5

Models

Chapter 4 described methods related to a central notion in inference, namely likelihood. This chapter and the next discuss how those ideas apply to some particular situations, beginning with the simplest model for the dependence of one variable on another, straight-line regression. There is then an account of exponential family distributions, which include many models commonly used in practice, such as the normal, exponential, gamma, Poisson and binomial densities, and which play a central role in statistical theory. We then briefly describe group transformation models, which are also important in statistical theory. This is followed by a description of models for data in the form of lifetimes, which are common in medical and industrial settings, and a discussion of missing data and the EM algorithm.

5.1 Straight-Line Regression

We have already met situations where we focus on how one variable depends on others. In such problems there are two or more variables, some of which are regarded as fixed, and others as random. The random quantities are known as *responses* and the fixed ones as *explanatory variables*. We shall suppose that only one variable is regarded as a response. Such models, known as regression models, are discussed extensively in Chapters 8, 9, and 10. Here we outline the basic results for the simplest regression model, where a single response depends linearly on a single covariate. We start with an example.

Example 5.1 (Venice sea level data) Table 5.1 and Figure 5.1 show annual maximum sea levels in Venice for 1931–1981. The most obvious feature is that the maximum sea level increased by about 25 cm over that period. A simple model is of linear trend in the sea level, y, so in year j,

$$y_j = \beta_0 + \beta_1 j + \varepsilon_j, \tag{5.1}$$

where β_0 (cm) represents the expected maximum sea level in year $j = 0$, β_1 the annual increase (cm/year) , and ε_j is a random variable with mean zero and variance

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

103	78	121	116	115	147	119	114	89	102
99	91	97	106	105	136	126	132	104	117
151	116	107	112	97	95	119	124	118	145
122	114	118	107	110	194	138	144	138	123
122	120	114	96	125	124	120	132	166	134
138									

Table 5.1 Annual maximum sea levels (cm) in Venice, 1931–1981 (Pirazzoli, 1982). To be read across rows.

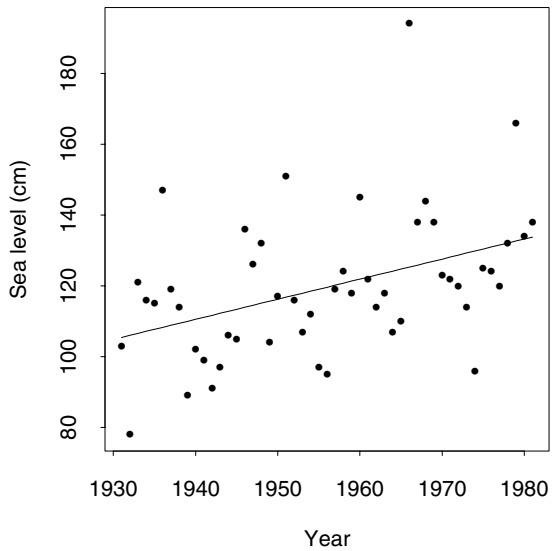

Figure 5.1 Annual maximum sea levels in Venice, 1931–1981, with fitted regression line.

σ^2 (cm^2) representing scatter about the trend. Here the response is sea level, y_j, and the year, j, is the sole explanatory variable. ■

The simplest linear model is that independent random variables Y_j satisfy

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, \ldots, n, \tag{5.2}$$

where the x_j are known constants, the $\varepsilon_j \overset{\text{iid}}{\sim} N(0, \sigma^2)$, and β_0, β_1 and σ^2 are unknown parameters, Thus Y_j is normal with mean $\beta_0 + \beta_1 x_j$ and variance σ^2. The data arise as pairs $(x_1, y_1), \ldots, (x_n, y_n)$, from which β_0, β_1, and σ^2 are to be estimated. In Example 5.1 the pairs are (1931, 103), ..., (1981, 138). If all the x_j are equal, we cannot estimate the slope of the dependence of y on x, so we assume that at least two x_j are distinct.

$\overset{\text{iid}}{\sim}$ means 'are independent and identically distributed as'.

A reparametrization of (5.2) is more convenient, so we consider instead

$$Y_j = \gamma_0 + \gamma_1(x_j - \overline{x}) + \varepsilon_j, \quad j = 1, \ldots, n, \tag{5.3}$$

where $\overline{x} = n^{-1} \sum x_j$. In terms of the original parameters, $\gamma_1 = \beta_1$, and $\gamma_0 = \beta_0 + \beta_1 \overline{x}$. This can make better statistical sense too. In (5.1) the interpretation of β_0 as a mean sea level at the start of the Christian era — when $j = 0$ — involves a ludicrous extrapolation of the straight-line model over two millenia, whereas γ_0 concerns its level when $j = \overline{x} = 1956$; this is clearly more sensible.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Under (5.3) the Y_j are independent and normal with means and variances $\gamma_0 + \gamma_1(x_j - \overline{x})$ and σ^2, so the likelihood based on $(x_1, y_1), \ldots, (x_n, y_n)$ is

$$\prod_{j=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}\{y_j - \gamma_0 - \gamma_1(x_j - \overline{x})\}^2\right],$$
$$-\infty < \gamma_0, \gamma_1 < \infty, \sigma^2 > 0.$$

The log likelihood is

$$\ell(\gamma_0, \gamma_1, \sigma^2) \equiv -\frac{1}{2}\left[n \log \sigma^2 + \frac{1}{\sigma^2}\sum_{j=1}^{n}\{y_j - \gamma_0 - \gamma_1(x_j - \overline{x})\}^2\right]. \tag{5.4}$$

For any σ^2, maximizing this over γ_0 and γ_1 is equivalent to minimizing the *sum of squares*

$$SS(\gamma_0, \gamma_1) = \sum_{j=1}^{n}\{y_j - \gamma_0 - \gamma_1(x_j - \overline{x})\}^2,$$

which is the sum of squared vertical deviations between the y_j and their means $\gamma_0 + \gamma_1(x_j - \overline{x})$ under the linear model. Its derivatives are

$$\frac{\partial SS}{\partial \gamma_0} = -2\sum_{j=1}^{n}\{y_j - \gamma_0 - \gamma_1(x_j - \overline{x})\},$$
$$\frac{\partial SS}{\partial \gamma_1} = -2\sum_{j=1}^{n}(x_j - \overline{x})\{y_j - \gamma_0 - \gamma_1(x_j - \overline{x})\},$$
$$\frac{\partial^2 SS}{\partial \gamma_0^2} = 2n, \quad \frac{\partial^2 SS}{\partial \gamma_1^2} = 2\sum_{j=1}^{n}(x_j - \overline{x})^2, \quad \frac{\partial^2 SS}{\partial \gamma_0 \partial \gamma_1} = 2\sum_{j=1}^{n}(x_j - \overline{x}) = 0.$$

The solutions to the equations $\partial SS/\partial\gamma_0 = \partial SS/\partial\gamma_1 = 0$ are the *least squares estimates*,

$$\widehat{\gamma}_0 = \overline{y}, \qquad \widehat{\gamma}_1 = \frac{\sum_{j=1}^{n} y_j(x_j - \overline{x})}{\sum_{j=1}^{n}(x_j - \overline{x})^2}. \tag{5.5}$$

As anticipated, γ_1 cannot be estimated if all the x_j are equal, for then $x_j \equiv \overline{x}$ and $\widehat{\gamma}_1$ is undefined. The matrix of second derivatives of SS is positive definite, so the estimates (5.5) minimize the sum of squares and hence maximize $\ell(\gamma_0, \gamma_1, \sigma^2)$ with respect to γ_0 and γ_1.

As the log likelihood may be written as $-\frac{1}{2}\left\{n \log \sigma^2 + SS(\gamma_0, \gamma_1)/\sigma^2\right\}$, the maximum likelihood estimate of σ^2 is

$$\widehat{\sigma}^2 = n^{-1} SS(\widehat{\gamma}_0, \widehat{\gamma}_1) = \frac{1}{n}\sum_{j=1}^{n}\{y_j - \widehat{\gamma}_0 - \widehat{\gamma}_1(x_j - \overline{x})\}^2.$$

The quantity $SS(\widehat{\gamma}_0, \widehat{\gamma}_1)$, known as the *residual sum of squares*, is the smallest sum of squares attainable by fitting (5.3) to the data.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

The least squares estimators are linear combinations of normal variables, so their distributions are also normal. If we rewrite them as

$$\widehat{\gamma}_0 = n^{-1}\sum_{j=1}^{n}\{\gamma_0 + \gamma_1(x_j - \overline{x}) + \varepsilon_j\} = \gamma_0 + n^{-1}\sum_{j=1}^{n}\varepsilon_j,$$

$$\widehat{\gamma}_1 = \frac{\sum_{j=1}^{n}\{\gamma_0 + \gamma_1(x_j - \overline{x}) + \varepsilon_j\}(x_j - \overline{x})}{\sum_{j=1}^{n}(x_j - \overline{x})^2} = \gamma_1 + \frac{\sum_{j=1}^{n}(x_j - \overline{x})\varepsilon_j}{\sum_{j=1}^{n}(x_j - \overline{x})^2},$$

we see that because the ε_j are independent with means zero and variances σ^2, $\widehat{\gamma}_0$ has mean γ_0 and variance σ^2/n, and that $\widehat{\gamma}_1$ has mean γ_1 and variance $\sigma^2/\sum(x_j - \overline{x})^2$. Moreover

$$\begin{aligned}\operatorname{cov}(\widehat{\gamma}_0, \widehat{\gamma}_1) &= \operatorname{cov}\left\{n^{-1}\sum \varepsilon_j, \frac{\sum_{j=1}^{n}(x_j - \overline{x})\varepsilon_j}{\sum_{j=1}^{n}(x_j - \overline{x})^2}\right\} \\ &= \frac{\sum_{j=1}^{n} n^{-1}(x_j - \overline{x})\operatorname{var}(\varepsilon_j)}{\sum_{j=1}^{n}(x_j - \overline{x})^2} = 0 :\end{aligned}$$

as $\widehat{\gamma}_0$ and $\widehat{\gamma}_1$ are uncorrelated normal random variables, they are independent.

If σ^2 is known, confidence intervals for the true values of γ_0 and γ_1 may be based on the normal distributions of $\widehat{\gamma}_0$ and $\widehat{\gamma}_1$. A $(1 - 2\alpha)$ confidence interval for γ_1, for example, is $\widehat{\gamma}_1 \pm \sigma z_\alpha/\{\sum(x_j - \overline{x})^2\}^{1/2}$.

We shall see in Chapter 8 that the residual sum of squares $SS(\widehat{\gamma}_0, \widehat{\gamma}_1) \sim \sigma^2\chi^2_{n-2}$, independent of $\widehat{\gamma}_0$ and $\widehat{\gamma}_1$. Thus when σ^2 is unknown, the estimator

$$S^2 = \frac{1}{n-2}SS(\widehat{\gamma}_0, \widehat{\gamma}_1)$$

satisfies $\mathrm{E}(S^2) = \sigma^2$, and as S^2 is independent of $\widehat{\gamma}_0$ and $\widehat{\gamma}_1$, a $(1 - 2\alpha)$ confidence interval for γ_1 is $\widehat{\gamma}_1 \pm St_{n-2}(\alpha)/\{\sum(x_j - \overline{x})^2\}^{1/2}$, because

$$\frac{\widehat{\gamma}_1 - \gamma_1}{\left\{S^2/\sum(x_j - \overline{x})^2\right\}^{1/2}} \sim t_{n-2}.$$

Example 5.2 (Venice sea level data) For the model $y_j = \beta_0 + \beta_1 j + \varepsilon_j$ of Example 5.1, we have $n = 51, x_1 = 1931, \ldots, x_n = 1981$, so $\overline{x} = 1956$. In parametrization (5.3), γ_0 is the expected annual maximum sea level in 1956 in cm, and γ_1 is the mean annual increase in maximum sea level in cm/year.

Straightforward calculation yields $\widehat{\gamma}_0 = 119.61$ cm and $\widehat{\gamma}_1 = 0.567$ cm/year, $SS(\widehat{\gamma}_0, \widehat{\gamma}_1) = 16988.1$, and $\sum(x_j - \overline{x})^2 = 11050$. The unbiased estimate of σ^2 is $s^2 = 16988.1/(51 - 2) = 346.7$, so we estimate σ by $s = 18.6$. This is very large relative to the annual increase in sea level, which as we see from Figure 5.1 is small relative to the overall vertical variation.

Standard errors for $\widehat{\gamma}_0$ and $\widehat{\gamma}_1$ are $s/n^{1/2} = 2.61$ and $s/\left\{\sum(x_j - \overline{x})^2\right\}^{1/2} = 0.177$, and a 95% confidence interval for γ_1 is $\widehat{\gamma}_1 \pm 0.177t_{49}(0.025)$, that is, $(0.213, 0.921)$. This does not include zero, confirming that the trend in Figure 5.1 is real. ■

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Linear combinations

Distributional results for linear functions of $\widehat{\gamma}_0$ and $\widehat{\gamma}_1$ are readily obtained. For example, in the original linear model (5.2) we have $\beta_0 = \gamma_0 - \gamma_1\overline{x}$, the maximum likelihood estimator of which is $\widehat{\beta}_0 = \widehat{\gamma}_0 - \widehat{\gamma}_1\overline{x}$. This has expected value $\gamma_0 - \gamma_1\overline{x}$ and variance

$$\mathrm{var}(\widehat{\gamma}_0 - \widehat{\gamma}_1\overline{x}) = \mathrm{var}(\widehat{\gamma}_0) - 2\overline{x}\mathrm{cov}(\widehat{\gamma}_0, \widehat{\gamma}_1) + \overline{x}^2\mathrm{var}(\widehat{\gamma}_1) = \sigma^2\left\{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{j=1}^n (x_j - \overline{x})^2}\right\}.$$

As

$$\mathrm{cov}(\widehat{\beta}_0, \widehat{\beta}_1) = \mathrm{cov}(\widehat{\gamma}_0 - \widehat{\gamma}_1\overline{x}, \widehat{\gamma}_1) = \mathrm{cov}(\widehat{\gamma}_0, \widehat{\gamma}_1) - \overline{x}\mathrm{var}(\widehat{\gamma}_1) = \frac{-\sigma^2\overline{x}}{\sum_{j=1}^n (x_j - \overline{x})^2},$$

the normal random variables $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are independent if and only if $\overline{x} = 0$.

Suppose we wish to predict the response value at x_+,

$$Y_+ = \gamma_0 + \gamma_1(x_+ - \overline{x}) + \varepsilon_+.$$

Here ε_+ represents the random variation about the expected value, which is independent of the other responses, because of our modelling assumptions. The random variable Y_+ has expected value $\gamma_0 + \gamma_1(x_+ - \overline{x})$. The maximum likelihood estimator of this, $\widehat{\gamma}_0 + \widehat{\gamma}_1(x_+ - \overline{x})$, has mean and variance

$$\gamma_0 + \gamma_1(x_+ - \overline{x}), \quad \sigma^2\left\{\frac{1}{n} + \frac{(x_+ - \overline{x})^2}{\sum_{j=1}^n (x_j - \overline{x})^2}\right\}.$$

This is the variance not of Y_+ but of $\widehat{\gamma}_0 + \widehat{\gamma}_1(x_+ - \overline{x})$: it does not account for the extra variability introduced by ε_+. The variance appropriate for the predicted response *actually observed* is

$$\mathrm{var}(Y_+) = \mathrm{var}\{\widehat{\gamma}_0 + \widehat{\gamma}_1(x_+ - \overline{x}) + \varepsilon_+\} = \sigma^2\left\{\frac{1}{n} + \frac{(x_+ - \overline{x})^2}{\sum_{j=1}^n (x_j - \overline{x})^2}\right\} + \sigma^2. \quad (5.6)$$

The final σ^2 is due to ε_+ and would remain even if the parameters were known.

Example 5.3 (Venice sea level data) For illustration we take $x_+ = 1993$. Our predicted value for Y_+ is $\widehat{\gamma}_0 + \widehat{\gamma}_1(x_+ - \overline{x}) = 140.59$, with estimated variance $49.75 + 346.70 = 396.45$, obtained by replacing σ^2 with s^2 in (5.6). The estimated variance of ε_+, 346.70, is much larger than the estimated variance 49.75 of the fitted value $\widehat{\gamma}_0 + \widehat{\gamma}_1(x_+ - \overline{x})$. A confidence interval for Y_+ could be obtained from the t statistic.

Our model (5.2) presupposes that the errors ε_j are normal, and that the dependence of y on x is linear. We discuss how to check these assumptions in Section 8.6.1, here noting that simple estimates of the errors ε_j are the *raw residuals* $e_j = y_j - \widehat{\beta}_0 - \widehat{\beta}_1 x_j$, which should be normal and approximately independent of x if the model is correct. We check linearity by looking for patterns in a plot of the e_j against the x_j, and check normality by a normal probability plot of the e_j; see Figure 5.2. Linearity seems justifiable, but the errors seem too skewed to be normally distributed.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

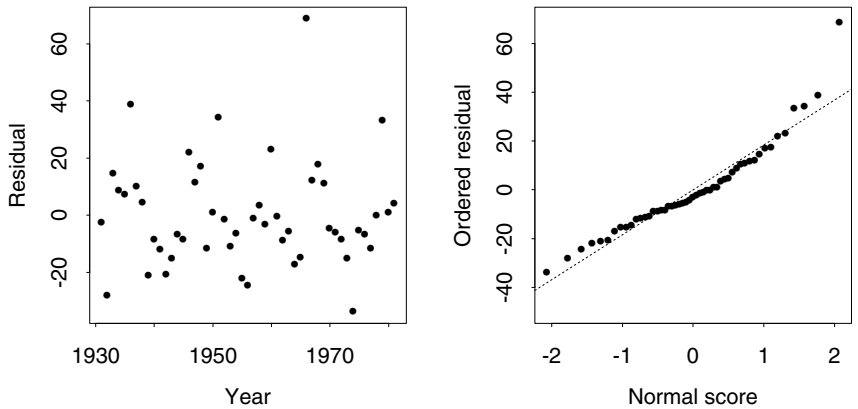

Figure 5.2 Straight-line regression fit to annual maximum sea levels in Venice, 1931–1981. Left: raw residuals plotted against time. Right: normal scores plot of raw residuals; the line has slope $\widehat{\sigma}$. The skewness of the residuals suggests that the errors are not normal.

The astute reader will realise that the changing sea level is due not to the rising waters of the Adriatic, but to the sinking of the marker that measures water height, along with Venice, to which it is attached. ■

Exercises 5.1

1 Find the observed and expected information matrices for the parameters in (5.4), and confirm that general likelihood theory gives the same variances and covariance for the least squares estimates as the direct argument on page 164.

2 Show that $(\widehat{\gamma}_0, \widehat{\gamma}_1, s^2)$ are minimal sufficient for the parameters of the straight-line regression model.

