomes,

$$x_l = \#\{\text{cases having outcome } l\}, \tag{5.33}$$

$\boldsymbol{x} = (9, 12, 7, 17)'$ for the ulcer data. It is convenient to code the outcomes in terms of the coordinate vectors $\boldsymbol{e}_l$ of length L,

$$\boldsymbol{e}_l = (0, 0, \ldots, 0, 1, 0, \ldots, 0)', \tag{5.34}$$

with a 1 in the lth place.

[4] Unless $\mathcal{I}_{\mu 1(2)}$ is a vector of zeros, a condition that amounts to approximate independence of $\hat{\mu}_1$ and $\hat{\mu}_{(2)}$.

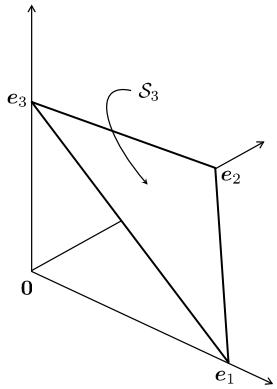

Figure 5.4 The simplex $\mathcal{S}_3$ is an equilateral triangle set at an angle to the coordinate axes in $\mathcal{R}^3$.

The multinomial probability model assumes that the n cases are independent of each other, with each case having probability π_l for outcome e_l,

$$\pi_l = \Pr\{\boldsymbol{e}_l\}, \qquad l = 1, 2, \ldots, L. \tag{5.35}$$

Let

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_L)' \tag{5.36}$$

indicate the vector of probabilities. The count vector $\boldsymbol{x}$ then follows the *multinomial distribution*,

$$f_{\boldsymbol{\pi}}(\boldsymbol{x}) = \frac{n!}{x_1! x_2! \ldots x_L!} \prod_{l=1}^{L} \pi_l^{x_l}, \tag{5.37}$$

denoted

$$\boldsymbol{x} \sim \text{Mult}_L(n, \boldsymbol{\pi}) \tag{5.38}$$

(for n observations, L outcomes, probability vector $\boldsymbol{\pi}$).

The parameter space Ω for $\boldsymbol{\pi}$ is the *simplex* $\mathcal{S}_L$,

$$\mathcal{S}_L = \left\{ \boldsymbol{\pi} : \pi_l \geq 0 \text{ and } \sum_{l=1}^{L} \pi_l = 1 \right\}. \tag{5.39}$$

Figure 5.4 shows $\mathcal{S}_3$, an equilateral triangle sitting at an angle to the coordinate axes $\boldsymbol{e}_1, \boldsymbol{e}_2,$ and $\boldsymbol{e}_3$. The midpoint of the triangle $\boldsymbol{\pi} = (1/3, 1/3, 1/3)$

corresponds to a multinomial distribution putting equal probability on the three possible outcomes.

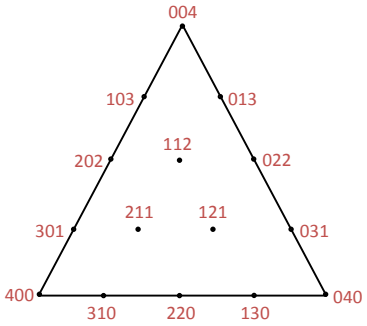

Figure 5.5 Sample space $\mathcal{X}$ for $\boldsymbol{x} \sim \text{Mult}_3(4, \boldsymbol{\pi})$; numbers indicate (x_1, x_2, x_3).

The sample space $\mathcal{X}$ for $\boldsymbol{x}$ is the subset of $n\mathcal{S}_L$ (the set of nonnegative vectors summing to n) having integer components. Figure 5.5 illustrates the case $n = 4$ and $L = 3$, now with the triangle of Figure 5.4 multiplied by 4 and set flat on the page. The point 121 indicates $\boldsymbol{x} = (1, 2, 1)$, with probability $12 \cdot \pi_1 \pi_2^2 \pi_3$ according to (5.37), etc.