3 Consider data from the straight-line regression model with n observations and

$$x_j = \begin{cases} 0, & j = 1, \ldots, m, \\ 1, & \text{otherwise,} \end{cases}$$

where $m \leq n$. Give a careful interpretation of the parameters β_0 and β_1, and find their least squares estimates. For what value(s) of m is $\text{var}(\widehat{\beta}_1)$ minimized, and for which maximized? Do your results make qualitative sense?

4 Let $Y_1, \ldots, Y_n$ be observations satisfying (5.2), with not all the x_j equal. Find $\text{var}(\widehat{\beta}_0 + x_+\widehat{\beta}_1)$, where x_+ is fixed. Hence give exact 0.95 confidence intervals for $\beta_0 + \beta_1 x_+$ when σ^2 is known and when it is unknown.

5.2 Exponential Family Models

Exponential families include most of the models we have met so far and are widely used in applications. Densities such as the normal, gamma, Poisson, multinomial, and so forth have the same underlying structure with elegant properties giving them a central role in statistical theory. This section outlines those properties, first giving the basic ideas for scalar random variables, then extending them to more complex models, and finally considering inference.

5.2.1 Basic notions

Let $f_0(y)$ be a given probability density, discrete or continuous, under which random variable Y has support $\mathcal{Y} = \{y : f_0(y) > 0\}$ that is a subset of the real line $\mathbb{R}$. For

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

example, $f_0(y)$ might be the uniform density on the unit interval $\mathcal{Y} = (0, 1)$, or might have probability mass function $e^{-1}/y!$ on $\mathcal{Y} = \{0, 1, \ldots\}$. Let $s(Y)$ be a function of Y, and let

When Y is discrete we interpret the integrals as sums over $y \in \mathcal{Y}$.

$$\mathcal{N} = \left\{\theta : \kappa(\theta) = \log \int e^{s(y)\theta} f_0(y)\, dy < \infty\right\}$$

denote the values of θ for which the cumulant-generating function $\kappa(\theta)$ of $s(Y)$ is finite. Evidently $0 \in \mathcal{N}$. To avoid trivial cases we suppose that $\mathcal{N}$ has at least one other element and that $\text{var}\{s(Y)\} > 0$ under f_0, so $s(Y)$ is not a degenerate random variable. In fact the set $\mathcal{N}$ is convex, because if $\theta_1, \theta_2 \in \mathcal{N}$ and $\alpha \in [0, 1]$, then $\alpha\theta_1 + (1 - \alpha)\theta_2 \in \mathcal{N}$:

$$\begin{aligned}\int e^{s(y)\{\alpha\theta_1 + (1-\alpha)\theta_2\}} f_0(y)\, dy &= \int \left\{e^{s(y)\theta_1}\right\}^{\alpha} \left\{e^{s(y)\theta_2}\right\}^{1-\alpha} f_0(y)\, dy \\ &\leq \left\{\int e^{s(y)\theta_1} f_0(y)\, dy\right\}^{\alpha} \left\{\int e^{s(y)\theta_2} f_0(y)\, dy\right\}^{1-\alpha} \\ &< \infty;\end{aligned}$$

the second line follows from Hölder's inequality (Exercise 5.2.1). Moreover, as $\kappa\{\alpha\theta_1 + (1 - \alpha)\theta_2\} \leq \alpha\kappa(\theta_1) + (1 - \alpha)\kappa(\theta_2)$, the function $\kappa(\theta)$ is convex on the set $\mathcal{N}$. Equality occurs only if $\theta_1 = \theta_2$, so in fact $\kappa(\theta)$ is strictly convex.

A single fixed density f_0 is not flexible enough to be useful in practice, for which we need families of distributions. Hence we embed f_0 in the larger class

$$f(y; \theta) = \frac{e^{s(y)\theta} f_0(y)}{\int e^{s(x)\theta} f_0(x)\, dx}, \quad y \in \mathcal{Y}, \theta \in \mathcal{N},$$

by *exponential tilting*: f_0 has been tilted by multiplication by $e^{s(y)\theta}$ and then the resulting positive function has been renormalized to have unit integral. Evidently $f(y; \theta)$ has support $\mathcal{Y}$ for every θ. If $s(Y) = Y$, we have a *natural exponential family of order 1*,

$$f(y; \theta) = \exp\{y\theta - \kappa(\theta)\} f_0(y), \quad y \in \mathcal{Y}, \theta \in \mathcal{N}. \tag{5.7}$$

The family is called *regular* if the *natural parameter space* $\mathcal{N}$ is an open set.

Example 5.4 (Uniform density) Let $f_0(y) = 1$ for $y \in \mathcal{Y} = (0, 1)$. Now

$$\kappa(\theta) = \log \int e^{y\theta} f_0(y)\, dy = \log \int_0^1 e^{y\theta}\, dy = \log\{(e^\theta - 1)/\theta\} < \infty$$

for all $\theta \in \mathcal{N} = (-\infty, \infty)$, and the natural exponential family

$$f(y; \theta) = \begin{cases} \theta e^{\theta y}/(e^\theta - 1), & 0 < y < 1, \\ 0, & \text{otherwise}, \end{cases} \tag{5.8}$$

is plotted in the left panel of Figure 5.3 for $\theta = -3, 0, 1$. For this or any natural exponential family with bounded $\mathcal{Y}$, $\mathcal{N} = (-\infty, \infty)$ and the family is regular.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

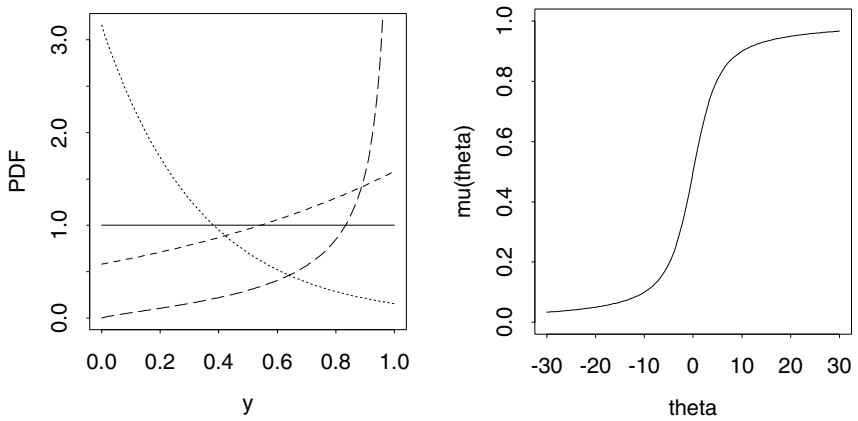

Figure 5.3 Exponential families generated by tilting the $U(0, 1)$ density. Left: original density (solid), natural exponential family when $\theta = -3$ (dots) and $\theta = 1$ (small dashes), and density generated when $s(y) = \log\{y/(1-y)\}$ when $\theta = 3/4$ (large dashes). Right: mean function $\mu(\theta)$ for the natural exponential family.

A different choice of $s(Y)$ will generate a different exponential family. With $s(Y) = \log\{Y/(1-Y)\}$, for example, the cumulant-generating function is given by

$$\begin{aligned}\int_0^1 e^{\theta \log\{y/(1-y)\}}\, dy &= \int_0^1 y^{(1+\theta)-1}(1-y)^{(1-\theta)-1}\, dy \\ &= B(1+\theta, 1-\theta) \\ &= \frac{\Gamma(1+\theta)\Gamma(1-\theta)}{\Gamma(1+\theta+1-\theta)}, \quad |\theta| < 1,\end{aligned}$$

For $a, b > 0$, $B(a, b) = \int_0^1 u^{a-1}(1-u)^{b-1}\, du$ is the beta function. It equals $\Gamma(a)\Gamma(b)/\Gamma(a+b)$, where $\Gamma(a) = \int_0^\infty u^{a-1}e^{-u}\, du$ is the gamma function; see Exercise 2.1.3.

and as $\Gamma(2) = 1$, we have $\kappa(\theta) = \log\Gamma(1+\theta) + \log\Gamma(1-\theta)$. Here the set $\mathcal{N} = (-1, 1)$ is open, so the resulting family is regular. Figure 5.3 shows how this family differs from the natural one, being unbounded unless $\theta = 0$. ■

The natural exponential family of order 1 generated by a tilted version of f_0 is the same as that generated by f_0 itself. To see why, note that if $s(Y)$ has density (5.7) for some $\theta = \theta_1$, say, exponential tilting generates a density proportional to $\exp\{s(y)\theta\}\exp\{s(y)\theta_1 - \kappa(\theta_1)\}f_0(y)$ with cumulant-generating function $\kappa(\theta+\theta_1) - \kappa(\theta_1)$ for $\theta + \theta_1 \in \mathcal{N}$. The new density is $\exp\{s(y)(\theta+\theta_1) - \kappa(\theta+\theta_1)\}f_0(y)$, for $\theta + \theta_1 \in \mathcal{N}$. This is (5.7) apart from replacement of θ by $\theta + \theta_1$. Hence just one family is generated by a specific choice of f_0 and $s(Y)$, and this family is obtained by tilting any of its members.

For many purposes discussion of an exponential family is simplified if it is expressed without reference to a baseline density f_0. If a density may be written as

$$f(y;\omega) = \exp\{s(y)\theta(\omega) - b(\omega) + c(y)\}, \quad y \in \mathcal{Y}, \omega \in \Omega, \tag{5.9}$$

where $\mathcal{Y}$ is independent of the parameter ω and θ is a function of ω, it is said to be an *exponential family of order 1*. Here θ and s are called the *natural parameter* and *natural observation*.

Example 5.5 (Exponential density) The exponential density with mean ω is $f(y;\omega) = \omega^{-1}\exp(-y/\omega)$, for $y > 0$ and $\omega > 0$. Here $\Omega = \mathcal{Y} = (0, \infty)$, with natural observation and parameter $s(y) = y$ and $\theta(\omega) = -1/\omega$, and $b(\omega) = \log\omega$. The cumulant-generating function is $\kappa(\theta) = b\{\omega^{-1}(\theta)\} = -\log(-\theta)$, which has

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

derivatives $(r-1)!(-1)^r\theta^{-r} = (r-1)!\omega^r$, the usual formula for cumulants of an exponential variable. ■

Example 5.6 (Binomial density) If R is binomial with denominator m and probability $0 < \pi < 1$, its density is

$$\binom{m}{r}\pi^r(1-\pi)^{m-r} = \exp\left\{r\log\left(\frac{\pi}{1-\pi}\right) + m\log(1-\pi) + \log\binom{m}{r}\right\},$$

for $r \in \mathcal{Y} = \{0, 1, \ldots, m\}$. This has form (5.9) with $\omega = \pi$,

$$s(r) = r, \quad \theta(\pi) = \log\left(\frac{\pi}{1-\pi}\right), \; b(\pi) = m\log(1-\pi), \; c(r) = \log\binom{m}{r}.$$

The natural parameter is the *log odds* $\theta = \log\{\pi/(1-\pi)\} \in (-\infty, \infty)$. This family is regular, with cumulant-generating function $\kappa(\theta) = m\log(1+e^\theta)$. ■

If the function $\theta(\omega)$ in (5.9) is 1–1, the density of $S = s(Y)$ has form

$\theta(\Omega)$ denotes the set $\{\theta(\omega) : \omega \in \Omega\}$.

$$f(s;\theta) = \exp\left[s\theta - b\left\{\omega^{-1}(\theta)\right\}\right]h(s), \quad s \in s(\mathcal{Y}), \theta \in \theta(\Omega).$$

If $\Theta = \theta(\Omega) = \mathcal{N}$ for some baseline density f_0 then this is a natural exponential family with cumulant-generating function $\kappa(\theta) = b\left\{\omega^{-1}(\theta)\right\}$.

Expressed as a function of θ rather than ω, the moment-generating function of $s(Y)$ under (5.9) is, if finite,

$$\begin{aligned}\mathrm{E}\left\{e^{ts(Y)}\right\} &= \int \exp\left\{ts(y) + \theta s(y) - \kappa(\theta) + c(y)\right\} dy \\ &= \exp\left\{\kappa(\theta+t) - \kappa(\theta)\right\} \int \exp\left\{(\theta+t)y - \kappa(\theta+t) + c(y)\right\} dy \\ &= \exp\left\{\kappa(\theta+t) - \kappa(\theta)\right\},\end{aligned}$$

because the second integral equals unity; here $\theta = \theta(\omega)$ and $\kappa(\theta) = b\left\{\omega^{-1}(\theta)\right\}$. Hence when Y has density (5.9), the cumulant-generating function of $s(Y)$ is $\kappa(\theta+t) - \kappa(\theta)$. The cumulants result from differentiating $\kappa(\theta+t) - \kappa(\theta)$ with respect to t and then setting $t = 0$, or equivalently differentiating $\kappa(\theta)$ with respect to θ.

Mean parameter

Under (5.7) the cumulant-generating function of Y is $\kappa(\theta+t) - \kappa(\theta)$, so its mean and variance are

$$\mathrm{E}(Y) = \frac{d\kappa(\theta)}{d\theta} = \kappa'(\theta), \quad \mathrm{var}(Y) = \frac{d^2\kappa(\theta)}{d\theta^2} = \kappa''(\theta),$$

say. As Y is non-degenerate under f_0, $\mathrm{var}(Y) > 0$ for all $\theta \in \mathcal{N}$, and hence $\kappa'(\theta)$ is a strictly monotonic increasing function of θ. Thus there is a smooth 1–1 mapping between θ and the *mean parameter* $\mu = \mu(\theta) = \kappa'(\theta)$, and as θ varies in $\mathcal{N}$, μ varies in the *expectation space* $\mathcal{M}$.

The function $\mu(\theta)$ is important for likelihood inference. A natural exponential family is called *steep* if $|\mu(\theta_i)| \to \infty$ for any sequence $\{\theta_i\}$ in $\mathrm{int}\,\mathcal{N}$ that converges

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

to a boundary point of $\mathcal{N}$. Let us define the *closed convex hull* of $\mathcal{Y}$ to be $C(\mathcal{Y})$, the smallest closed set containing

$$\{y : y = \alpha y_1 + (1-\alpha)y_2, 0 \le \alpha \le 1, y_1, y_2 \in \mathcal{Y}\}.$$

The interior of a set, $\text{int}\,\mathcal{N}$, is what remains when its boundary is subtracted from its closure.

Now $\mathcal{M} \subseteq C(\mathcal{Y})$, because every density (5.7) reweights elements of $\mathcal{Y}$. It can be shown that a regular natural exponential family is steep, and that for such a family, steepness is equivalent to $\mathcal{M} = \text{int}\, C(\mathcal{Y})$. Thus there is a duality between $\text{int}\, C(\mathcal{Y})$ and the expectation space $\mathcal{M}$, and hence between $\text{int}\, C(\mathcal{Y})$ and $\text{int}\,\mathcal{N}$: for every $\mu \in \text{int}\, C(\mathcal{Y})$ there is a unique $\theta \in \mathcal{N}$ such that $f(y;\theta)$ has mean μ. This equivalence applies widely because most natural exponential families are regular. As we shall see below, it implies that there is a unique maximum likelihood estimator of θ except for pathological samples.

Example 5.7 (Uniform density) The mean function for the natural exponential family generated by the $U(0,1)$ density, $\mu(\theta) = (1-e^{-\theta})^{-1} - \theta^{-1}$, is shown in the right panel of Figure 5.3. Here $\mathcal{Y} = (0,1)$, so $C(\mathcal{Y}) = [0,1]$ and $\text{int}\, C(\mathcal{Y}) = (0,1) = \mathcal{M}$. The family is steep because the only boundary points of $\mathcal{N} = (-\infty, \infty)$ are $\pm\infty$, to which no sequence $\{\theta_i\} \subset \mathcal{N}$ can converge.

The family with $\Theta = [0, \infty)$ is not steep, because $\mu(\theta) \to 1/2$ as $\theta \downarrow 0$. ■

Example 5.8 (Poisson density) If $\mathcal{Y} = \{0, 1, \ldots\}$ and $f_0(y) = e^{-1}/y!$, then

$$\kappa(\theta) = \log\left(\sum_{y=0}^{\infty} e^{\theta y - 1}/y!\right) = e^{\theta} - 1$$

is finite for all $\theta \in \mathcal{N} = (-\infty, \infty)$. Hence

$$f(y;\theta) = \exp(\theta y - e^{\theta})/y!, \quad y \in \mathcal{Y},\ \theta \in \mathcal{N},$$

is a regular natural exponential family. Here $C(\mathcal{Y}) = [0, \infty)$, and the mean function is $\mu(\theta) = \kappa'(\theta) = e^{\theta}$, so $\mathcal{M} = (0, \infty) = \text{int}\, C(\mathcal{Y})$; the family is steep.

In terms of μ we have the familiar expression

$$f(y;\mu) = \exp(y \log \mu - \mu)/y! = \mu^y e^{-\mu}/y!, \quad y = 0, 1, \ldots, \mu > 0.$$

■

Variance function

When Y has a natural exponential family density with cumulant-generating function $\kappa(\theta)$, its mean is $\mu(\theta) = \kappa'(\theta)$. Now $\kappa(\theta)$ is smooth and strictly convex, so the mapping between θ and $\mu = \mu(\theta) = \kappa'(\theta)$ is smooth and monotone. It follows that the density (5.7) can be reparametrized in terms of μ, setting $\theta = \theta(\mu)$. In terms of μ, $\kappa(\theta) = \kappa\{\theta(\mu)\}$, so

$$\text{var}(Y) = \kappa''(\theta) = \left.\frac{d\mu}{d\theta}\right|_{\theta=\theta(\mu)} = V(\mu), \quad \mu \in \mathcal{M},$$

say, where $V(\mu)$ is the *variance function* of the family. As we saw in Section 3.1.2, the variance function determines the variance-stabilizing transformation for Y. It plays a

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

central role in generalized linear models, which we shall study in Section 10.3. The variance function and its domain $\mathcal{M}$ together determine their exponential family, as we shall now see.

On differentiating the identity $\mu\{\theta(\mu)\} = \mu$ with respect to μ, we obtain $\mu'\{\theta(\mu)\}d\theta/d\mu = 1$, and this implies that

$$\frac{d\theta(\mu)}{d\mu} = \frac{1}{\mu'\{\theta(\mu)\}} = \frac{1}{V(\mu)}. \tag{5.10}$$

As $\text{var}(Y) > 0$, this derivative is finite for any $\mu \in \mathcal{M}$, so

$$\int_{\mu_0}^{\mu} \frac{1}{V(u)}\, du = \theta(\mu) - \theta(\mu_0),$$

and as $0 \in \mathcal{N}$ we can choose $\mu_0 \in \mathcal{M}$ to give $\theta(\mu_0) = 0$. Now

$$\kappa(\theta) = \int_0^{\theta} \kappa'(t)\, dt = \int_0^{\theta} \mu(t)\, dt = \int_{\mu_0}^{\mu} \mu \frac{dt}{d\mu}\, d\mu = \int_{\mu_0}^{\mu} \frac{u}{V(u)}\, du,$$

where we have used (5.10). Hence

$$\kappa\left\{\int_{\mu_0}^{\mu} \frac{1}{V(u)}\, du\right\} = \int_{\mu_0}^{\mu} \frac{u}{V(u)}\, du, \tag{5.11}$$

and given $\mathcal{M}$ and $V(\mu)$, we have expressed κ in terms of μ; this determines $\kappa(\theta)$ implicitly. The natural parameter space $\mathcal{N}$ is traced out by $\theta(\mu) = \int_{\mu_0}^{\mu} V(u)^{-1}\, du$ as μ varies in $\mathcal{M}$.

Example 5.9 (Linear variance function) Let Y be a random variable with $V(\mu) = \mu$ and $\mathcal{M} = (0, \infty)$. Then

$$\int_{\mu_0}^{\mu} \frac{1}{V(u)}\, du = \int_{\mu_0}^{\mu} \frac{du}{u} = \log(\mu/\mu_0), \qquad \int_{\mu_0}^{\mu} \frac{u}{V(u)}\, du = \mu - \mu_0,$$

and if $\mu_0 = 1$, (5.11) gives $\kappa(\log \mu) = \mu - 1$. On setting $\theta = \log \mu$, we have $\kappa(\theta) = e^{\theta} - 1$, and as μ varies in $\mathcal{M}, \theta = \log \mu$ varies in $(-\infty, \infty)$. As $e^{\theta} - 1$ is the cumulant-generating function of the Poisson density with mean e^{θ} and there is a 1–1 correspondence between cumulant-generating functions and distributions, Y is Poisson with mean $\mu = e^{\theta}$. ■

5.2.2 Families of order p

To generalize the preceding discussion to models with several parameters, we again start from a base density $f_0(y)$, now supposing that its support $\mathcal{Y} \subseteq \mathbb{R}^d$, for $d \geq 1$, is not a subset of any space of dimension lower than d. Let the $p \times 1$ vector $s(y) = (s_1(y), \ldots, s_p(y))^{\mathrm{T}}$ consist of functions of y for which the set $\{1, s_1(y), \ldots, s_p(y)\}$ is linearly independent, and define

$$\mathcal{N} = \left\{\theta \in \mathbb{R}^p : \kappa(\theta) = \log \int e^{s(y)^{\mathrm{T}}\theta} f_0(y)\, dy < \infty\right\},$$

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

where $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$. In general $\theta = \theta(\omega)$ may depend on a parameter ω taking values in $\Omega \subset \mathbb{R}^q$, where $\theta(\Omega) \subseteq \mathcal{N}$.

An *exponential family of order p* has density

$$f(y;\omega) = \exp\{s(y)^{\mathrm{T}}\theta(\omega) - b(\omega)\} f_0(y), \quad y \in \mathcal{Y}, \omega \in \Omega, \tag{5.12}$$

where $b(\omega) = \kappa\{\theta(\omega)\}$. This is called a *minimal representation* if the set $\{1, \theta_1(\omega), \ldots, \theta_p(\omega)\}$ is linearly independent. If there is a 1–1 mapping between $\mathcal{N}$ and Ω the family can be written as a *natural exponential family of order p*,

$$f(y;\omega) = \exp\{s(y)^{\mathrm{T}}\theta - \kappa(\theta)\} f_0(y), \quad y \in \mathcal{Y}, \theta \in \mathcal{N}. \tag{5.13}$$

Terms such as natural observation, natural parameter space, expectation space, regular model, and steep family generalize to families of order p and we shall use them below without further comment. Our proofs that the natural parameter space $\mathcal{N}$ is convex, that the family may be generated by any of its members, that $\kappa(\theta)$ is strictly convex, and that $s(Y)$ has cumulant-generating function $\kappa(\theta + t) - \kappa(\theta)$ also generalize with minor changes. The mean vector and covariance matrix of $s(Y)$ are now the $p \times 1$ vector and $p \times p$ matrix

$$\mathrm{E}\{s(Y)\} = \frac{d\kappa(\theta)}{d\theta}, \quad \mathrm{var}\{s(Y)\} = \frac{d^2\kappa(\theta)}{d\theta d\theta^{\mathrm{T}}}.$$

Example 5.10 (Beta density) If $f_0(y)$ is uniform on $(0, 1)$ and $s(y)$ equals $(\log y, \log(1 - y))^{\mathrm{T}}$, then

$$\kappa(\theta) = \log \int_0^1 \exp\{\theta_1 \log y + \theta_2 \log(1 - y)\}\, dy = \log B(1 + \theta_1, 1 + \theta_2),$$

where $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ is the beta function; see Example 5.4. The resulting model is usually written in terms of $a = \theta_1 + 1$ and $b = \theta_2 + 1$, giving the beta density

$$f(y; a, b) = \frac{y^{a-1}(1 - y)^{b-1}}{B(a, b)}, \quad 0 < y < 1, \quad a, b > 0. \tag{5.14}$$

In this parametrization the natural parameter space is $\mathcal{N} = (0, \infty) \times (0, \infty)$. In Example 5.4 we took $s(y) = \log\{y/(1 - y)\}$, thereby generating the one-parameter subfamily in which $b = 2 - a$. This subfamily is also obtained by taking $s(y) = (\log y, \log(1 - y))^{\mathrm{T}}$ and $\theta(\omega) = (\omega, -\omega)^{\mathrm{T}}$, but this representation is not minimal because $(1, 1)\theta(\omega) = 0$.

Comparison of Figures 5.4 and 5.3 shows how tilting with two parameters broadens the variety of densities the family contains. ■

Example 5.11 (von Mises density) Directional data are those where the observations y_j are angles — see Table 5.2, which gives the bearings of 29 homing pigeons 30, 60, and 90 seconds after release and on vanishing from sight. Another example is a wind direction, while the position of a star in the sky is an instance of directional data on a sphere.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Table 5.2 Homing pigeon data (Artes, 1997). Bearings (degrees) of 29 homing pigeons 30, 60 and 90 seconds after release, with their bearings on vanishing from sight.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
30	240	300	225	285	210	265	310	330	325	290	15	330	100	35	340
60	250	290	210	325	205	240	330	315	285	335	10	305	95	65	345
90	270	305	215	295	195	210	335	315	135	10	5	325	90	70	330
van	275	285	185	290	195	225	335	285	120	30	10	85	90	80	350

	16	17	18	19	20	21	22	23	24	25	26	27	28	29
30	320	340	355	40	225	50	200	330	325	330	280	180	50	20
60	325	335	25	330	220	50	195	320	315	290	285	155	25	0
90	15	320	30	335	215	55	185	325	345	285	280	160	15	25
van	60	345	35	65	250	60	175	325	330	280	350	185	20	30

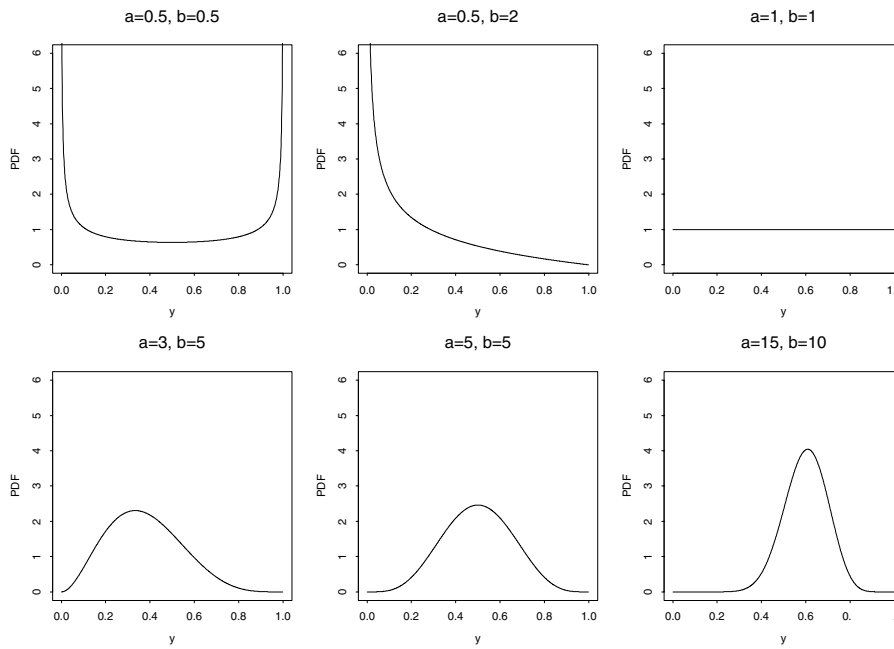

Figure 5.4 Beta densities for different values of a and b. Swapping a and b reflects the densities about $y = 0.5$.

To build a class of densities for circular data we start from the uniform density on the circle, $f_0(y) = (2\pi)^{-1}$ for $0 \le y < 2\pi$, and take

$$s(y) = (\cos y, \sin y)^{\mathrm{T}}, \quad \theta(\omega) = (\tau \cos \gamma, \tau \sin \gamma)^{\mathrm{T}},$$

where $\omega = (\tau, \gamma)$ lies in $\Omega = [0, \infty) \times [0, 2\pi)$. This choice of $s(y)$ ensures the desirable property $f(y) = f(y \pm 2k\pi)$ for all integer k. Now $s(y)^{\mathrm{T}}\theta(\omega) = \tau \cos(y - \gamma)$ and

$$\int e^{s(y)^{\mathrm{T}}\theta(\omega)} f_0(y)\, dy = \frac{1}{2\pi} \int_0^{2\pi} e^{\tau \cos(y-\gamma)}\, dy = \frac{1}{2\pi} \int_0^{2\pi} e^{\tau \cos y}\, dy = I_0(\tau),$$

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

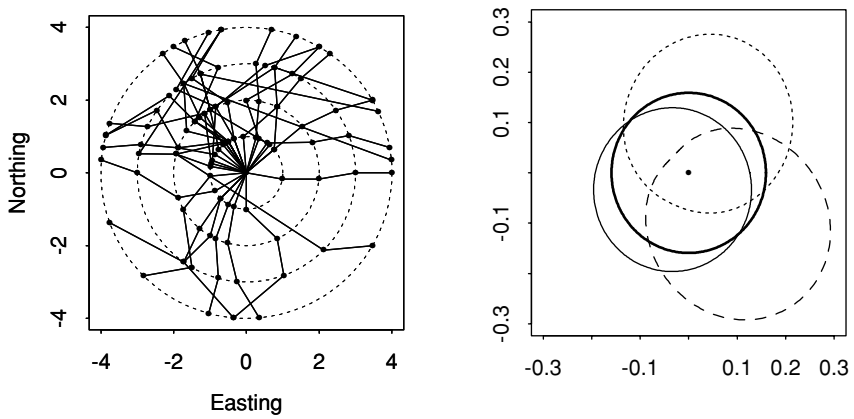

Figure 5.5 Circular data. Left: bearings of 29 homing pigeons at various intervals after release. Right: von Mises densities for different values of γ and τ. Shown are the baseline uniform density (heavy) $(2\pi)^{-1}$, and von Mises densities with $\tau = 0.3$, $\gamma = 5\pi/4$ (solid), $\tau = 0.7$, $\gamma = 3\pi/8$ (dots), and $\tau = 1$, $\gamma = 7\pi/4$ (dashes). In each case the density $f(y; \tau, \gamma)$ is given by the distance from the origin to the curve, so the areas do not integrate to one.

where $I_\nu(\tau)$ is the modified Bessel function of the first kind and order ν. The resulting exponential family is the von Mises density

$$f(y; \tau, \gamma) = \{2\pi I_0(\tau)\}^{-1} e^{\tau \cos(y-\gamma)}, \quad 0 \le y < 2\pi,\ \tau > 0, 0 \le \gamma < 2\pi;$$

see Figure 5.5. The *mean direction* γ gives the direction in which observations are concentrated, and the *precision* τ gives the strength of that concentration. Notice that $\tau = 0$ gives the uniform distribution on the circle, whatever the value of γ. Here interest focuses on Y rather than on $s(Y)$, which is introduced purely in order to generate a natural class of densities for y.

The estimates and standard errors for the data in Table 5.2 are $\widehat{\gamma} = 320$ (15) and $\widehat{\tau} = 1.08$ (0.32) at 30 seconds, with corresponding figures 316 (15) and 1.05 (0.32) at 60 seconds, 329 (21) and 0.75 (0.29) at 90 seconds, and 357 (29) and 0.52 (0.28) on vanishing. Thus as Figure 5.5 shows, the bearings of the pigeons become more dispersed as they fly away. The likelihood ratio statistics that compare the fitted two-parameter model with the uniform density are 13.80, 13.34, 7.33, and 3.75. As the mean direction γ vanishes under the uniform model, the situation is non-regular (Section 4.6), but the evidence against uniformity clearly weakens as time passes. ■

Richard von Mises (1883–1953) was born in Lvov and educated in Vienna and Brno. He became professor of applied mathematics in Strasbourg, Dresden and Berlin, then left for Istanbul to escape the Nazis, finishing his career at Harvard. A man of wide interests, he spent the 1914–18 war as a pilot in the Austro-Hungarian army, gave the first university course on powered flight, and made contributions to aeronautics, aerodynamics and fluid dynamics as well as philosophy, probability and statistics; he was also an authority on the Austrian poet Rainer Maria Rilke. He is now perhaps best known for his frequency theory basis for probability.

Curved exponential families

In the examples above, the natural parameter $\theta = (\theta_1(\omega), \ldots, \theta_p(\omega))^{\mathrm{T}}$ is a 1–1 function of $\omega = (\omega_1, \ldots, \omega_q)^{\mathrm{T}}$, so of course $p = q$. Another possibility is that $q > p$, in which case ω cannot be identified from data. Such models are not useful in practice, and it is more interesting to consider the case $q < p$. Now $\theta(\omega)$ varies in the q-dimensional subspace $\theta(\Omega)$ of $\mathcal{N}$. If $\theta = a + B\omega$ is a linear function of ω, where a and B are a $p \times 1$ vector and a $p \times q$ matrix of constants, then $s(y)^{\mathrm{T}}\theta(\omega) = s(y)^{\mathrm{T}}a + \{s(y)^{\mathrm{T}}B\}\omega$, and the exponential family may be generated from $f_0'(y) \propto e^{a^{\mathrm{T}}s(y)} f_0(y)$ by taking $s'(y) = B^{\mathrm{T}}s(y)$. Hence it is just an exponential family of order q and no new issues arise: the original representation was not minimal. If $\theta(\omega)$ is a nonlinear function, however, and the representation is minimal, we have a (p, q) *curved exponential family*.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Example 5.12 (Multinomial density) The multinomial density with denominator m and probability vector $\pi = (\pi_1, \ldots, \pi_p)^{\mathrm{T}}$ is

$$\begin{aligned}\frac{m!}{y_1! \cdots y_p!} \pi_1^{y_1} \cdots \pi_p^{y_p} &\propto \exp\{y_1 \log \pi_1 + \cdots + y_p \log \pi_p\} \\ &= \exp\{y_1 \log \pi_1 + \cdots + y_{p-1} \log \pi_{p-1} \\ &\quad + (m - y_1 - \cdots - y_{p-1}) \log(1 - \pi_1 - \cdots - \pi_{p-1})\} \\ &= \exp\{y_1\theta_1 + \cdots + y_{p-1}\theta_{p-1} - \kappa(\theta)\},\end{aligned}$$

where

$$\pi_r = \frac{e^{\theta_r}}{1 + e^{\theta_1} + \cdots + e^{\theta_{p-1}}}, \quad \kappa(\theta) = m \log(1 + e^{\theta_1} + \cdots + e^{\theta_{p-1}}).$$

This is a minimal representation of a natural exponential family of order $p - 1$ with $s(y) = (y_1, \ldots, y_{p-1})^{\mathrm{T}}$, $\mathcal{N} = (-\infty, \infty)^{p-1}$ and

$$f_0(y) = \frac{p^{-m} m!}{y_1! \cdots y_p!}, \quad \mathcal{Y} = \left\{(y_1, \ldots, y_p) : y_1, \ldots, y_p \in \{0, \ldots, m\}, \sum y_r = m\right\};$$

$\mathcal{Y}$ is a subset of the scaled p-dimensional simplex

$$C(\mathcal{Y}) = \left\{(y_1, \ldots, y_p) : 0 \le y_1, \ldots, y_p \le m, \sum y_r = m\right\}.$$

Now

$$\mathrm{E}\{s(Y)\} = \frac{m}{1 + e^{\theta_1} + \cdots + e^{\theta_{p-1}}}(e^{\theta_1}, \ldots, e^{\theta_{p-1}}),$$

and as $\mathrm{E}(Y_p) = m - \mathrm{E}(Y_1) - \cdots - \mathrm{E}(Y_{p-1})$, the expectation space in which $\mu(\theta) = \mathrm{E}(Y)$ varies equals $\operatorname{int} C(\mathcal{Y})$: the model is steep.

Many multinomial models are curved exponential families. In Example 4.38, for instance, the ABO blood group data had $p = 4$ groups with

$$\pi_A = \lambda_A^2 + 2\lambda_A\lambda_O, \quad \pi_B = \lambda_B^2 + 2\lambda_B\lambda_O, \quad \pi_O = \lambda_O^2, \quad \pi_{AB} = 2\lambda_A\lambda_B, \quad (5.15)$$

where $\lambda_A + \lambda_B + \lambda_O = 1$. This is a $(3, 2)$ curved exponential family. In the full family of order p, the probabilities π_A, π_B and π_{AB} vary in the set

$$\mathcal{A} = \{(\pi_A, \pi_B, \pi_{AB}) : 0 \le \pi_A, \pi_B, \pi_{AB} \le 1, 0 \le \pi_A + \pi_b + \pi_{AB} \le 1\},$$

shown in Figure 5.6. In the sub-family given by (5.15), when λ_O is fixed we have $\lambda_A + \lambda_B = 1 - \lambda_O$, and as λ_A varies from 0 to $1 - \lambda_O$, (π_A, π_B, π_{AB}) traces a curve from $(0, 1 - \lambda_O^2, 0)$ to $(1 - \lambda_O^2, 0, 0)$ shown in the figure. As λ_O varies from 0 to 1,

$$\begin{aligned}(\pi_A, \pi_B, \pi_{AB}) = \big(&\lambda_A^2 + 2p\lambda_O, (1 - \lambda_A - \lambda_O)^2 + 2(1 - \lambda_A - \lambda_O)\lambda_O, \\ &2\lambda_A(1 - \lambda_A - \lambda_O)\big)\end{aligned}$$

traces out the intersection of a cone with the set $\mathcal{A}$. Thus although any value of (π_A, π_B, π_{AB}) inside the tetrahedron with corners $(0, 0, 0)$, $(0, 0, 1)$, $(0, 1, 0)$ and $(1, 0, 0)$ is possible under the full model, the curved submodel restricts the probabilities to the hatched surface. ■

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

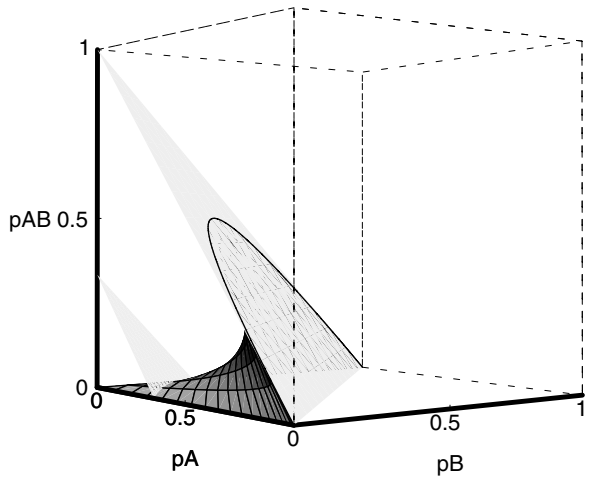

Figure 5.6 Parameter space for four-category multinomial model. The full parameter space for (π_A, π_B, π_{AB}) is the tetrahedron with corners (0, 0, 0), (0, 0, 1), (0, 1, 0) and (1, 0, 0), whose outer face is shaded. The other parameter $\pi_O = 1 - \pi_A - \pi_B - \pi_{AB}$. The two-parameter sub-model given by (5.15) is shown by the hatched surface.