In the *dichotomous* case, $L = 2$, the multinomial distribution reduces to the binomial, with (π_1, π_2) equaling $(\pi, 1 - \pi)$ in line 3 of Table 5.1, and (x_1, x_2) equaling $(x, n - x)$. The mean vector and covariance matrix of $\text{Mult}_L(n, \boldsymbol{\pi})$, for any value of L, are[†] †6

$$\boldsymbol{x} \sim \left(n\boldsymbol{\pi}, n\left[\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'\right]\right) \tag{5.40}$$

($\text{diag}(\boldsymbol{\pi})$ is the diagonal matrix with diagonal elements π_l), so $\text{var}(x_l) = n\pi_l(1 - \pi_l)$ and covariance $(x_l, x_j) = -n\pi_l\pi_j$; (5.40) generalizes the binomial mean and variance $(n\pi, n\pi(1 - \pi))$.

There is a useful relationship between the multinomial distribution and the Poisson. Suppose $S_1, S_2, \ldots, S_L$ are independent Poissons having possibly different parameters,

$$S_l \overset{\text{ind}}{\sim} \text{Poi}(\mu_l), \qquad l = 1, 2, \ldots, L, \tag{5.41}$$

or, more concisely,

$$\boldsymbol{S} \sim \text{Poi}(\boldsymbol{\mu}) \tag{5.42}$$

with $\boldsymbol{S} = (S_1, S_2, \ldots, S_L)'$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_L)'$, the independence

being assumed in notation (5.42). Then the conditional distribution of $\boldsymbol{S}$ †7 given the sum $S_+ = \sum S_l$ is multinomial,†

$$\boldsymbol{S}\,|S_+ \sim \text{Mult}_L(S_+, \boldsymbol{\mu}/\mu_+), \tag{5.43}$$

$\mu_+ = \sum \mu_l$.

Going in the other direction, suppose $N \sim \text{Poi}(n)$. Then the unconditional or marginal distribution of $\text{Mult}_L(N, \boldsymbol{\pi})$ is Poisson,

$$\text{Mult}_L(N, \boldsymbol{\pi}) \sim \text{Poi}(n\boldsymbol{\pi}) \qquad \text{if } N \sim \text{Poi}(n). \tag{5.44}$$

Calculations involving $\boldsymbol{x} \sim \text{Mult}_L(n, \boldsymbol{\pi})$ are sometimes complicated by the multinomial's correlations. The approximation $\boldsymbol{x} \doteq \text{Poi}(n\boldsymbol{\pi})$ removes the correlations and is usually quite accurate if n is large.

There is one more important thing to say about the multinomial family: it contains *all* distributions on a sample space $\mathcal{X}$ composed of L discrete categories. In this sense it is a model for *nonparametric* inference on $\mathcal{X}$. The nonparametric bootstrap calculations of Chapter 10 use the multinomial in this way. Nonparametrics, and the multinomial, have played a larger role in the modern environment of large, difficult to model, data sets.

5.5 Exponential Families

Classic parametric families dominated statistical theory and practice for a century and more, with an enormous catalog of their individual properties—means, variances, tail areas, etc.—being compiled. A surprise, though a slowly emerging one beginning in the 1930s, was that all of them were examples of a powerful general construction: *exponential families*. What follows here is a brief introduction to the basic theory, with further development to come in subsequent chapters.

To begin with, consider the Poisson family, line 2 of Table 5.1. The ratio of Poisson densities at two parameter values μ and μ_0 is

$$\frac{f_\mu(x)}{f_{\mu_0}(x)} = e^{-(\mu-\mu_0)}\left(\frac{\mu}{\mu_0}\right)^x, \tag{5.45}$$

which can be re-expressed as

$$f_\mu(x) = e^{\alpha x - \psi(\alpha)} f_{\mu_0}(x), \tag{5.46}$$

where we have defined

$$\alpha = \log\{\mu/\mu_0\} \quad \text{and} \quad \psi(\alpha) = \mu_0(e^\alpha - 1). \tag{5.47}$$

Looking at (5.46), we can describe the Poisson family in three steps.