5.2.3 Inference

Let $Y_1, \ldots, Y_n$ be a random sample from an exponential family of order p. Their joint density is

$$\prod_{j=1}^{n} f(y_j;\omega) = \exp\left\{\sum_{j=1}^{n} s(y_j)^{\mathrm{T}}\theta(\omega) - nb(\omega)\right\}\prod_{j=1}^{n} f_0(y_j), \quad \omega \in \Omega, \tag{5.16}$$

and consequently the density of $S = \sum s(Y_j)$ is

$$\begin{aligned} f(s;\omega) = \int \prod_{j=1}^{n} f(y_j;\omega)\,dy &= \exp\{s^{\mathrm{T}}\theta(\omega) - nb(\omega)\}\int \prod_{j=1}^{n} f_0(y_j)\,dy \\ &= \exp\{s^{\mathrm{T}}\theta(\omega) - nb(\omega)\}g_0(s), \end{aligned}$$

say, where the integral is over

$$\left\{(y_1, \ldots, y_n) : y_1, \ldots, y_n \in \mathcal{Y}, \sum_{j=1}^{n} s(y_j) = s\right\}.$$

Hence S too has an exponential family density of order p. That is, the sum of n independent variables from an exponential family belongs to the same family, with cumulant-generating function $n\kappa(\theta) = nb(\omega)$. The factorization criterion (4.15) applied to (5.16) implies that S is a sufficient statistic for ω based on $Y_1, \ldots, Y_n$, and if $f(y;\omega)$ is a minimal representation, S is minimal sufficient (Exercise 5.2.12). Thus inference for ω may be based on the density of S, while the joint density of $Y_1, \ldots, Y_n$ given the value of S is independent of ω:

$$f(y_1, \ldots, y_n;\omega) = f(y_1, \ldots, y_n \mid s) f(s;\omega). \tag{5.17}$$

This decomposition allows us to split the inference into two parts, corresponding to the factors on its right, the first of which may be used to assess model adequacy. If satisfied of an adequate fit, we use the second term for inference on ω. We now discuss these aspects in turn.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Model adequacy

The argument for using the first factor on the right of (5.17) to assess model adequacy is that the value of ω is irrelevant to deciding if $f(y;\omega)$ fits the random sample $Y_1, \ldots, Y_n$. Hence we should assess fit using the conditional distribution of Y given S; see Example 4.10.

Example 5.13 (Poisson density) If $Y_1, \ldots, Y_n$ is a random sample from a Poisson density with mean μ, their common cumulant-generating function is $\mu(e^t - 1)$ and the natural observation is $s(y_j) = y_j$. Hence $S = \sum s(Y_j) = \sum Y_j$ has cumulant-generating function $n\mu(e^t - 1)$. The joint conditional density of $y_1, \ldots, y_n$ given that $S = s$,

$$\begin{aligned} f(y_1, \ldots, y_n \mid s) &= \frac{f(y_1, \ldots, y_n; \theta)}{f(s;\theta)} \\ &= \frac{\prod_{j=1}^n \mu^{y_j} e^{-\mu}/y_j!}{(n\mu)^s e^{-n\mu}/s!} \\ &= \begin{cases} \frac{s!}{y_1! \cdots y_n!} n^{-s}, & y_1 + \cdots + y_n = s, \\ 0, & \text{otherwise}, \end{cases} \end{aligned}$$

is multinomial with denominator s and $n \times 1$ probability vector $(n^{-1}, \ldots, n^{-1})$. This density is independent of μ by its construction.

The mean and variance of a Poisson variable both equal μ, so Poissonness of a random sample of counts can be assessed by comparing their average $\overline{Y}$ and sample variance $(n-1)^{-1}\sum(Y_j - \overline{Y})^2$. A common problem with such data is overdispersion, which is suggested if $P = \sum(Y_j - \overline{Y})^2/\overline{Y}$ greatly exceeds $n-1$. How big is 'greatly'? As $\widehat{\mu} = \overline{Y}$ is the maximum likelihood estimate of μ, P is Pearson's statistic (Section 4.5.3) and has an asymptotic χ^2_{n-1} distribution. The argument above suggests that we assess if P is large compared to its conditional distribution given the value of $S = \sum Y_j = n\overline{Y}$, so the distribution we seek is that of P conditional on $\overline{Y}$. The conditional mean and variance of P are $(n-1)$ and $2(n-1)(1-s^{-1}) \doteq 2(n-1)$, and the conditional distribution of P is very close to χ^2_{n-1} unless s and n are both very small. Hence the *Poisson dispersion test* compares P to the χ^2_{n-1} distribution, with large values suggesting that the counts are more variable than Poisson data would be.

In Table 2.1, for example, the daily numbers of arrivals are 16, 16, 13, 11, 14, 13, 12, so P takes value 1.6, to be treated as χ^2_6, so the counts seem under- rather than overdispersed. In Example 4.40, by constrast, with counts 1, 5, 3, 2, 2, 1, 0, 0, 2, 1, 1, 7, 11, 4, 7, 10, 16, 16, 9, 15, we have $P = 99.92$, which is very large compared to the χ^2_{19} distribution; and in fact $\Pr(P \geq 99.92) \doteq 0$ to 12 decimal places. As one might expect, these data are highly overdispersed relative to the Poisson model.

Another possibility is that although all Poisson, the Y_j have different means. In Example 4.40 we compared the changepoint model under which $Y_1, \ldots, Y_\tau$ and $Y_{\tau+1}, \ldots, Y_n$ have different means with the model of equal means. The comparison involved the likelihood ratio statistic, whose exact conditional distribution was simulated under the simpler model; see Figure 4.9. ■

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Example 5.14 (Normal model) The normal density may be written

$$f(y;\mu,\sigma^2) = \frac{1}{(2\pi)^{1/2}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$
$$= \exp\left\{\frac{\mu}{\sigma^2}y - \frac{1}{2\sigma^2}y^2 - \frac{\mu^2}{2\sigma^2} - \log\sigma - \frac{1}{2}\log(2\pi)\right\}. \quad (5.18)$$

This is a minimal representation of an exponential family of order 2 with

$$\omega = (\mu,\sigma^2) \in \Omega = (-\infty,\infty)\times(0,\infty),$$
$$\theta(\omega)^{\mathrm{T}} = (\mu/\sigma^2, 1/(2\sigma^2)) \in \mathcal{N} = (-\infty,\infty)\times(0,\infty),$$
$$s(y)^{\mathrm{T}} = (y,-y^2),$$
$$\kappa(\theta) = \theta_1^2/(4\theta_2) - \frac{1}{2}\log(2\theta_2),$$

arising from tilting the standard normal density $(2\pi)^{-1/2}e^{-y^2/2}$.

We now consider how decomposition (5.17) applies for the normal model with $n > 2$. When $Y_1,\ldots,Y_n$ is a random sample from (5.18), our general discussion implies that $(\sum Y_j, -\sum Y_j^2)$ is minimal sufficient. As this is in 1–1 correspondence with $\overline{Y}$, $S^2 = (n-1)^{-1}\sum(Y_j-\overline{Y})^2$, our old friends the average and sample variance are also minimal sufficient. When $n > 1$ the joint distribution of $\overline{Y}$ and S^2 is nondegenerate with probability one, and (3.15) states that they are independently distributed as $N(\mu,\sigma^2/n)$ and $(n-1)^{-1}\sigma^2\chi^2_{n-1}$.

In order to compute the conditional density of $Y_1,\ldots,Y_n$ given $\overline{Y}$ and S, it is neatest to set $E_j = (Y_j-\overline{Y})/S$ and consider the conditional density of $E_1,\ldots,E_n$. As $\sum E_j = 0$ and $\sum E_j^2 = n-1$, the random vector $(E_1,\ldots,E_n) \in \mathbb{R}^n$ lies on the intersection of the hypersphere of radius $n-1$ and the hyperplane $\sum E_j = 0$. As this is a $(n-2)$-dimensional subset of $\mathbb{R}^n$, the joint density of $E_1,\ldots,E_n$ is degenerate but that of $E_3,\ldots,E_n$ is not.

To find the joint density of $T_3 = E_3,\ldots,T_n = E_n$ given $T_1 = \overline{Y}$ and $T_2 = S$, we need the Jacobian of the transformation from $y_1,\ldots,y_n$ to $t_1,\ldots,t_n$. In order to obtain this Jacobian, we first note that $y_j = t_1 + t_2t_j$, for $j = 3,\ldots,n$. As $\sum e_j = 0$ and $\sum e_j^2 = n-1$, we can write

$$e_1 + e_2 = -\sum_{j=3}^n t_j, \quad n-1-e_1^2-e_2^2 = \sum_{j=3}^n t_j^2,$$

implying that there are functions h_1 and h_2 such that

$$e_1 = h_1(t_3,\ldots,t_n), \quad e_2 = h_2(t_3,\ldots,t_n),$$

which in turn gives

$$y_1 = t_1 + t_2h_1(t_3,\ldots,t_n), \quad y_2 = t_1 + t_2h_2(t_3,\ldots,t_n).$$

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Let $h_{ij} = \partial h_i(t_3, \ldots, t_n)/\partial t_j$. The Jacobian we seek is

$$\left|\frac{\partial(y_1, \ldots, y_n)}{\partial(t_1, \ldots, t_n)}\right| = \begin{vmatrix} 1 & h_1 & t_2 h_{13} & t_2 h_{14} & \cdots & t_2 h_{1n} \\ 1 & h_2 & t_2 h_{23} & t_2 h_{24} & \cdots & t_2 h_{2n} \\ 1 & t_3 & t_2 & 0 & \cdots & 0 \\ 1 & t_4 & 0 & t_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & 0 & 0 & \cdots & t_2 \end{vmatrix} = t_2^{n-2} h'(t_3, \ldots, t_n)$$

$$= s^{n-2} H(e), \qquad (5.19)$$

say. Hence

$$f(e_3, \ldots, e_n \mid \overline{y}, s) = \frac{f(y_1, \ldots, y_n; \mu, \sigma^2) s^{n-2} H(e)}{f(\overline{y}; \mu, \sigma^2) f(s; \sigma^2)} \propto H(e)$$

after a straightforward calculation. As this depends on $e_1, \ldots, e_n$ alone, the corresponding random variables $E_1, \ldots, E_n$ are independent of $\overline{Y}$ and S^2.

Thus assessment of fit of the normal model should be based on the *raw residuals* $e_1, \ldots, e_n$. One simple tool is a normal probability plot of the e_j, which should be a straight line of unit gradient through the origin. Such plots and variants are common in regression (Section 8.6.1). Further support for use of the e_j for model checking is given in Section 5.3. ■

Likelihood

Let $Y_1, \ldots, Y_n$ be a random sample from an exponential family of order p. Inference for the parameter may be based on the sufficient statistic $\overline{S} = n^{-1} \sum s(Y_j)$, which also belongs to a natural exponential family of order p, with support $\mathcal{S}$, say. Hence the log likelihood may be written

$$\ell(\omega) \equiv n \{\overline{S}^{\mathrm{T}} \theta(\omega) - b(\omega)\} = n[\overline{S}^{\mathrm{T}} \theta(\omega) - \kappa \{\theta(\omega)\}], \quad \omega \in \Omega,$$

and the score vector and observed information matrix are given by

$$U(\omega) = \frac{\partial \ell(\omega)}{\partial \omega} = \frac{\partial \theta^{\mathrm{T}}}{\partial \omega} n \left\{\overline{S} - \frac{\partial \kappa(\theta)}{\partial \theta}\right\},$$

$$J(\omega)_{rs} = -\frac{\partial^2 \ell(\omega)}{\partial \omega_r \partial \omega_s} = -\frac{\partial^2 \theta^{\mathrm{T}}}{\partial \omega_r \partial \omega_s} n \left\{\overline{S} - \frac{\partial \kappa(\theta)}{\partial \theta}\right\} + \frac{\partial \theta^{\mathrm{T}}}{\partial \omega_r} \left\{n \frac{\partial^2 \kappa(\theta)}{\partial \theta \partial \theta^{\mathrm{T}}}\right\} \frac{\partial \theta}{\partial \omega_s}.$$

The observed information is random unless the family is in natural form, in which case $\theta = \omega$ and hence $\partial^2 \theta / \partial \omega_r \partial \omega_s = 0$; then $I(\theta) = \mathrm{E}\{J(\theta)\} = J(\theta)$.

If the family is steep, there is a 1–1 relation between the interior of the closure of $\mathcal{S}$, $\mathrm{int}\, C(\mathcal{S})$, the expectation space $\mathcal{M}$ of $\overline{S}$, and the natural parameter space $\mathcal{N} = \theta(\Omega)$. Thus if $\overline{S} \in \mathrm{int}\, C(\mathcal{S})$, there is a single value of θ such that $\overline{S} = \mu(\theta)$ and $u(\theta) = 0$, and moreover there is a 1–1 map between $\widehat{\theta}$ and $\widehat{\omega}$. Hence the maximum likelihood estimators satisfy

$$\widehat{\mu} = \mu(\widehat{\theta}) = \mu\{\theta(\widehat{\omega})\} = \overline{S}.$$

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Thus the likelihood equation has just one solution, which maximizes the log likelihood. Moreover, as Ω is open and $\widehat{\omega} \in \Omega$, standard likelihood asymptotics will apply, so $\widehat{\omega} \mathrel{\dot\sim} N\{\omega, I(\omega)^{-1}\}$ and $2\{\ell(\widehat{\omega}) - \ell(\omega)\} \mathrel{\dot\sim} \chi^2_p$. If the model permits $\overline{S} \notin \mathcal{M}$, standard asymptotics will break down. The same difficulty could arise if the true parameter lies on the boundary of the parameter space.

Example 5.15 (Uniform density) The average $\overline{y}$ of a random sample from (5.8) must lie in the interval $(0, 1)$. Given $\overline{y}$, the maximum likelihood estimate $\widehat{\theta}$ is read off from the right panel of Figure 5.3 as the value of θ on the horizontal axis for which $\mu(\theta) = \overline{y}$ on the vertical axis.

As mentioned in Example 5.7, when θ is restricted to $\Theta = [0, \infty)$ the family is not steep, because $\mathcal{M} = [1/2, 1) \neq (0, 1) = \operatorname{int} C(\mathcal{Y})$. A value $\overline{y} < 1/2$ is possible for any sample size and any $\theta \in \Theta$, and as $\widehat{\theta} = 0$ is the maximum likelihood estimate for any such $\overline{y}$, the 1–1 mapping between $\overline{y}$ and $\widehat{\theta}$ is destroyed. Furthermore, this Θ is not open, so the limiting distribution of $\widehat{\theta}$ and the likelihood ratio statistic are non-standard if $\theta = 0$; see Example 4.39. ■

Example 5.16 (Binomial density) The binomial model with denominator m, probability $0 < \pi < 1$ and natural parameter $\theta = \log\{\pi/(1-\pi)\} \in (-\infty, \infty)$ has $\mathcal{Y} = \{0, 1, \ldots, m\}$ and $\operatorname{int} C(\mathcal{Y}) = \mathcal{M} = (0, m)$. The average $\overline{R}$ of a random sample $R_1, \ldots, R_n$ lies outside $(0, m)$ with probability

$$\Pr(R_1 = \cdots = R_n = 0) + \Pr(R_1 = \cdots = R_n = m) = (1-\pi)^{mn} + \pi^{mn} > 0,$$

so the maximum likelihood estimator $\widehat{\theta} = \log\left\{\overline{R}/(m - \overline{R})\right\}$ may not be finite. As the family is steep, a unique value of θ corresponds to each $\overline{R} \in \mathcal{M}$, so the only problem that can arise is that $\widehat{\theta} = \pm\infty$ with small probability. On the other hand $\Pr(|\widehat{\theta}| = \infty) \to 0$ exponentially fast as $n \to \infty$, so infinite $\widehat{\theta}$ is rare in practice, though not unknown. It corresponds to $\widehat{\pi} = 0$ or $\widehat{\pi} = 1$.

This difficulty also arises with other discrete exponential families. ■

Example 5.17 (Normal density) Example 4.18 gives the score and information quantities for a sample from the normal model in terms of μ and σ^2; in this parametrization the observed information is random. In Example 4.22 we saw that the log likelihood $\ell(\mu, \sigma^2)$ is unimodal and that the maximum likelihood estimators are the sole solution to the likelihood equation; this is an instance of the general result above. ■

Derived densities

Various models derived from exponential families are themselves exponential families, and this can be useful in inference.

Consider a natural exponential family of order p with S^{T} and θ^{T} partitioned as $(S_1^{\mathrm{T}}, S_2^{\mathrm{T}})$ and $(\psi^{\mathrm{T}}, \lambda^{\mathrm{T}})$, where S_1 and ψ have dimension $q < p$. The marginal density

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

of S_2, obtained by integration over the values of S_1, is

$$\begin{aligned} f(s_2;\theta) &= \int \exp\left\{s_1^{\mathrm{T}}\psi + s_2^{\mathrm{T}}\lambda - \kappa(\theta)\right\} g_0(s_1,s_2)\,ds_1 \\ &= \exp\left\{s_2^{\mathrm{T}}\lambda - \kappa(\theta)\right\} \int \exp\left(s_1^{\mathrm{T}}\psi\right) g_0(s_1,s_2)\,ds_1 \\ &= \exp\left\{s_2^{\mathrm{T}}\lambda - \kappa(\theta) + d_\psi(s_2)\right\}, \end{aligned}$$

say, so for fixed ψ the marginal density of S_2 is an exponential family with natural parameter λ.

The conditional density of S_1 given $S_2 = s_2$ is

$$\begin{aligned} f_{S_1|S_2}(s_1 \mid s_2;\theta) &= \frac{\exp\left\{s_1^{\mathrm{T}}\psi + s_2^{\mathrm{T}}\lambda - \kappa(\theta)\right\} g_0(s_1,s_2)}{\exp\left\{s_2^{\mathrm{T}}\lambda - \kappa(\theta) + d_\psi(s_2)\right\}} \\ &= \exp\left\{s_1^{\mathrm{T}}\psi - \kappa_{s_2}(\psi)\right\} g_{s_2}(s_1), \end{aligned}$$

say. This is an exponential family of order q with natural parameter ψ, but the base density and cumulant-generating function depend on s_2. Such a removal of λ by conditioning is a powerful way to deal with nuisance parameters.

Example 5.18 (Gamma density) Independent gamma variables $Y_1, \ldots, Y_n$ with scale parameter λ and shape parameters $\kappa_1, \ldots, \kappa_n$ have joint density

$$\prod_{j=1}^{n} \frac{\lambda^{\kappa_j} y_j^{\kappa_j - 1}}{\Gamma(\kappa_j)} \exp(-\lambda y_j) = \lambda^{\sum \kappa_j} \exp\left(-\lambda \sum_{j=1}^{n} y_j\right) \prod_{j=1}^{n} \frac{y_j^{\kappa_j - 1}}{\Gamma(\kappa_j)}.$$

As Y_j has cumulant-generating function $-\kappa_j \log(1-\lambda t)$, $S_1 = S = \sum Y_j$ is gamma with parameters λ and $\sum \kappa_j$. The conditional density of $Y_1, \ldots, Y_n$ given $S = s$ is

$$\frac{\Gamma\left(\sum \kappa_j\right)}{\prod_{j=1}^{n} \Gamma(\kappa_j)} s^{-n} \prod_{j=1}^{n} \left(\frac{y_j}{s}\right)^{\kappa_j - 1}, \quad y_j > 0, \sum_{j=1}^{n} y_j = s.$$

Thus the joint density of $U_1 = Y_1/S, \ldots, U_n = Y_n/S$,

$$f(u_1, \ldots, u_n; \kappa_1, \ldots, \kappa_n) = \frac{\Gamma\left(\sum \kappa_j\right)}{\prod_{j=1}^{n} \Gamma(\kappa_j)} \prod_{j=1}^{n} u_j^{\kappa_j - 1}, \quad u_j > 0, \sum_{j=1}^{n} u_j = 1, \quad (5.20)$$

lies on the simplex in n dimensions; it is called the *Dirichlet density*. Hence we may base inferences for $\kappa_1, \ldots, \kappa_n$ on the conditional density of $Y_1, \ldots, Y_n$ given their sum, or equivalently on the observed values of the U_j. ■

The discussion above suggests that we may write

$$f(s;\theta) = f_{S_1|S_2}(s_1 \mid s_2;\psi) f_{S_2}(s_2;\psi,\lambda). \quad (5.21)$$

If the model can be reparametrized in terms of a $(p-q)\times 1$ vector $\rho = \rho(\psi,\lambda)$ which is variation independent of ψ, in such a way that the second term on the right

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

of (5.21) depends only on ρ, then S_2 is said to be a *cut*. The log likelihood based on (5.21) then has form $\ell_1(\psi) + \ell_2(\rho)$, maximum likelihood estimates of ρ and ψ do not depend on each other, and the observed information matrix is block diagonal. Inferences on ψ and ρ may be made separately, using the conditional density of S_1 given S_2 and the marginal density of S_2. The cut most commonly encountered in practice arises with Poisson variables; see Example 7.34 and page 501.

Exercises 5.2

1 Here is a version of Hölder's inequality: let $f(x)$ be a density supported in $[a, b]$, let $p > 1$, and let $g(y)$ and $h(y)$ be any two real functions such that the integrals

$$\int_a^b |g(y)|^p f(y)\,dy, \quad \int_a^b |h(y)|^q f(y)\,dy,$$

are finite, where $p^{-1} + q^{-1} = 1$. Then

$$\int g(y)h(y)f(y)\,dy \le \left\{\int_a^b |g(y)|^p f(y)\,dy\right\}^{1/p} \left\{\int_a^b |h(y)|^q f(y)\,dy\right\}^{1/q}.$$

If g and h are both non-zero, there is equality if and only if $c|g(y)|^p = d|h(y)|^q$ for positive constants c and d.
Show strict convexity of the cumulant-generating function $\kappa(\theta)$ of an exponential family.

2 What natural exponential families are generated by (a) $f_0(y) = e^{-y}$, $y > 0$, and (b) $f_0(y) = \frac{1}{2}e^{-|y|}$, $-\infty < y < \infty$?

3 Which of Examples 4.1–4.6 are exponential families? What about the $U(0, \theta)$ density?

4 Show that the gamma density (2.7) is an exponential family. What about the inverse gamma density, for $1/Y$ when Y is gamma?

5 Show that the inverse Gaussian density

$$f(y; \mu, \lambda) = \left(\frac{\lambda}{2\pi y^3}\right)^{1/2} \exp\{-\lambda(y-\mu)^2/(2\mu^2 y)\}, \quad y > 0, \lambda, \mu > 0,$$

is an exponential family of order 2. Give a general form for its cumulants.

6 Find the exponential families with variance functions (i) $V(\mu) = a\mu(1-\mu)$, $\mathcal{M} = (0, 1)$, (ii) $V(\mu) = a\mu^2$, $\mathcal{M} = (0, \infty)$, and (iii) $V(\mu) = a\mu^2$, $\mathcal{M} = (-\infty, 0)$.

7 For what values of a is there an exponential family with variance function $V(\mu) = a\mu$, $\mathcal{M} = (0, \infty)$?

8 Show that the $N(\mu, \mu^2)$ model is a curved exponential family and sketch how the density changes as μ varies in $(-\infty, 0) \cup (0, \infty)$. Sketch also the subset of the natural parameter space for the $N(\mu, \sigma^2)$ distribution generated by this model.

9 Find a connection between Example 4.11 and (5.20), and hence suggest methods of checking the fit of the exponential model.

10 Explain how (5.20) may be generated as an exponential family, by showing that it generalizes (5.14).

11 Use Example 5.18 to construct a simulation algorithm for Dirichlet random variables.

12 Show that $\sum s(Y_j)$ is minimal sufficient for the parameter ω of an exponential family of order p in a minimal representation.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

5.3 Group Transformation Models

Another important class of models stems from observing that many inferences should have invariance properties. If, for instance, data y are recorded in degrees Celsius, one might obtain a conclusion $s(y)$ directly from the original data, or one might transform them to degrees Fahrenheit, giving $g(y)$, say, obtain the conclusion $s\{g(y)\}$ in these terms, and then back-transform to Celsius scale, giving conclusion $g^{-1}[s\{g(y)\}]$. It is clearly essential that $g^{-1}[s\{g(y)\}] = s(y)$. The transformation from Celsius to Fahrenheit is just one of many possible invertible linear transformations that might be applied to y, however, any of which should leave the inference unchanged. More generally we might insist that inferences be invariant when any element g of a group of transformations acts on the sample space. This section explores some consequences of this requirement.

A *group* $\mathcal{G}$ is a mathematical structure having an operation $\circ$ such that:

- if $g, g' \in \mathcal{G}$, then $g \circ g' \in \mathcal{G}$;
- $\mathcal{G}$ contains an identity element e such that $e \circ g = g \circ e = g$ for each $g \in \mathcal{G}$; and
- each $g \in \mathcal{G}$ possesses an inverse $g^{-1} \in \mathcal{G}$ such that $g \circ g^{-1} = g^{-1} \circ g = e$.

A subgroup is a subset of $\mathcal{G}$ that is also a group.

A *group action* arises when elements of a group act on those of a set $\mathcal{Y}$. In the present case the group elements g_θ typically correspond to elements of a parameter space Θ and $\mathcal{Y}$ is the sample space of a random variable Y. The action of g on y, $g(y)$, say, is defined for each $y \in \mathcal{Y}$ and $g(y)$ is an element of $\mathcal{Y}$ for each $g \in \mathcal{G}$.

Setting $y \approx y'$ if and only if there is a $g \in \mathcal{G}$ such that $y = g(y')$ gives an equivalence relation, which partitions $\mathcal{Y}$ into equivalence classes called *orbits* and labelled by an index a, say. Each y belongs to precisely one orbit, and can be represented by a and its position on the orbit. Hence we can write $y = g(a)$ for some $g \in \mathcal{G}$. If this representation is unique for a given choice of index, the group action is said to be *free*.

Example 5.19 (Location model) Let $Y = \theta + \varepsilon$, where $\theta \in \Theta = \mathbb{R}$ and ε is a scalar random variable with known density $f(y)$, where $y \in \mathbb{R}$. The density of Y is $f(y - \theta) = f(y; \theta)$, say, and that of $\theta' + Y = \theta' + \theta + \varepsilon$ is $f(y; \theta + \theta')$. Thus adding θ' to Y changes the parameter of the density. Taking $\theta' = -\theta$ gives the baseline density $f(y; 0) = f(y)$ of ε.

Here group elements may be written g_θ, corresponding to the parameters θ, and the group operation is equivalent to addition. Hence $g_\theta \circ g_{\theta'} = g_{\theta+\theta'}$, the identity e is g_0 and the inverse of g_θ is $g_{-\theta}$. Each element of the group corresponds to a point in Θ, but it induces a group action $g_\theta(y) = \theta + y$ on the sample space.

For a random sample $Y_1, \ldots, Y_n$, we take $\mathcal{Y} = \mathbb{R}^n$ and interpret expressions such as $g_\theta(Y) = \theta + Y$ as vectors, with $\theta \equiv \theta 1_n$ and $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$. Then y and y' belong to the same orbit if there exists a g_θ such that $g_\theta(y) = y'$, that is, there exists a θ such that $\theta + y = y'$, and this implies that y' is a location shift of y. On taking $\theta = \overline{y}' - \overline{y}$ we see that $y - \overline{y} = y' - \overline{y}'$, implying that we can represent the orbit by

1_n is the $n \times 1$ vector of ones.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

the vector $a(y) = y - \overline{y}$, because this choice of index gives $a(y) = a(y')$. Thus y is equivalently written as $(y - \overline{y}, \overline{y})$, where the first term indexes the orbit and the second the position of y within it. In terms of this representation we write y as $g_{\overline{y}}(a) = \overline{y} + a = \overline{y} + y - \overline{y} = y$. The group action is free because $g_\theta(a) = y$ implies that $\theta = \overline{y}$.

In geometric terms, $a(y)$ lies on the $(n-1)$-dimensional hyperplane $\sum a_j = 0$, each point of which determines a different orbit. The orbits themselves are lines $\theta + a(y)$ passing through these points, with $\theta \in \mathbb{R}$. When $n = 2$, each point (y_1, y_2) in $\mathbb{R}^2$ is indexed by a point on the line $y_1 + y_2 = 0$, which determines the orbit, a straight line perpendicular to this. ■

Two points y and y' on the same orbit have the same index $a = a(y)$, which is said to be *invariant* to the action of the group because its value does not depend on whether y or $g(y)$ was observed, for any $g \in \mathcal{G}$. It is *maximal invariant* if every other invariant statistic is a function of it, or equivalently

$$a(y) = a(y') \text{ implies that } y' = g(y) \text{ for some } g \in \mathcal{G}.$$

The distribution of $A = a(Y)$ does not depend on the elements of $\mathcal{G}$. In the present context these are identified with parameter values, so the distribution of A does not depend on parameters and is known in principle; A is said to be *distribution constant*. A maximal invariant can be thought of as a reduced version of the data that represents it as closely as possible while remaining invariant to the action of $\mathcal{G}$. In some sense it is what remains of Y once minimal information about the parameter values has been extracted.

Often there is a 1–1 correspondence between the elements of $\mathcal{G}$ and the parameter space Θ, and then the action of $\mathcal{G}$ on $\mathcal{Y}$ induces a group action on Θ. If we can write g_θ for a general element of $\mathcal{G}$, then $g \circ g_\theta = g_{\theta'}$ for some $\theta' \in \Theta$. Hence g has mapped θ to θ', thereby inducing an action on Θ. In principle the action of g on Θ might be different from its action on $\mathcal{Y}$, and it is clearer to think of two related groups $\mathcal{G}$ and $\mathcal{G}^*$, the second of which acts on Θ. We use g^*_θ to denote the element of $\mathcal{G}^*$ that corresponds to $g_\theta \in \mathcal{G}$. In many cases the action of $\mathcal{G}^*$ is *transitive*, that is, each parameter can be obtained by applying an element of the group to a single baseline parameter.

Example 5.20 (Permutation group) Permutation of the indices of a random sample $Y_1, \ldots, Y_n$ should leave any inference unaffected. Hence we may consider the group of permutations π, with $g_\pi(y)$ representing the permuted version of $y \in \mathbb{R}^n$. Note that π^{-1} is also a permutation, as is the operation that leaves the indices of y unchanged. In the location model we might let $\mathcal{G}$ be the group containing all $n!$ of the g_π in addition to the g_θ. Though well-defined on the sample space, g_π has no counterpart in the parameter space, and so the enlarged group is not transitive.

To check that $a(y) = (y_{(1)} - \overline{y}, \ldots, y_{(n)} - \overline{y})^{\mathrm{T}}$ is a maximal invariant, note that if $a(y) = a(y')$, then permutations π, π' exist such that $g_\pi \circ g_{-\overline{y}}(y) = g_{\pi'} \circ g_{-\overline{y}'}(y')$. This in turn implies that $g^{-1}_{-\overline{y}'} \circ g^{-1}_{\pi'} \circ g_\pi \circ g_{-\overline{y}}(y) = y'$. Hence a is a maximal invariant.

If permutations are not included in the group, the same argument shows that $(y_1 - \overline{y}, \ldots, y_n - \overline{y})^{\mathrm{T}}$ is a maximal invariant. Thus the maximal invariant depends on the chosen group. ■

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

We shall usually ignore permutations of the order of a random sample, because the discussion below is simpler if the group considered is transitive.

Equivariance

A statistic $S = s(Y)$ defined on $\mathcal{Y}$ and taking values in the parameter space Θ is said to be *equivariant* if $s(g_\theta(Y)) = g^*_\theta(s(Y))$ for all $g_\theta \in \mathcal{G}$. Often S is chosen to be an estimator of θ, and then it is called an *equivariant estimator*. Maximum likelihood estimators are equivariant, because of their transformation property, that if $\phi = \phi(\theta)$ is a 1–1 transformation of the parameter θ, then $\widehat{\phi} = \phi(\widehat{\theta})$, where $\widehat{\theta} = s(Y)$ is the maximum likelihood estimator of θ. If the transformation ϕ corresponds to $g^*_\phi \in \mathcal{G}^*$, and $g_\phi(Y)$ is the transformation of Y whose maximum likelihood estimator is $\widehat{\phi}$, then $\widehat{\phi} = s(g_\phi(Y))$, while $\phi(\widehat{\theta}) = g^*_\phi(s(Y))$. Hence $s(g_\phi(Y)) = g^*_\phi(s(Y))$ for all such g_ϕ, which is the requirement for equivariance.

An equivariant estimator can be used to construct a maximal invariant. Note first that as $s(Y) \in \Theta$, the corresponding group elements $g^*_{s(Y)} \in \mathcal{G}^*$ and $g_{s(Y)} \in \mathcal{G}$ exist. Now consider $a(Y) = g^{-1}_{s(Y)}(Y)$. If $a(Y) = a(Y')$, then $g^{-1}_{s(Y)}(Y) = g^{-1}_{s(Y')}(Y')$, and it follows that $Y' = g_{s(Y')} \circ g^{-1}_{s(Y)}(Y)$. Hence $A = a(Y) = g^{-1}_{s(Y)}(Y)$ is maximal invariant.

Example 5.21 (Location-scale model) Let $Y = \eta + \tau\varepsilon$, where as before ε has a known density f, and the parameter $\theta = (\eta, \tau) \in \Theta = \mathbb{R} \times \mathbb{R}_+$. The group action is $g_\theta(y) = g_{(\eta,\tau)}(y) = \eta + \tau y$, so

$$g_{(\eta,\tau)} \circ g_{(\mu,\sigma)}(y) = g_{(\eta,\tau)}(\mu + \sigma y) = \eta + \tau\mu + \tau\sigma y = g_{(\eta+\tau\mu,\tau\sigma)}(y). \tag{5.22}$$

The set of such transformations is closed with identity $g_{(0,1)}$. It is easy to check that $g_{(\eta,\tau)}$ has inverse $g_{(-\eta/\tau,\tau^{-1})}$. Therefore

$$\mathcal{G} = \left\{ g_{(\eta,\tau)} : (\eta, \tau) \in \mathbb{R} \times \mathbb{R}_+ \right\}$$

is indeed a group under the operation $\circ$ defined above.

The action of $g_{(\eta,\tau)}$ on a random sample is $g_{(\eta,\tau)}(Y) = \eta + \tau Y$, with $\eta \equiv \eta 1_n$ and Y an $n \times 1$ vector, as in Example 5.19. Expression (5.22) implies that the implied group action on Θ is

$$g^*_{(\eta,\tau)}((\mu, \sigma)) = (\, \eta + \tau\mu, \tau\sigma \,).$$

The sample average and standard deviation are equivariant, because with $s(Y) = (\overline{Y}, V^{1/2})$, where $V = (n-1)^{-1} \sum (Y_j - \overline{Y})^2$, we have

$$\begin{aligned} s(g_{(\eta,\tau)}(Y)) &= \left(\overline{\eta + \tau Y}, \left\{ (n-1)^{-1} \sum (\eta + \tau Y_j - \overline{\eta + \tau Y})^2 \right\}^{1/2} \right) \\ &= \left(\eta + \tau\overline{Y}, \left\{ (n-1)^{-1} \sum (\eta + \tau Y_j - \eta - \tau\overline{Y})^2 \right\}^{1/2} \right) \\ &= \left(\eta + \tau\overline{Y}, \tau V^{1/2} \right) \\ &= g^*_{(\eta,\tau)}\,(s(Y)). \end{aligned}$$

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

A maximal invariant is $A = g_{s(Y)}^{-1}(Y)$, and the parameter corresponding to $g_{s(Y)}^{-1}$ is $(-\overline{Y}/V^{1/2}, V^{-1/2})$. Hence a maximal invariant is the vector of residuals

$$A = (Y - \overline{Y})/V^{1/2} = \left(\frac{Y_1 - \overline{Y}}{V^{1/2}}, \ldots, \frac{Y_n - \overline{Y}}{V^{1/2}} \right)^{\mathrm{T}}, \tag{5.23}$$

also called the *configuration*. It can be checked directly that the distribution of A depends on n and f but not on θ. Any function of A is invariant. If permutations are added to $\mathcal{G}$, a maximal invariant is $A = (Y_{(\cdot)} - \overline{Y})/V^{1/2}$, where $Y_{(\cdot)} = (Y_{(1)}, \ldots, Y_{(n)})$ represents the vector of ordered values of Y.

The orbits are determined by different values a of the statistic A, and Y has a unique representation as $g_{s(Y)}(A) = \overline{Y} + V^{1/2}A$. Hence the group action is free.