1 Start with any one Poisson distribution $f_{\mu_0}(x)$.
2 For any value of $\mu > 0$ let $\alpha = \log\{\mu/\mu_0\}$ and calculate

$$\tilde{f}_\mu(x) = e^{\alpha x} f_{\mu_0}(x) \qquad \text{for } x = 0, 1, 2, \ldots. \tag{5.48}$$

3 Finally, divide $\tilde{f}_\mu(x)$ by $\exp(\psi(\alpha))$ to get the Poisson density $f_\mu(x)$.

In other words, we "tilt" $f_{\mu_0}(x)$ with the exponential factor $e^{\alpha x}$ to get $\tilde{f}_\mu(x)$, and then renormalize $\tilde{f}_\mu(x)$ to sum to 1. Notice that (5.46) gives $\exp(-\psi(\alpha))$ as the renormalizing constant since

$$e^{\psi(\alpha)} = \sum_0^\infty e^{\alpha x} f_{\mu_0}(x). \tag{5.49}$$

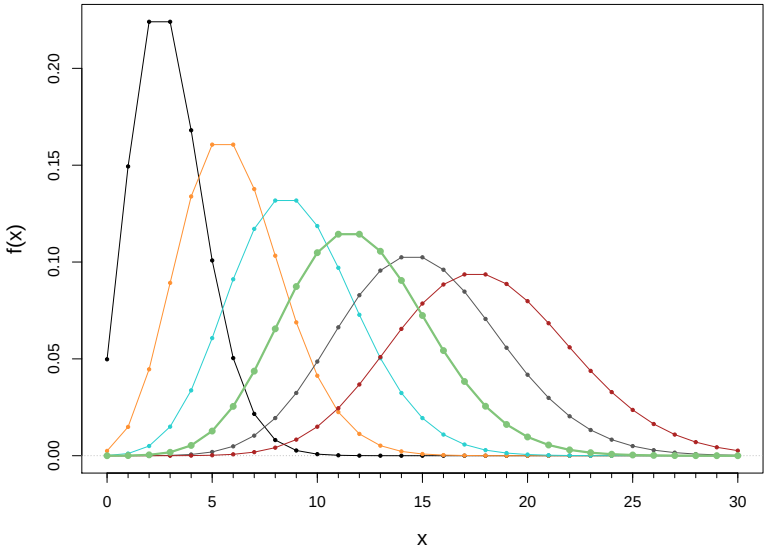

Figure 5.6 Poisson densities for $\mu = 3, 6, 9, 12, 15, 18$; heavy green curve with dots for $\mu = 12$.

Figure 5.6 graphs the Poisson density $f_\mu(x)$ for $\mu = 3, 6, 9, 12, 15, 18$. Each Poisson density is a renormalized exponential tilt of any other Poisson density. So for instance $f_6(x)$ is obtained from $f_{12}(x)$ via the tilt $e^{\alpha x}$ with $\alpha = \log\{6/12\} = -0.693$.[5]

[5] Alternate expressions for $f_\mu(x)$ as an exponential family are available, for example $\exp(\alpha x - \psi(\alpha)) f_0(x)$, where $\alpha = \log \mu$, $\psi(\alpha) = \exp(\alpha)$, and $f_0(x) = 1/x!$. (It isn't necessary for $f_0(x)$ to be a member of the family.)