The elements of a satisfy the equations $\sum a_j = 0$ and $\sum a_j^2 = n - 1$, so A lies on an $(n-2)$-dimensional surface in $\mathbb{R}^n$. When $n = 3$ this is easily visualized; it is the circle that forms the intersection of the sphere of radius 2 with the plane $a_1 + a_2 + a_3 = 0$. The entire space $\mathbb{R}^3$ is generated by first choosing an element of this circle, then multiplying it by a positive number to rescale it to lie on a ray passing through the origin, and finally adding the vector $\overline{y}1_3$.

Another equivariant estimator is $(Y_{(1)}, Y_{(2)} - Y_{(1)})$, where $Y_{(r)}$ is the rth order statistic, and the argument above shows that the vector $(Y - Y_{(1)})/(Y_{(2)} - Y_{(1)})$ is corresponding maximal invariant. Evidently this is just one of many possible location-scale shifts of A, which can be thought of as the 'shape' of the sample, shorn of information about its location and scale. ■

The group-averse reader may wonder whether the generality of the discussion above is needed to deal with our motivating example of temperatures in Celsius and Fahrenheit. In fact we have not yet raised a crucial distinction between invariances intrinsic to a context and those stemming only from the mathematical structure of the model. Invariances of the first sort are more defensible than are the second, because not every mathematical expression of a statistical problem successfully preserves aspects such the interpretation of key parameters. Thus the sensible choice of group in a particular context may not be mathematically most natural. Furthermore appeal to invariance is not sensible if external information suggests that some parameter values should be favoured over others. Invariance arguments require careful thought.

Example 5.22 (Venice sea level data) The straight-line regression model (5.2) can be expressed as

$$y = X\beta + \varepsilon,$$

where

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

An $n \times n$ orthogonal matrix of real numbers O has the properties that $O^{\mathrm{T}}O = OO^{\mathrm{T}} = I_n$.

If the ε_j are independent normal variables then $Y \sim N_n(X\beta, \sigma^2 I_n)$. Hence $OY \sim N_p(OX\beta, \sigma^2 I_n)$ for any $n \times n$ orthogonal matrix O that preserves the column space of X, that is, such that $X(X^{\mathrm{T}}X)^{-1}XOX = OX$. It is straightforward to check that such matrices form a group. Now $\mathrm{E}(OY) = X\gamma$, where $\gamma = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}OX\beta = A^{-1}\beta$, say, is the result of applying the corresponding group element in the parameter space.

The transformation giving (5.3), with

$$\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \beta = A\gamma = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \gamma = \begin{pmatrix} 1 & -\overline{x} \\ 0 & 1 \end{pmatrix} \gamma = \begin{pmatrix} \gamma_0 - \gamma_1 \overline{x} \\ \gamma_1 \end{pmatrix},$$

preserves the interpretation of $\beta_1 = a_{22}\gamma_1$ as a rate of change of $\mathrm{E}(Y)$ with respect to time, though the time origin is shifted. From a mathematical viewpoint there is no reason not to take more general invertible transformations $\beta = A\gamma$, for example with $a_{21} \neq 0$, but this makes no sense statistically. Moreover even with $a_{21} = 0$ not every choice of a_{22} makes sense: taking $a_{22} < 0$ or such that the units of γ_1 were seconds would have little appeal. ■

In some cases the full parameter space does not give a useful group of transformations, but subspaces of it do. If the parameter space has form $\Psi \times \Lambda$, with the same group of transformations $\mathcal{G} = \{g_\lambda : \lambda \in \Lambda\}$ acting on the sample space for each value of ψ, then we have a *composite group transformation model*.

Example 5.23 (Location-scale model) In the previous example, suppose that the density f_ψ of ε depends on a further parameter ψ. An example is the t_ψ density. Then for each fixed ψ we have a location-scale model in terms of $\lambda = (\eta, \tau)$, with $g_\lambda(y) = \eta + \tau y$, and our previous discussion applies.

For each ψ a maximal invariant based on a random sample $Y_1, \ldots, Y_n$ is $A = (Y - \overline{Y})/V^{1/2}$, whose distribution depends on the sample size and on f_ψ but not on λ. ■

Exercises 5.3

1 Show that $\approx$ is an equivalence relation.

2 Suppose $Y = \tau\varepsilon$, where $\tau \in \mathbb{R}_+$ and ε is a random variable with known density f. Show that this scale model is a group transformation model with free action $g_\tau(y) = \tau y$. Show that $s_1(Y) = \overline{Y}$ and $s_2(Y) = (\sum Y_j^2)^{1/2}$ are equivariant and find the corresponding maximal invariants. Sketch the orbits when $n = 2$.

3 Suppose that ε has known density f with support on the unit circle in the complex plane, and that $Y = e^{i\theta}\varepsilon$ for $\theta \in \mathbb{R}$. Show that this is a group transformation model. Is it transitive? Is the action free?

4 Write the configuration (5.23) in terms of $\varepsilon_1, \ldots, \varepsilon_n$, where $Y_j = \mu + \sigma\varepsilon_j$, and thereby show that its distribution does not depend on the parameters.

5 Show that the gamma density with shape and scale parameters ψ and λ, is a composite transformation model under the mapping from Y to τY, where $\tau > 0$.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

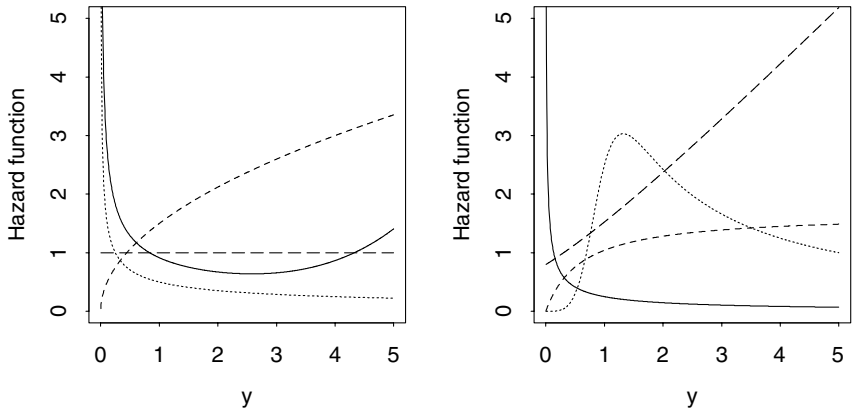

Figure 5.7 Hazard functions. Left panel: Weibull hazards with $\theta = 1$ and $\alpha = 0.5$ (dots), $\alpha = 1$ (large dashes), $\alpha = 1.5$ (dashes), and bi-Weibull hazard with $\theta_1 = 0.3$, $\alpha_1 = 0.5$, $\theta_2 = \alpha_2 = 5$ (solid). Right panel: Log-logistic hazards with $\lambda = 1$ and $\alpha = 0.5$ (solid), $\alpha = 5$ (dots), gamma hazard with $\lambda = 0.6$ and $\alpha = 2$ (dashes), and standard normal hazard (large dashes).

5.4 Survival Data

5.4.1 Basic ideas

The focus of interest in survival data is the time to an event. An important area of application is medicine, where, for example, interest may centre on whether a new treatment lengthens the life of a cancer patient, relative to those who receive existing treatments. Other common applications are in industrial reliability, where the aim may be to estimate the distribution of time to failure for a fridge, a computer program, or a pacemaker. Examples also abound in the social sciences, where for example the length of a period of unemployment may be of interest. In each case the time Y to the event is non-negative and may be censored. For example, a patient may be lost to follow-up for some reason unrelated to his disease, so that it is unknown whether or not he died from the cause under study. In general discussion we refer to the items liable to fail as *units*; these may be persons, widgets, marriages, cars, or whatever.

This section outlines some basic notions in survival analysis, concentrating on single samples. More complex models are discussed in Section 10.8.

Hazard and survivor functions

A central concept is the *hazard function* of Y, defined loosely as the probability density of failure at time y, given survival to then. If Y is a continuous random variable this is

$$h(y) = \lim_{\delta y \to 0} \frac{1}{\delta y} \Pr(y \le Y < y + \delta y \mid Y \ge y) = \frac{f(y)}{\mathcal{F}(y)},$$

where $\mathcal{F}(y) = \Pr(Y \ge y) = 1 - F(y)$ is the *survivor function* of Y. An older term for $h(y)$ is the *force of mortality*, and it is also called the *age-specific failure rate*. Evidently $h(y) \ge 0$; some example hazard functions are shown in Figure 5.7.

The exponential density with rate λ has $\mathcal{F}(y) = \exp(-\lambda y)$ and constant hazard function $h(y) = \lambda$, and although data are rarely so simple, this model of a constant failure rate independent of the past is a natural baseline from which to develop more realistic models.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Or integrated hazard function.

The *cumulative hazard function* is

$$H(y) = \int_0^y h(u)\,du = \int_0^y \frac{f(u)}{1-F(u)}\,du = -\log\{1-F(y)\},$$

as $F(0) = 0$. Thus the survivor function may be written as $\mathcal{F}(y) = \exp\{-H(y)\}$, and $f(y) = h(y)\exp\{-H(y)\}$. If $\lim_{y\to\infty} H(y) < \infty$, then $\mathcal{F}(\infty) > 0$ and the distribution is *defective*, putting positive probability on an infinite survival time. This may arise in practice if, for example, the endpoint for a study is death from a disease, but complete recovery is possible.

For a discrete distribution with probabilities f_i at $0 \le t_1 < t_2 < \cdots$, we may write $h(y) = \sum h_i \delta(y - t_i)$, where

$$h_i = \Pr(Y = t_i \mid Y \ge t_i) = \frac{f_i}{f_i + f_{i+1} + \cdots}.$$

Thus

$$\Pr(Y > t_i \mid Y \ge t_i) = 1 - h_i, \quad f_i = h_i \prod_{j=1}^{i-1}(1-h_j), \tag{5.24}$$

and if $t_i < y \le t_{i+1}$ then

$$\begin{aligned}\mathcal{F}(y) &= \Pr(Y > t_i \mid Y \ge t_i)\Pr(Y > t_{i-1} \mid Y \ge t_{i-1}) \cdots \Pr(Y > t_1) \\ &= \prod_{i:t_i<y}(1-h_i).\end{aligned} \tag{5.25}$$

We define the cumulative hazard as $H(y) = -\sum_{i:t_i<y}\log(1-h_i)$, again giving $\mathcal{F}(y) = \exp\{-H(y)\}$. The more natural definition $\sum_{i:t_i<y} h_i$ is approximately equal to $H(y)$ if the individual h_i are small.

Mixed discrete-continuous variables are important in a general treatment of survival data — for example, a patient may die so fast from complications after an operation that the survival time is effectively zero, but otherwise may live for years — but here we avoid them and the complications they bring.

Suppose that $Y = \min(Y_1, \ldots, Y_k)$, where the Y_i are continuous times to failure from k independent causes, and that their hazard functions are $h_i(y)$. Then Y exceeds y if and only if all the Y_i exceed y, so

$$\mathcal{F}(y) = \prod_{i=1}^k \Pr(Y_i \ge y) = \exp\left\{-\sum_{i=1}^k \int_0^y h_i(u)\,du\right\},$$

and it follows that Y has hazard function $h(y) = \sum h_i(y)$. That is, hazards for independent causes of failure are added.

Example 5.24 (Weibull density) The Weibull density (4.4) has survivor function $\mathcal{F}(y) = \exp\{-(y/\theta)^\alpha\}$, so its hazard function is $\alpha\theta^{-\alpha}y^{\alpha-1}$. This is constant when $\alpha = 1$, decreasing when $\alpha < 1$, and increasing when $\alpha > 1$, as shown in the left panel of Figure 5.7. This flexibility and the tractability of its density and distribution functions make the Weibull a popular choice in reliability studies.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

This density is the basis of the bi-Weibull model, which corresponds to the minimum of two independent Weibull variables, shown by the argument above to have hazard function $\alpha_1\theta_1^{-\alpha_1}y^{\alpha_1-1} + \alpha_2\theta_2^{-\alpha_2}y^{\alpha_2-1}$. If the shape parameters lie on opposite sides of unity, so $0 < \alpha_1 < 1 < \alpha_2$, say, $h(y)$ is bathtub-shaped: there is a high early failure rate during a 'burn-in period', then a flattish hazard and an eventual increase in failure rate; see Figure 5.7. If $\alpha_1 = \alpha_2$ the hazard is indistinguishable from the Weibull hazard and θ_1 and θ_2 are not identifiable. ■

Example 5.25 (Log-logistic density) The log-logistic distribution has survivor and hazard functions

$$\mathcal{F}(y) = \{1 + (\lambda y)^\alpha\}^{-1}, \quad h(y) = \alpha\frac{\lambda^\alpha y^{\alpha-1}}{1 + (\lambda y)^\alpha}, \quad y > 0, \alpha, \lambda > 0.$$

Two examples of $h(y)$ are shown in the right panel of Figure 5.7. It is decreasing for $\alpha \leq 1$ and unimodal otherwise. The log-normal distribution, that is, the distribution of $Y = e^Z$, where Z has a normal distribution, is similar to the log-logistic, and its hazard can take similar shapes. The normal hazard, also shown, increases very rapidly due to the light tails of the normal density. ■

Example 5.26 (Gamma density) The gamma survivor and hazard functions are

$$\mathcal{F}(y) = \int_y^\infty \frac{\lambda^\alpha u^{\alpha-1}}{\Gamma(\alpha)}e^{-\lambda u}\,du, \quad h(y) = \frac{\lambda^\alpha y^{\alpha-1}e^{-\lambda y}}{\int_y^\infty \lambda^\alpha u^{\alpha-1}e^{-\lambda u}\,du}.$$

Figure 5.7 shows an example of the gamma hazard function. ■

Censoring

The simplest form of censoring occurs when a random variable Y is watched until a pre-determined time c. If $Y \leq c$, we observe the value y of Y, but if $Y > c$, we know only that Y survived beyond c. This is known as *Type I censoring*. *Type II censoring* arises when n independent variables are observed until there have been r failures, so the first r order statistics $0 < Y_{(1)} < \cdots < Y_{(r)}$ are observed, All that is known about the $n - r$ remaining observations is that they exceed $Y_{(r)}$. This scheme is typically used in industrial life-testing.

For simplicity we assume no ties.

Under *random censoring* we suppose that the jth of n independent units has an associated censoring time C_j drawn from a distribution G, independent of its survival time Y_j^0. The time actually observed is $Y_j = \min(Y_j^0, C_j)$, and it is known whether or not $Y_j = Y_j^0$, an event indicated by D_j. Thus a pair (y_j, d_j) is observed for each unit, with $d_j = 1$ if y_j is the survival time and $d_j = 0$ if y_j is the censoring time. This type of censoring is important in medical applications, where a patient may die of a cause unrelated to the reason they are being studied, may withdraw from the study or be lost to follow-up, or the study may end before their survival time is observed.

Figure 5.8 shows the relation between calendar time and time on trial for a medical study, with censoring both before and at the end of the trial. We assume below that failure does not depend on the calendar time at which an individual enters the study;

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

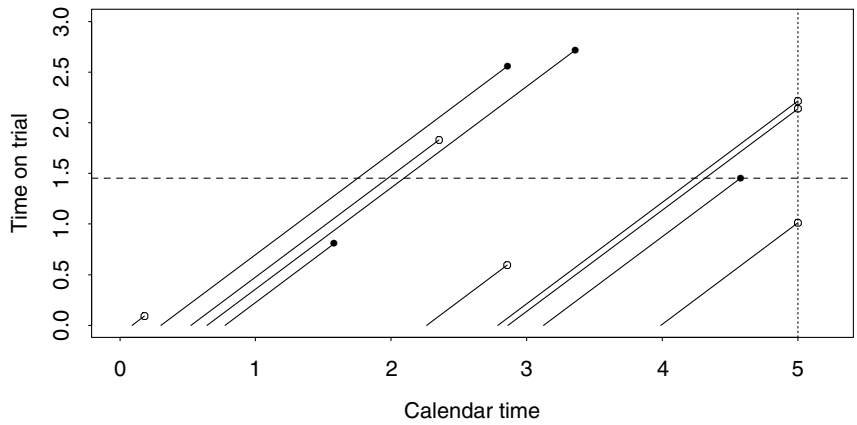

Figure 5.8 Lexis diagram showing typical pattern of censoring in a medical study. Each individual is shown as a line whose x coordinates run from the calendar time of entry to the trial to the calendar time of failure (blob) or censoring (circle). Censoring occurs at the end of the trial, marked by the vertical dotted line, or earlier. The vertical axis shows time on trial, which starts when individuals enter the study. The risk set for the failure at calendar time 4.5 comprises those individuals whose lines touch the horizontal dashed line; see page 543.

thus we study events on the vertical axis. Calendar time may be used to account for changes in medical practice over the course of a trial.

In applications the assumption that C_j and Y_j^0 are independent is critical. There would be serious bias if the illest patients drop out of a trial because the treatment makes them feel even worse, thereby inducing association between survival and censoring variables because patients die soon after they withdraw.

The examples above all involve *right-censoring*. Less common is left-censoring, where the time of origin is not known exactly, for example if time to death from a disease is observed, but the time of infection is unknown.

In practice a high proportion of the data may be censored, and there may be a serious loss of efficiency if they are ignored (Example 4.20). There will also be bias, as survival probabilities will be underestimated if censoring is not taken into account. Hence it is crucial to make proper allowance for censoring.

5.4.2 Likelihood inference

Suppose that the survival times are continuous, that data $(y_1, d_1), \ldots, (y_n, d_n)$ on n independent units are available, and that there is a parametric model for survival times, with survivor and hazard functions $\mathcal{F}(y;\theta)$ and $h(y;\theta)$. Recall that the density may be written $f(y;\theta) = h(y;\theta)\mathcal{F}(y;\theta)$ and that in terms of the integrated hazard function, $\mathcal{F}(y;\theta) = \exp\{-H(y;\theta)\}$. Under random censoring in which the censoring variables have density and distribution functions g and G, the likelihood contribution from y_j is

$$f(y_j;\theta)\{1 - G(y_j)\} \quad \text{if } d_j = 1, \quad \text{and} \quad \mathcal{F}(y_j;\theta)g(y_j) \quad \text{if } d_j = 0.$$

If the censoring distribution does not depend on θ, then $g(y_j)$ and $G(y_j)$ are constant and the overall log likelihood is

$$\ell(\theta) \equiv \sum_u \log f(y_j;\theta) + \sum_c \log \mathcal{F}(y_j;\theta),$$

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

0+	1+	1+	3+	3+	7	10+	11+	12+	12+	15+	18+
20+	22+	22+	24+	25+	26+	31+	36+	36+	36	38	40
47+	47+	49+	53+	53+	55+	56+	57+	61+	67+	67+	70
73	75+	77+	83+	84+	88+	89+	99	121+	122+	123+	141+
0+	0+	2+	2+	2+	2+	3	3+	4+	5+	9+	10+
11	12+	13	13+	18+	22+	22+	24+	24+	24+	25+	26+
27	28	32+	35+	36	40+	43+	50+	54			

Table 5.3
Blalock–Taussig shunt data (Oakes, 1991). The table gives survival time of shunt (months after operation) for 48 infants aged over one month at time of operation, followed by times for 33 infants aged 30 or fewer days at operation. Infants whose shunt has not yet failed are marked +.

where the sums are over uncensored and censored units. This amounts to treating the censoring pattern as fixed, and encompasses Type I censoring, for which G puts all its probability at c. In terms of the hazard function and its integral, the log likelihood is

$$\ell(\theta) = \sum_{j=1}^{n} \{d_j \log h(y_j;\theta) - H(y_j;\theta)\}. \tag{5.26}$$

Inference for θ is based on this in the usual way. As calculation of expected information involves assumptions about the censoring mechanism, standard errors for parameter estimates are based on observed information.

Example 5.27 (Exponential distribution) When $f(y;\lambda) = \lambda e^{-\lambda y}$, the hazard is $h(y;\lambda) = \lambda$, and hence the log likelihood for a random sample $(y_1, d_1), \ldots, (y_n, d_n)$ is

$$\ell(\lambda) = \sum_{j=1}^{n} (d_j \log\lambda - \lambda y_j) = \log\lambda \sum_{j=1}^{n} d_j - \lambda \sum_{j=1}^{n} y_j,$$

giving maximum likelihood estimate $\widehat{\lambda} = \sum d_j / \sum y_j$ and observed information $J(\lambda) = \sum d_j/\lambda^2$; see Example 4.20. Hence the estimate of λ is zero if there are no failures, and censored data contribute no information about λ.

The expected information $I(\lambda) = \mathrm{E}\{J(\lambda)\}$ involves $\mathrm{E}(D_j)$, where D_j indicates whether a failure or censoring time is observed for the jth observation, but this expectation cannot be obtained without some assumption about the censoring distribution G. Although this is feasible for theoretical calculations such as those in Example 4.20, in practice the inverse observed information is used to give a standard error $J(\widehat{\lambda})^{-1/2}$ for $\widehat{\lambda}$.

The mean of the exponential density is $\theta = \lambda^{-1}$, and its maximum likelihood estimate is $\widehat{\theta} = \sum y_j / \sum d_j$, with observed information $J(\widehat{\theta}) = \widehat{\theta}^2 / \sum d_j$ and maximized log likelihood $\ell(\widehat{\theta}) = -(1 + \log\widehat{\theta}) \sum d_j$. ■

Example 5.28 (Blalock–Taussig shunt data) The Blalock–Taussig shunt is an operative procedure for infants with congential cyanotic heart disease. Table 5.3 contains data from the University of Rochester on survival times for the shunt for 81 infants, divided into two age groups. Many of the survival times are censored, meaning that the shunt was still functioning after the given survival time; its time to failure is not known for these children, whereas it is known for the others. There are just seven failures in each group. The table suggests that the shunt fails sooner for younger children, and it is of interest to see how failure depends on age.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

A simple model for these data is that the failure times are independent exponential variables, with common mean θ for both groups. Formulae from Example 5.27 show that $\widehat{\theta} = 209.1$ and the maximized log likelihood is -88.79. If the means are different, θ_1 and θ_2, say, then the maximized log likelihood is -85.98, so the likelihood ratio statistic for comparing these models is $2 \times (88.79 - 85.98) = 5.62$, to be compared with the χ^2_1 distribution. As $\Pr(\chi^2_1 \geq 5.62) \doteq 0.018$, there is strong evidence that the mean survival time is shorter for the younger group, if the exponential model is correct.

If the data were uncensored, it would be straightforward to assess the fit of this model using probabability plots, but the amount of censoring is so high that this is not sensible. More specialized methods are needed, and they are discussed in Section 5.4.3.

One way to judge adequacy of the exponential model is to embed it in a larger one. A simple alternative is to suppose that the data are Weibull, with $H(y) = (y/\theta)^\alpha$. The maximized log likelihoods are -83.72 when this model is fitted separately to each group, and -83.74 when the same value of α is used for both groups. The likelihood ratio statistic for comparison of these is $2 \times (83.74 - 83.72) = 0.04$, which is negligible, but that for comparison with the best exponential model, $2 \times (85.98 - 83.74) = 4.48$, suggests that the Weibull model gives the better fit. The corresponding estimates and their standard errors are $\widehat{\theta}_1 = 181.1$ (52.7), $\widehat{\theta}_2 = 57.6$ (15.1), and $\widehat{\alpha} = 1.64$ (0.35). The value of $\widehat{\alpha}$ corresponds to an increasing hazard. ■

Discrete data

Suppose that events could occur at pre-assigned times $0 \leq t_1 < t_2 < \cdots$, and that under a parametric model of interest the hazard function at t_i is $h_i = h_i(\theta)$. We adopt the convention that a unit censored at time t_i could have been observed to fail there, so giving likelihood contribution

$$\lim_{y \downarrow t_i} \mathcal{F}(y) = (1 - h_1) \cdots (1 - h_i),$$

from (5.25); one way to think of this is that censoring at t_i in fact takes place immediately afterwards. The contribution to the likelihood from a unit that fails at t_i is $(1 - h_1) \cdots (1 - h_{i-1})h_i$; see (5.24). Although the likelihood can be written down directly, it is more useful to express it in terms of the r_i units still in the *risk set* — that is not yet failed or censored — at time t_i and the number d_i of units who fail there. This modifies our previous notation: now d_i is the sum of the indicators of unit failures at time t_i, and can take one of values $0, 1, \ldots, r_i$. Each of the d_i failures at t_i contributes h_i to the likelihood, and the other units then still in view each contribute $1 - h_i$. It follows that the log likelihood may be written as

$$\ell(\theta) = \sum_i \{d_i \log h_i + (r_i - d_i) \log (1 - h_i)\}, \tag{5.27}$$

with the interpretation that the probability of failure at t_i conditional on survival to t_i is h_i, and d_i of the r_i units in view at t_i fail then. Thus (5.27) is a sum of contributions from independent binomial variables representing the numbers of failures d_i at each

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Age group	Hungary 900–1100	England 1640–89	Breslau 1687–91	England & Wales, 1841		England & Wales, 1980–82	
				Males	Females	Males	Females
30–35	0.0235	0.0171	0.0164	0.0108	0.0107	0.0010	0.0006
35–40	0.0291	0.0205	0.0195	0.0123	0.0118	0.0014	0.0009
40–45	0.0337	0.0195	0.0233	0.0140	0.0131	0.0024	0.0016
45–50	0.0402	0.0244	0.0282	0.0159	0.0145	0.0043	0.0028
50–55	0.0696	0.0307	0.0342	0.0181	0.0162	0.0079	0.0047
55–60	0.0814	0.0459	0.0383	0.0254	0.0220	0.0138	0.0076
60–65	0.1033	0.0513	0.0474	0.0375	0.0331	0.0227	0.0119
65–70	0.1485	0.0701	0.0630	0.0553	0.0493	0.0365	0.0187
70–75	0.1877	0.1129	0.0995	0.0815	0.0736	0.0587	0.0308
75–80	0.3008	0.1445	0.1589	0.1201	0.1097	0.0930	0.0527
80–85		0.1974		0.1771	0.1638	0.1432	0.0919
85–90				0.2617	0.2448	0.2110	0.1567
90–95				0.3884	0.3674	0.2900	0.2374
95–100						0.3894	0.3215
Deaths	2300	3133	2675	71,000	74,000	834,000	828,000

Table 5.4 Historical estimates of the force of mortality (year^{-1}), averaged for 5-year age groups (Thatcher, 1999). The bottom line gives the estimated number of deaths at age 30 years and above, on which the force of mortality is based.

time t_i, with denominators r_i and failure probabilities h_i. In fact r_i depends on the history of failures and censorings up to time t_i, so the d_i are not independent, but it turns out that for large sample inference we may proceed as if they were. This can be formalized using the theory of counting processes and martingales; see the bibliographic notes to this chapter and to Chapter 10.

Example 5.29 (Human lifetime data) The virtual elimination of many infectious diseases due to improved medical care and living conditions have led to increased life expectancy in the developed world. If the trend continues there are potentially major consequences for social security systems. Some physicians have asserted that an upper limit to the length of human life is imposed by physical constraints, and that the consequence of improved health care is that senescence will eventually be compressed into a short period just prior to death at or near this upper limit. This view is controversial, however, and there is a lively debate about the future of old age.

A natural way to assess the plausibility of the hypothesized upper limit is to examine data on mortality. Table 5.4 contains historical snapshots of the force of mortality, obtained from census data, records of births and deaths, and other sources. The earliest data were obtained by forensic examination of adult skeletons in Hungarian graveyards, using a procedure that probably underestimates ages over 60 years and overestimates those below. The table shows estimates of the average probability of dying per year, conditional on survival to then, using the following argument. For continuous-time data with survivor function $\mathcal{F}(y)$ and corresponding hazard function $h(y)$, the probability of failure in the period $[t_i, t_{i+1})$ given survival to t_i would be

$$\frac{\mathcal{F}(t_i) - \mathcal{F}(t_{i+1})}{\mathcal{F}(t_i)} = 1 - \exp\left\{-(t_{i+1} - t_i)\frac{1}{t_{i+1} - t_i}\int_{t_i}^{t_{i+1}} h(y)\,dy\right\},$$

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

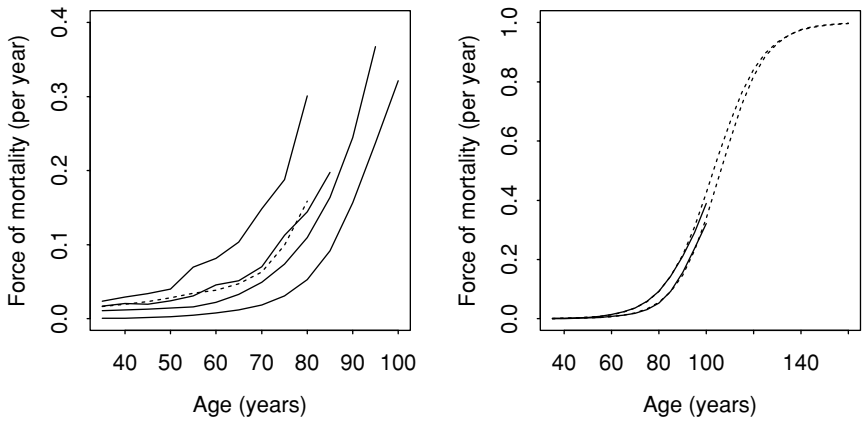

Figure 5.9 Force of mortality for historical data, in units of deaths per person-year. Left panel, from top to bottom: data for medieval Hungary, England 1640–89, Breslau 1687–91 (dots), English and Welsh females 1841 and 1980–82. Right panel: data for England and Wales, 1980–82, males (above) and females (below) and fitted hazard functions (dots).

where $(t_{i+1} - t_i)^{-1} \int_{t_i}^{t_{i+1}} h(y)\, dy$ is the average hazard over the interval. Given discretized data with r_i people alive at time t_i, of whom d_i fail in $[t_i, t_{i+1})$, the corresponding empirical hazard is $-(t_{i+1} - t_i)^{-1} \log(1 - d_i/r_i)$, and this is reported in the table; the corresponding d_i and r_i are unavailable to us. For British males dying in 1980 the empirical hazard rose from about 0.001 year^{-1} at age 30 years to about 0.1 year^{-1} at 80 years to about 0.4 year^{-1} at 95 years; for females the probabilities were slightly lower. Figure 5.9 shows the force of mortality of some of the columns of the table; it is no surprise that it is lower in later than in earlier periods.

One model for such data is that

$$h(y;\theta) = \lambda + \frac{\alpha e^{\beta y}}{1 + \alpha e^{\beta y}},$$

where $\theta = (\alpha, \beta, \lambda)$, corresponding to integrated hazard and survivor functions

$$H(y;\theta) = \lambda y + \frac{1}{\beta} \log\left(\frac{1 + \alpha e^{\beta y}}{1 + \alpha}\right), \quad \mathcal{F}(y;\theta) = e^{-\lambda y} \times \left(\frac{1 + \alpha}{1 + \alpha e^{\beta y}}\right)^{1/\beta}, \quad y \geq 0.$$

One interpretation of this model is that there are two competing causes of death, one with a constant hazard, and the other with a logistic hazard.

In order to use (5.27) to fit this model to the data given in Table 5.4, we must calculate $h_i(\theta)$ and (d_i, r_i). The probability of dying in $[t_i, t_{i+1})$ conditional on survival to t_i is

$$\begin{aligned} h_i(\theta) &= \Pr(t_i \leq Y \leq t_{i+1} \mid Y \geq t_i) \\ &= \frac{\mathcal{F}(t_i;\theta) - \mathcal{F}(t_{i+1};\theta)}{\mathcal{F}(t_i;\theta)} \\ &= 1 - \exp\{H(t_i;\theta) - H(t_{i+1};\theta)\}, \end{aligned}$$

and this is calculated using the logistic hazard given above. The empirical values of the hazard function $h_i = d_i/r_i$, where d_i is the number of deaths among the r_i persons at risk, can be obtained from the columns of Table 5.4. Some calculation gives

$$d_1 = nh_1, \quad d_i = nh_i(1 - h_1) \cdots (1 - h_{i-1}), \quad i = 2, \ldots, k,$$

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Data set	Deaths at age 30 years and over	Estimate (standard error)		
		$10^4\widehat{\alpha}$	$10^2\widehat{\beta}$	$10^2\widehat{\lambda}$
Hungary, 900–1100	2300	8.76 (3.78)	7.68 (0.65)	1.27 (0.32)
England, 1640–89	3133	1.87 (0.66)	8.65 (0.48)	1.40 (0.12)
Breslau, 1687–91	2675	1.44 (0.76)	8.88 (0.73)	1.57 (0.15)
England & Wales, 1841, males	71,000	0.50 (0.03)	10.08 (0.08)	0.97 (0.01)
England & Wales, 1841, females	74,000	0.32 (0.02)	10.50 (0.08)	0.97 (0.01)
England & Wales, 1980–82, males	834,000	0.46 (0.00)	9.93 (0.01)	−0.04 (0.00)
England & Wales, 1980–82, females	828,000	0.12 (0.00)	10.92(0.01)	0.03 (0.00)

Table 5.5 Maximum likelihood estimates for fits of logistic hazard model to the data in Table 5.4. Standard errors given as 0.00 are smaller than 0.005.

where $n = r_1$ is the number initially at risk, an estimate of which is given at the foot of the table; once the d_i are known the r_i are given by d_i / h_i. When these pieces are put together, maximum likelihood estimates of θ may be obtained by numerical maximization of (5.27), with standard errors based on the inverse observed information matrix, also obtained numerically. Table 5.5 shows that $\widehat{\alpha}$ and $\widehat{\lambda}$ decrease systematically with time, while the value of $\widehat{\beta}$ increases slightly but is broadly constant, close to 0.1. These are consistent with the overall decrease in the hazard function, but no change in its shape, that we see in the left panel of Figure 5.9. The values of $\widehat{\lambda}$ are generally similar to the observed force of mortality at age 30–35, and one interpretation is that $\widehat{\lambda}$ represents the danger from the principal risks at this age, namely infectious diseases and child-bearing, which has sharply reduced over the last 150 years.

The fits for the 1980–82 data are shown in the right panel of Figure 5.9. Although the fit is good, the extrapolation beyond the range of the data must be treated skeptically. It shows that although the model imposes no absolute upper limit on lifetimes, for a person dying in 1980–82 there was an effective limit of about 140 years, well beyond the limits of 110 or 115 years which have been suggested by physicians. In fact the longest life for which there is good documentation is that of Mme Jeanne Calment, who died in 1997 aged 122 years, and there is unlikely ever to be enough data to see if there is an upper limit well above this.

Example 5.32 gives further discussion of this model. ■

5.4.3 Product-limit estimator

Graphical procedures are essential for initial data inspection, for suggesting plausible models and for checking their fit. One standard tool is a nonparametric estimator of the survivor function, in effect extending the empirical distribution function (Example 2.7) to censored data.

The simplest derivation of it is based on the model for failures at discrete prespecified times given above (5.25), though the estimator is useful more widely. We therefore start with expression (5.27), which gives the log likelihood for such data in terms of the hazard function $h_1, h_2, \ldots$. For parametric analysis of a discrete failure distribution the h_i are functions of a parameter θ, but for nonparametric estimation we treat each h_i as a separate parameter and estimate it by maximum likelihood.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Differentiation of (5.27) with respect to h_i gives $\widehat{h}_i = d_i/r_i$ and hence

$$\widehat{\mathcal{F}}(y) = \prod_{i:t_i<y} (1-\widehat{h}_i) = \prod_{i:t_i<y} \left(1 - \frac{d_i}{r_i}\right).$$

This is known as the *product-limit* or *Kaplan–Meier* estimator. Note that

$$-\frac{\partial^2 \ell}{\partial h_i \partial h_j} = \begin{cases} \frac{r_i}{\widehat{h}_i(1-\widehat{h}_i)}, & i = j, \\ 0, & \text{otherwise}, \end{cases}$$

implying that that the $\widehat{h}_i$ are asymptotically independent, with diagonal variance matrix whose ith element is $\widehat{h}_i(1-\widehat{h}_i)/r_i$.

Edward Kaplan and Paul Meier were former students of John Tukey who submitted separate papers to *Journal of the American Statistical Association*. They were encouraged to merge them by the editor. Despite mixed reviews the editor decided to publish the joint paper (Kaplan and Meier, 1958), which has become one of the most-cited articles in statistics.

This derivation extends to continuous failure times by supposing that the y_j are ordered and that there are no ties, giving $t_1 = y_1 < \cdots < t_n = y_n$. Then d_j simply indicates whether y_j is a failure or a censoring time, and

$$\widehat{\mathcal{F}}(y) = \prod_{j:y_j<y} \left(1 - \frac{1}{r_j}\right)^{d_j}, \tag{5.28}$$

so the estimate decreases only at those values of t_j with $d_j = 1$. This estimate is valid also when the y_j are not pre-specified, but full justification of this would take us too far afield. If there is no censoring, then $1 - \widehat{\mathcal{F}}(y)$ is the empirical distribution function.