The Poisson is a *one-parameter exponential family*, in that α and x in expression (5.46) are one-dimensional. A *p-parameter exponential family* has the form

$$f_\alpha(x) = e^{\alpha' y - \psi(\alpha)} f_0(x) \qquad \text{for } \alpha \in A, \tag{5.50}$$

where α and y are p-vectors and A is contained in $\mathcal{R}^p$. Here α is the "canonical" or "natural" parameter vector and $y = t(x)$ is the "sufficient statistic" vector. The normalizing function $\psi(\alpha)$, which makes $f_\alpha(x)$ integrate (or sum) to one, satisfies

$$e^{\psi(\alpha)} = \int_{\mathcal{X}} e^{\alpha' y} f_0(x)\, dx, \tag{5.51}$$

and it can be shown that the parameter space A for which the integral is finite is a convex set[†] in $\mathcal{R}^p$. As an example, the gamma family on line 4 of Table 5.1 is a two-parameter exponential family, with α and $y = t(x)$ given by

†8

$$(\alpha_1, \alpha_2) = \left(-\frac{1}{\sigma}, \nu\right), \quad (y_1, y_2) = (x, \log x), \tag{5.52}$$

and

$$\begin{aligned} \psi(\alpha) &= \nu \log \sigma + \log \Gamma(\nu) \\ &= -\alpha_2 \log\{-\alpha_1\} + \log\{\Gamma(\alpha_2)\}. \end{aligned} \tag{5.53}$$

The parameter space A is $\{\alpha_1 < 0 \text{ and } \alpha_2 > 0\}$.

Why are we interested in exponential tilting rather than some other transformational form? The answer has to do with repeated sampling. Suppose $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is an iid sample from a p-parameter exponential family (5.50). Then, letting $y_i = t(x_i)$ denote the sufficient vector corresponding to x_i,

$$\begin{aligned} f_\alpha(\boldsymbol{x}) &= \prod_{i=1}^{n} e^{\alpha' y_i - \psi(\alpha)} f_0(x_i) \\ &= e^{n(\alpha' \bar{y} - \psi(\alpha))} f_0(\boldsymbol{x}), \end{aligned} \tag{5.54}$$

where $\bar{y} = \sum_1^n y_i/n$. This is still a p-parameter exponential family, now with natural parameter $n\alpha$, sufficient statistic $\bar{y}$, and normalizer $n\psi(\alpha)$. No matter how large n may be, the statistician can still compress all the inferential information into a p-dimensional statistic $\bar{y}$. Only exponential families enjoy this property.

Even though they were discovered and developed in quite different contexts, and at quite different times, all of the distributions discussed in this

chapter exist in exponential families. This isn't quite the coincidence it seems. Mathematical tractability was the prized property of classic parametric distributions, and tractability was greatly facilitated by exponential structure, even if that structure went unrecognized.

In one-parameter exponential families, the normalizer $\psi(\alpha)$ is also known as the *cumulant generating function*. Derivatives of $\psi(\alpha)$ yield the cumulants of y,[6] the first two giving the mean and variance† †9

$$\dot{\psi}(\alpha) = E_\alpha\{y\} \quad \text{and} \quad \ddot{\psi}(\alpha) = \text{var}_\alpha\{y\}. \tag{5.55}$$

Similarly, in p-parametric families

$$\dot{\psi}(\alpha) = (\ldots \partial\psi/\partial\alpha_j \ldots)' = E_\alpha\{y\} \tag{5.56}$$

and

$$\ddot{\psi}(\alpha) = \left(\frac{\partial^2\psi(\alpha)}{\partial\alpha_j\,\partial\alpha_k}\right) = \text{cov}_\alpha\{y\}. \tag{5.57}$$

The p-dimensional *expectation parameter*, denoted

$$\beta = E_\alpha\{y\}, \tag{5.58}$$

is a one-to-one function of the natural parameter α. Let V_α indicate the $p \times p$ covariance matrix,

$$V_\alpha = \text{cov}_\alpha(y). \tag{5.59}$$

Then the $p \times p$ derivate matrix of β with respect to α is

$$\frac{d\beta}{d\alpha} = (\partial\beta_j/\partial\alpha_k) = V_\alpha, \tag{5.60}$$

this following from (5.56)–(5.57), the inverse mapping being $d\alpha/d\beta = V_\alpha^{-1}$. As a one-parameter example, the Poisson in Table 5.1 has $\alpha = \log\mu$, $\beta = \mu$, $y = x$, and $d\beta/d\alpha = 1/(d\alpha/d\beta) = \mu = V_\alpha$.

The maximum likelihood estimate for the expectation parameter β is simply y (or $\bar{y}$ under repeated sampling (5.54)), which makes it immediate to calculate in most situations.† Less immediate is the MLE for the natural †10 parameter α: the one-to-one mapping $\beta = \dot{\psi}(\alpha)$ (5.56) has inverse $\alpha = \dot{\psi}^{-1}(\beta)$, so

$$\hat{\alpha} = \dot{\psi}^{-1}(y), \tag{5.61}$$

[6] The simplified dot notation leads to more compact expressions: $\dot{\psi}(\alpha) = d\psi(\alpha)/d\alpha$ and $\ddot{\psi}(\alpha) = d^2\psi(\alpha)/d\alpha^2$.

e.g., $\hat{\alpha} = \log y$ for the Poisson. The trouble is that $\dot{\psi}^{-1}(\cdot)$ is usually unavailable in closed form. Numerical approximation algorithms are necessary to calculate $\hat{\alpha}$ in most cases.

All of the classic exponential families have closed-form expressions for $\psi(\alpha)$ (and $f_\alpha(x)$), yielding pleasant formulas for the mean β and covariance V_α, (5.56)–(5.57). Modern computational technology allows us to work with general exponential families, designed for specific tasks, without concern for mathematical tractability.

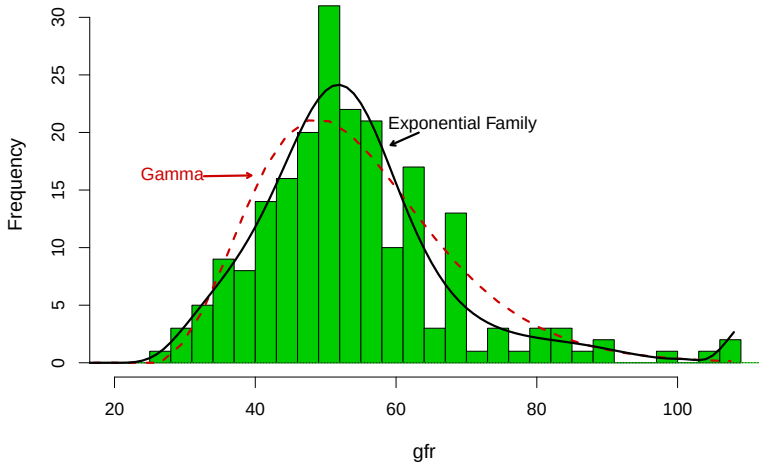

Figure 5.7 A seven-parameter exponential family fit to the **gfr** data of Figure 2.1 (solid) compared with gamma fit of Figure 4.1 (dashed).

As an example we again consider fitting the **gfr** data of Figure 2.1. For our exponential family of possible densities we take $f_0(x) \equiv 1$, and sufficient statistic vector

$$y(x) = (x, x^2, \ldots, x^7), \tag{5.62}$$

so $\alpha' y$ in (5.50) can represent all 7th-order polynomials in x, the **gfr** measurement.[7] (Stopping at power 2 gives the $\mathcal{N}(\mu, \sigma^2)$ family, which we already know fits poorly from Figure 4.1.) The heavy curve in Figure 5.7 shows the MLE fit $f_{\hat{\alpha}}(x)$ now following the **gfr** histogram quite closely. Chapter 15 discusses "Lindsey's method," a simplified algorithm for calculating the MLE $\hat{\alpha}$.

[7] Any intercept in the polynomial is absorbed into the $\psi(\alpha)$ term in (5.57).

A more exotic example concerns the generation of random graphs on a fixed set of N nodes. Each possible graph has a certain total number E of edges, and T of triangles. A popular choice for generating such graphs is the two-parameter exponential family having $y = (E, T)$, so that larger values of α_1 and α_2 yield more connections.

5.6 Notes and Details

The notion of *sufficient statistics*, ones that contain all available inferential information, was perhaps Fisher's happiest contribution to the classic corpus. He noticed that in the exponential family form (5.50), the fact that the parameter α interacts with the data x only through the factor $\exp(\alpha' y)$ makes $y(x)$ sufficient for estimating α. In 1935–36, a trio of authors, working independently in different countries, Pitman, Darmois, and Koopmans, showed that exponential families are the only ones that enjoy fixed-dimensional sufficient statistics under repeated independent sampling. Until the late 1950s such distributions were called Pitman–Darmois–Koopmans families, the long name suggesting infrequent usage.

Generalized linear models, Chapter 8, show the continuing impact of sufficiency on statistical practice. Peter Bickel has pointed out that *data compression*, a lively topic in areas such as image transmission, is a modern, less stringent, version of sufficiency.

Our only nonexponential family so far was (4.39), the Cauchy translational model. Efron and Hinkley (1978) analyze the Cauchy family in terms of *curved exponential families*, a generalization of model (5.50).

Properties of classical distributions (lots of properties and lots of distributions) are covered in Johnson and Kotz's invaluable series of reference books, 1969–1972. Two classic multivariate analysis texts are Anderson (2003) and Mardia *et al.* (1979).

†1 [p. 57] *Formula* (5.12). From $z = \boldsymbol{T}^{-1}(x - \mu)$ we have $dz/dx = \boldsymbol{T}^{-1}$ and

$$f_{\mu,\boldsymbol{\Sigma}}(x) = f(z)|\boldsymbol{T}^{-1}| = (2\pi)^{-\frac{p}{2}}|\boldsymbol{T}^{-1}|e^{-\frac{1}{2}(x-\mu)'\boldsymbol{T}^{-1'}\boldsymbol{T}^{-1}(x-\mu)}, \quad (5.63)$$

so (5.12) follows from $\boldsymbol{T}\boldsymbol{T}' = \boldsymbol{\Sigma}$ and $|\boldsymbol{T}| = |\boldsymbol{\Sigma}|^{1/2}$.

†2 [p. 58] *Formula* (5.18). Let $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ be partitioned as in (5.17). Then

$$\begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix} = \begin{pmatrix} \left(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)^{-1} & -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Lambda}_{22} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Lambda}_{11} & \left(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right)^{-1} \end{pmatrix}, \quad (5.64)$$

direct multiplication showing that $\boldsymbol{\Lambda}\boldsymbol{\Sigma} = \boldsymbol{I}$, the identity matrix. If $\boldsymbol{\Sigma}$ is

symmetric then $\mathbf{\Lambda}_{21} = \mathbf{\Lambda}'_{12}$. By redefining x to be $x - \mu$ we can set $\mu_{(1)}$ and $\mu_{(2)}$ equal to zero in (5.18). The quadratic form in the exponent of (5.12) is

$$(x'_{(1)}, x'_{(2)})\mathbf{\Lambda}\left(x_{(1)}, x_{(2)}\right) = x'_{(2)}\mathbf{\Lambda}_{22}x_{(2)} + 2x'_{(1)}\mathbf{\Lambda}_{12}x_{(2)} + x'_{(1)}\mathbf{\Lambda}_{11}x_{(1)}. \tag{5.65}$$

But, using (5.64), this matches the quadratic form from (5.18),

$$\left(x_{(2)} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}x_{(1)}\right)'\mathbf{\Lambda}_{22}\left(x_{(2)} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}x_{(1)}\right) \tag{5.66}$$

except for an added term that does *not* involve $x_{(2)}$. For a multivariate normal distribution, this is sufficient to show that the conditional distribution of $x_{(2)}$ given $x_{(1)}$ is indeed (5.18) (see †3).

†3 [p. 59] *Formulas* (5.21) *and* (5.23). Suppose that the continuous univariate random variable z has density of the form

$$f(z) = c_0 e^{-\frac{1}{2}Q(z)}, \qquad \text{where } Q(z) = az^2 + 2bz + c_1, \tag{5.67}$$

a, b, c_0 and c_1 constants, $a > 0$. Then, by "completing the square,"

$$f(z) = c_2 e^{-\frac{1}{2}a\left(z-\frac{b}{a}\right)^2}, \tag{5.68}$$

and we see that $z \sim \mathcal{N}(b/a, 1/a)$. The key point is that form (5.67) specifies z as normal, with mean and variance uniquely determined by a and b. The multivariate version of this fact was used in the derivation of formula (5.18).

By redefining μ and x as $\mu - M$ and $x - M$, we can take $M = 0$ in (5.21). Setting $B = A/(A + \sigma^2)$, density (5.21) for $\mu|x$ is of form (5.67), with

$$Q(\mu) = \frac{\mu^2}{B\sigma^2} - \frac{2x\mu}{\sigma^2} + \frac{Bx^2}{\sigma^2}. \tag{5.69}$$

But Bayes' rule says that the density of $\mu|x$ is proportional to $g(\mu)\, f_\mu(x)$, also of form (5.67), now with

$$Q(\mu) = \left(\frac{1}{A} + \frac{1}{\sigma^2}\right)\mu^2 - \frac{2x\mu}{\sigma^2} + \frac{x^2}{\sigma^2}. \tag{5.70}$$

A little algebra shows that the quadratic and linear coefficients of μ match in (5.69)–(5.70), verifying (5.21).

We verify the multivariate result (5.23) using a different argument. The $2p$ vector $(\mu, x)'$ has joint distribution

$$\mathcal{N}\left(\begin{pmatrix} M \\ M \end{pmatrix}, \begin{pmatrix} \boldsymbol{A} & \boldsymbol{A} \\ \boldsymbol{A} & \boldsymbol{A} + \mathbf{\Sigma} \end{pmatrix}\right). \tag{5.71}$$

Now we employ (5.18) and a little manipulation to get (5.23).

†4 [p. 60] *Formula* (5.30). This is the matrix identity (5.64), now with $\boldsymbol{\Sigma}$ equaling $\mathcal{I}_\mu$.

†5 [p. 61] *Multivariate Gaussian and nuisance parameters.* The cautionary message here—that increasing the number of unknown nuisance parameters decreases the accuracy of the estimate of interest—can be stated more positively: if some nuisance parameters are actually known, then the MLE of the parameter of interest becomes more accurate. Suppose, for example, we wish to estimate μ_1 from a sample of size n in a bivariate normal model $x \sim \mathcal{N}_2(\mu, \boldsymbol{\Sigma})$ (5.14). The MLE $\bar{x}_1$ has variance σ_{11}/n in notation (5.19). But if μ_2 is known then the MLE of μ_1 becomes $\bar{x}_1 - (\sigma_{12}/\sigma_{22})(\bar{x}_2 - \mu_2)$ with variance $(\sigma_{11}/n)\cdot(1-\rho^2)$, ρ being the correlation $\sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$.

†6 [p. 63] *Formula* (5.40). $\boldsymbol{x} = \sum_{i=1}^{n} \boldsymbol{x}_i$, where the $\boldsymbol{x}_i$ are iid observations having $\Pr\{\boldsymbol{x}_i = \boldsymbol{e}_i\} = \pi_l$, as in (5.35). The mean and covariance of each $\boldsymbol{x}_i$ are

$$E\{\boldsymbol{x}_i\} = \sum_{1}^{L} \pi_l \boldsymbol{e}_l = \boldsymbol{\pi} \tag{5.72}$$

and

$$\begin{aligned} \operatorname{cov}\{\boldsymbol{x}_i\} &= E\{\boldsymbol{x}_i \boldsymbol{x}_i'\} - E\{\boldsymbol{x}_i\}E\{\boldsymbol{x}_i'\} = \sum \pi_l \boldsymbol{e}_l \boldsymbol{e}_l' - \boldsymbol{\pi}\boldsymbol{\pi}' \\ &= \operatorname{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'. \end{aligned} \tag{5.73}$$

Formula (5.40) follows from $E\{\boldsymbol{x}\} = \sum E\{\boldsymbol{x}_i\}$ and $\operatorname{cov}(\boldsymbol{x}) = \sum \operatorname{cov}(\boldsymbol{x}_i)$.

†7 [p. 64] *Formula* (5.43). The densities of $\boldsymbol{S}$ (5.42) and $S_+ = \sum S_l$ are

$$f_{\boldsymbol{\mu}}(\boldsymbol{S}) = \prod_{l=1}^{L} e^{-\mu_l} \mu_l^{S_l} / S_l! \quad \text{and} \quad f_{\mu_+}(S_+) = e^{-\mu_+} \mu_+^{S_+} / S_+!. \tag{5.74}$$

The conditional density of $\boldsymbol{S}$ given S_+ is the ratio

$$f_{\boldsymbol{\mu}}(\boldsymbol{S}|S_+) = \left(\frac{S_+!}{\prod_1^L S_l!}\right) \prod_{l=1}^{L} \left(\frac{\mu_l}{\mu_+}\right)^{S_l}, \tag{5.75}$$

which is (5.43).

†8 [p. 66] *Formula* (5.51) *and the convexity of A.* Suppose α_1 and α_2 are any two points in A, i.e., values of α having the integral in (5.51) finite. For any value of c in the interval $[0, 1]$, and any value of y, we have

$$c e^{\alpha_1' y} + (1-c) e^{\alpha_2' y} \geq e^{[c\alpha_1 + (1-c)\alpha_2]' y} \tag{5.76}$$

because of the convexity in c of the function on the right (verified by showing that its second derivative is positive). Integrating both sides of (5.76)

over $\mathcal{X}$ with respect to $f_0(x)$ shows that the integral on the right must be finite: that is, $c\alpha_1 + (1-c)\alpha_2$ is in A, verifying A's convexity.

†9 [p. 67] *Formula* (5.55). In the univariate case, differentiating both sides of (5.51) with respect to α gives

$$\dot{\psi}(\alpha)e^{\psi(\alpha)} = \int_{\mathcal{X}} y e^{\alpha y} f_0(x)\, dx; \tag{5.77}$$

dividing by $e^{\psi(\alpha)}$ shows that $\dot{\psi}(\alpha) = E_\alpha\{y\}$. Differentiating (5.77) again gives

$$\left(\ddot{\psi}(\alpha) + \dot{\psi}(\alpha)^2\right) e^{\psi(\alpha)} = \int_{\mathcal{X}} y^2 e^{\alpha y} f_0(x)\, dx, \tag{5.78}$$

or

$$\ddot{\psi}(\alpha) = E_\alpha\{y^2\} - E_\alpha\{y\}^2 = \mathrm{var}_\alpha\{y\}. \tag{5.79}$$

Successive derivatives of $\psi(\alpha)$ yield the higher cumulants of y, its skewness, kurtosis, etc.

†10 [p. 67] *MLE for* β. The gradient with respect to α of $\log f_\alpha(y)$ (5.50) is

$$\nabla_\alpha\left(\alpha' y - \psi(\alpha)\right) = y - \dot{\psi}(\alpha) = y - E_\alpha\{y^*\}, \tag{5.80}$$

(5.56), where y^* represents a hypothetical realization $y(x^*)$ drawn from $f_\alpha(\cdot)$. We achieve the MLE $\hat{\alpha}$ at $\nabla_{\hat{\alpha}} = 0$, or

$$E_{\hat{\alpha}}\{y^*\} = y. \tag{5.81}$$

In other words the MLE $\hat{\alpha}$ is the value of α that makes the expectation $E_\alpha\{y^*\}$ match the observed y. Thus (5.58) implies that the MLE of parameter β is y.