We find the variance of $\widehat{\mathcal{F}}(y)$ by arguing that if the d_i are asymptotically independent binomial variables with denominators r_i, then

$$\begin{aligned} \operatorname{var}\{\log \widehat{\mathcal{F}}(y)\} &= \operatorname{var}\left\{\sum_{i:y_i<y} \log(1-\widehat{h}_i)\right\} \\ &\doteq \sum_{i:y_i<y} \operatorname{var}\{\log(1-\widehat{h}_i)\} \\ &\doteq \sum_{i:y_i<y} \frac{1}{(1-\widehat{h}_i)^2} \frac{\widehat{h}_i(1-\widehat{h}_i)}{r_i} \\ &= \sum_{i:y_i<y} \frac{d_i}{r_i(r_i-d_i)}, \end{aligned} \tag{5.29}$$

Major Greenwood (1880–1949) qualified as a physician before turning to statistics and epidemiology under the influence of Karl Pearson. He was the first resident statistician at any medical research institute, and worked for the British Medical Research Council and the London School of Hygiene and Tropical Medicine. He studied infant mortality, tuberculosis and hospital fatality rates, pioneered clinical trials and gradually persuaded sceptical physicians of the value of statistical thinking. Major was not his military rank but his first name.

where the first approximation uses the asymptotic independence of the $\widehat{h}_i$ and the second uses the delta method. As $\operatorname{var}\{\log \widehat{\mathcal{F}}(y)\} \doteq \operatorname{var}\{\widehat{\mathcal{F}}(y)\}/\widehat{\mathcal{F}}(y)^2$, we obtain *Greenwood's formula*,

$$\operatorname{var}\{\widehat{\mathcal{F}}(y)\} \doteq \widehat{\mathcal{F}}(y)^2 \sum_{i:y_i<y} \frac{d_i}{r_i(r_i-d_i)},$$

variants of which are widely used to assess the uncertainty of $\widehat{\mathcal{F}}(y)$. In practice it is better to use (5.29) to compute approximate normal confidence intervals for $\log \mathcal{F}(y)$, and then to transform these intervals back to the original scale.

The cumulative hazard function can be estimated as $\widehat{H}(y) = \sum_{i:y_i<y} d_i/r_i$; this is a step function with jumps at failure times and approximate variance (5.29).

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Failure time, y_i	7	36	38	40	70	73	99
Number in view, r_i	43	29	26	25	13	12	5
Number failing, d_i	1	1	1	1	1	1	1
$1 - d_i/r_i$	0.977	0.966	0.962	0.960	0.923	0.916	0.8
$\widehat{\mathcal{F}}(y_i+)$	0.977	0.944	0.908	0.872	0.804	0.737	0.590
Standard error	0.023	0.040	0.052	0.062	0.086	0.102	0.155

Table 5.6 Product-limit estimator for older group of infants in Table 5.3.

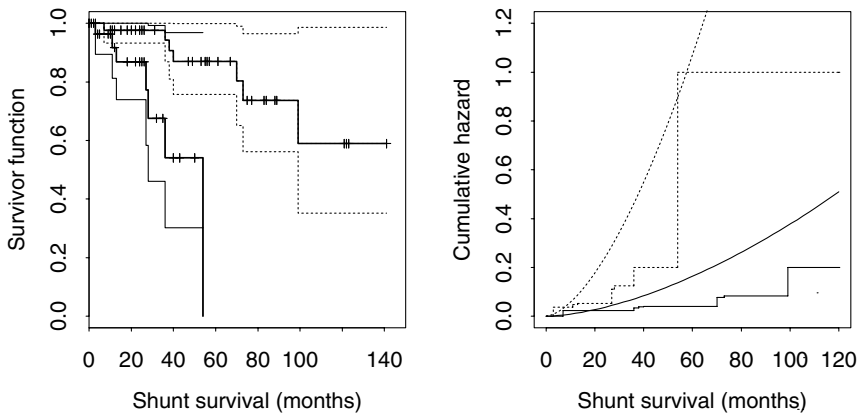

Figure 5.10 Nonparametric analysis of shunt data. Left panel: product-limit estimates of survivor function for older (upper heavy line) and younger infants (lower heavy line), with 95% confidence intervals (dots and light solid). Pluses on the product-limit estimates mark times of censored data. Right panel: estimated cumulative hazard functions for older (solid) and younger (dots) infants, using nonparametric estimate and fitted Weibull model (smooth curves).

Example 5.30 (Blalock–Taussig shunt data) Table 5.6 illustrates the calculation of the product-limit estimator using data from Table 5.3. As the estimator changes only at times of failures, it need not be calculated at censoring times. The estimate does not approach zero for large y because the largest observation in the sample is censored.

Estimated survivor functions for both groups are shown in the left panel of Figure 5.10, together with approximate 95% confidence intervals. There is a strong effect of age, with shunts failing appreciably sooner for the younger children. The right panel compares the cumulative hazard function estimators $\widehat{H}(y) = \sum_{i:y_i \le y} \widehat{h}_i$ with their parametric counterparts under the best Weibull model of Example 5.28. The parametric fits overstate the hazards appreciably. The apparent large difference after 60 months is largely due to a single failure in the younger group that strongly influences the analysis. ■

5.4.4 Other ideas

Competing risks

In some applications there may be different types of failure due to k different causes, say, and each failure time Y is accompanied by an indicator I showing which type of failure occurred. We can then define *cause-specific hazard functions*

$$h_i(y) = \lim_{\delta y \to 0} \frac{\Pr(y \le Y \le y + \delta y, I = i \mid Y \ge y)}{\delta y}, \quad y \ge 0,\ i = 1, \ldots, k,$$

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

corresponding to the rate at which failure of type i occurs, given survival to y. The overall hazard, cumulative hazard and survivor functions may be written

$$h(y) = \sum_{i=1}^{k} h_i(y), \quad H(y) = \sum_{i=1}^{k} \int_0^y h_i(u)\,du, \quad \mathcal{F}(y) = \exp\left\{ \sum_{i=1}^{k} \int_0^y h_i(u)\,du \right\}.$$

If we imagine observing a population of values of (Y, I), then each of the $h_i(y)$ would be known, but we would observe no other aspect of the population. Thus without further assumptions the only estimable quantities are functions of the $h_i(y)$ such as $H(y)$ and $\mathcal{F}(y)$.

The likelihood contribution from an uncensored failure of type i is $h_i(y)\mathcal{F}(y)$, while provided censoring is independent, that from a censored failure is $\mathcal{F}(y)$, because the corresponding I is unknown. Suppose that we have independent triplets $(y_1, i_1, d_1), \ldots, (y_n, i_n, d_n)$, where y_j is the jth survival time and $d_j = 1$ if it is uncensored. If so, i_j indicates its failure type, while $i_j = 0$, say, if $d_j = 0$. The likelihood based on these data is

$$\prod_{j=1}^{n} \mathcal{F}(y_j) \prod_{i=1}^{k} h_{i_j}(y_j)^{d_j} = \prod_{i=1}^{k} \left[\prod_{j=1}^{n} \exp\left\{ -\int_0^{y_j} h_i(y)\,du \right\} h_i(y_j)^{d_j I(i_j = i)} \right],$$

so it follows that to estimate $h_i(y)$ we treat any failure not of type i as a censoring. Thus, for example, the survivor function for $h_i(y)$ may be estimated by the product-limit estimator (5.28) with d_j replaced by $d_j I(i_j = i)$. Failures of types other than i are treated as censorings. Likewise for estimation of a parametric h_i.

For simplicity let $k = 2$. One way to think of competing risks is in terms of latent or potential failure times Y_1, Y_2 corresponding to the failure types. The observed quantities are $Y = \min(Y_1, Y_2)$ and $I = \{i : Y_i = Y\}$. Here Y_1 is interpreted as the time to failure that would be observed if cause 2 was removed, assuming that the failure time distribution for cause 1 when both causes of failure operate remains unchanged if cause 2 is eliminated. This assumption may be plausible in situations such as a reliability study where different types of failure are due to physically separate sub-systems and it is possible to imagine that all but one of these have been perfected, but the elimination of one failure type may alter the risk for others, particularly in medical contexts, where the assumption is often unsustainable. If it can be justified by appeal to subject-matter considerations it is very useful — the case for vaccination against infectious diseases, for example, presumes that removal of their risks increases overall survival.

An even stronger assertion is that Y_1 and Y_2 actually exist for each unit under study, with independence of causes of failure equivalent to independence of Y_1 and Y_2. In fact it is impossible to contradict this model. As mentioned above, the only observable quantities are functions of the cause-specific hazards $h_1(y)$ and $h_2(y)$. The joint survivor function

$$\mathcal{F}(y_1, y_2) = \Pr(Y_1 > y_1, Y_2 > y_2) = \exp\left\{ -\int_0^{y_1} h_1(u)\,du - \int_0^{y_2} h_2(u)\,du \right\}$$

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Lymphoma	159	189	191	198	200	207	220	235	245	250
	256	261	265	266	280	343	356	383	403	414
	428	432								
Sarcoma	317	318	399	495	525	536	549	552	554	557
	558	571	586	594	596	605	612	621	628	631
	636	643	647	648	649	661	663	666	670	695
	697	700	705	712	713	738	748	753		
Other	163	179	206	222	228	249	252	282	324	333
	341	366	385	407	420	431	441	461	462	482
	517	517	524	564	567	586	619	620	621	622
	647	651	686	761	763					
Lymphoma	158	192	193	194	195	202	212	215	229	230
	237	240	244	247	259	300	301	321	337	415
	434	444	485	496	529	537	624	707	800	
Sarcoma	430	590	606	638	655	679	691	693	696	747
	752	760	778	821	986					
Other	136	246	255	376	421	565	616	617	652	655
	658	660	662	675	681	734	736	737	757	769
	777	800	807	825	855	857	864	868	870	870
	873	882	895	910	934	942	1015	1019		

Table 5.7 Mouse data (Hoel and Walburg, 1972). Age at death (days) of RFM male mice exposed to 300 rads of x-radiation at 5–6 weeks of age. The causes of death were thymic lymphoma, reticulum cell sarcoma and other. The upper group of 95 mice were kept in a conventional environment; the lower 82 in a germ-free environment.

is a model for independent failures that yields cause-specific hazard functions h_1 and h_2, so whatever the form of these functions, data of form (Y, I) cannot give evidence against independent risks. Dependence can only be inferred from data in which both Y_1 and Y_2 are observed for certain units, or from subject-matter considerations. This is important because interest often focuses on the effect of eliminating failures of one type, say type 2, in which case the survivor function is $\mathcal{F}(y, 0)$. As this is not a function of h_1 and h_2 it is inestimable unless assumptions, typically unverifiable ones, are made about the relation between the risks. Some statisticians therefore insist that the only valid inferences from competing risk data concern the h_i and quantities derived from them.

Example 5.31 (Mouse data) The data in Table 5.7 are from a experiment in which two groups of RFM strain male mice were exposed to 300 rad of radiation at age 5–6 weeks. The first group lived in a conventional laboratory environment, and the second group lived in a germ-free environment. After their deaths, a pathologist ascertained whether the death was due to one of two types of cancer or to other causes. One purpose of the experiment was to assess the effect of environment on different causes of death. As irradiation took place when the mice were aged between 35 and 42 days old, it might be better to take age since irradiation as the response, but its exact value is unknown.

The panels of Figure 5.11 shows the estimated cumulative hazard functions for death from lymphoma and from other causes. Mortality from the lymphoma arises early, and seems to depend little on the environment. Deaths from other causes arise earlier in the conventional environment than in the germ-free one. See also Example 10.38. ■

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

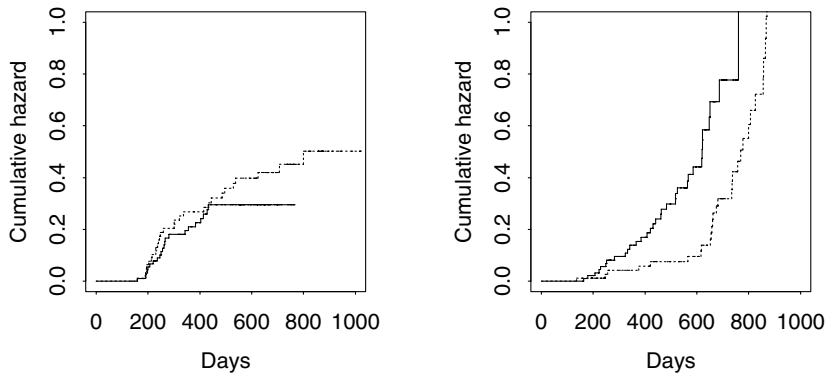

Figure 5.11 Estimated cumulative hazard functions for deaths from lymphoma (left) and other causes (right), in conventional (solid) and germ-free (dots) environments.

Frailty

The discussion above presupposes that all units have the same propensity to fail. In practice this is unrealistic — some cars are more reliable than others, some persons healthier than others, and so forth — and it may be important to build heterogeneity into models for survival. One reason for this is that allowing the failure rate to vary across units may greatly change the interpretation of the hazard function. It is tempting to view the population hazard function as a measure of how the risk for each unit changes as a function of time. For example, the fact that the divorce rate typically increases to a maximum a few years after marriage and thereafter decreases is sometimes interpreted as meaning that the typical marriage experiences increasing difficulties, but that if these are resolved there is eventually a more stable union. A unimodal divorce rate can be generated, however, by supposing that the hazard of failure increases with the duration of each marriage, but that the initial value of this hazard varies randomly from couple to couple. If this second interpretation is correct, then the population hazard function depends both on hazards for individual marriages and on variation across them, and reflects a selection process whereby the marriages most at risk tend to fail quickly, leaving those that were more stable to begin with. Thus the hazard rate is a more complicated quantity than it might seem at first sight.

One approach is to represent heterogeneity using the outcome of a positive random variable, Z, known as a *frailty*. We suppose that Z varies across units according to a density $f_Z(z)$, and that at time y the hazard function for a unit for whom $Z = z$ is $zh(y)$; thus the cumulative hazard to that time is $zH(y)$. Units whose z is large have high hazard functions and tend to fail sooner than those whose frailty is low. If known, the value of z could be incorporated into the analysis by modifying the likelihood, but we suppose it is unobserved, perhaps representing unobserveable genetic and environmental differences among units, and use it to model heterogeneity in the data.

As the survivor function for a unit with frailty z may be expressed as $\Pr(Y \geq y \mid Z = z) = \exp\{-zH(y)\}$, the survivor function for a unit taken randomly from the

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

population is

$$\begin{aligned}\Pr(Y \geq y) &= \int_0^\infty \Pr(Y \geq y \mid Z = z) f_Z(z)\, dz \\ &= \int_0^\infty \exp\{-zH(y)\}\, f_Z(z)\, dz \\ &= M\{-H(y)\},\end{aligned}$$

where M is the moment-generating function of Z. Thus the cumulative hazard function for the population is $-\log M\{-H(y)\}$. The densities of Z conditional on failure at y and conditional on survival at least to y,

$$f(z \mid Y = y) = \frac{z f_Z(z) \exp\{-zH(y)\}}{\int_0^\infty z f_Z(z) \exp\{-zH(y)\}\, dz},$$

$$f_Z(z \mid Y \geq y) = \frac{e^{-zH(y)} f_Z(z)}{\int_0^\infty \exp\{-zH(y)\}\, f_Z(z)\, dz}, \quad z > 0,$$

can be used to see how frailty depends on failure and on survival.

Example 5.32 (Logistic hazard) Let $\beta > 0$ and $H(y) = e^{\beta y} - 1$, so a unit with frailty z has hazard $z\beta e^{\beta y}$; this increases exponentially. Suppose also that Z has the gamma density with mean $\alpha\beta^{-1}/(1+\alpha)$ and shape parameter β^{-1}. Then $M(u) = \{1 - \alpha u/(1+\alpha)\}^{-1/\beta}$, and the population cumulative hazard function,

$$-\log M\{-H(y)\} = \frac{1}{\beta} \log\left(\frac{1+\alpha e^{\beta y}}{1+\alpha}\right),$$

is the same as that fitted to the data on old age in Example 5.29. Thus although each unit has a constant hazard, the effect of frailty is that the population hazard has an S-shaped logistic form, because of the selective effect of the early failure of the weakest units.

Simple calculations show that the density of frailties among those units failing at time y is gamma with mean $\alpha(1+\beta^{-1})/(1+\alpha e^{\beta y})$ and shape parameter $1+\beta^{-1}$, while that among those units who have not failed at time y is gamma with corresponding parameters $\alpha\beta^{-1}/(1+\alpha e^{\beta y})$ and β^{-1}. Both of these are decreasing in y, showing how the tendency for units with high frailties to fail first leads to survival of the fittest.

Information on unit hazard functions would be needed before such a model could be regarded as a serious explanation of the good fit of the logistic hazard for the data on old age. Absent such knowledge, the model is best regarded as suggesting a possible mechanism for the observed phenomenon, and as indicating the type of data needed for a more detailed investigation. ■

Evidently frailty has the potential to greatly complicate the analysis of population phenomena. It also complicates group comparisons (Problem 5.15).

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Exercises 5.4

1 Show that if there is no censoring, the product-limit estimator may be written $\widehat{\mathcal{F}}(y) = n^{-1}\#\{i : y_i > y\}$, and hence show that in this case $1 - \widehat{\mathcal{F}}(y)$ equals the empirical distribution function (2.3). Find Greenwood's formula, and comment.

If in doubt, think of failures of your car, fridge, computer, ...

2 Suggest physical phenomena that might give increasing, decreasing, and bathtub-shaped hazard functions. Sketch the corresponding survivor functions.

3 Use the relation $\mathcal{F}(y) = \exp\{-\int_0^y h(u)du\}$ between the survivor and hazard functions to find the survivor functions corresponding to the following hazards: (a) $h(y) = \lambda$; (b) $h(y) = \lambda y^\alpha$; (c) $h(y) = \alpha y^{\kappa-1}/(\beta + y^\kappa)$. In each case state what the distribution is.
Show that $\mathrm{E}\{1/h(Y)\} = \mathrm{E}(Y)$ and hence find the means in (a), (b), and (c).

4 The *mean excess life function* is defined as $e(y) = \mathrm{E}(Y - y \mid Y > y)$. Show that

$$e(y) = \mathcal{F}(y)^{-1} \int_y^\infty \mathcal{F}(u)\, du$$

and deduce that $e(y)$ satisfies the equation $e(y)Q'(y) + Q(y) = 0$ for a suitable $Q(y)$. Hence show that provided the underlying density is continuous,

$$\mathcal{F}(y) = \frac{e(0)}{e(y)} \exp\left\{-\int_0^y \frac{1}{e(u)}\, du\right\}.$$

As a check on this, find $e(y)$ and hence $\mathcal{F}(y)$ for the exponential density.
One approach to modelling survival is in terms of $e(y)$. For human lifetime data, let $e(y) = \gamma(1 - y/\theta)^\beta$, where θ is an upper endpoint and $\beta, \gamma > 0$. Find the corresponding survivor and hazard functions, and comment.

5 If $\mathcal{F}_1(y), \ldots, \mathcal{F}_k(y)$ are the survivor functions of independent positive random variables and $\beta_1, \ldots, \beta_k > 0$, show that $\prod \mathcal{F}_i(y)^{\beta_i}$ is also a survivor function, and find the corresponding hazard and cumulative hazard functions.
Suppose that $k = 2$ and the survivor functions are (i) log-logistic, (ii) log-normal and (iii) Weibull. Show that in the first two cases new models are obtained, but that in the third the parameters are not identifiable.

6 An empirical estimate of the survivor function $\mathcal{F}(y)$ when data $y_1, \ldots, y_n$ are not censored is given by $\widehat{\mathcal{F}}(y) = \#\{j : y_j > y\}/(n + 1)$. Suggest how plots of $\log\{-\log \widehat{\mathcal{F}}(y_j)\}$ against $\log y_j$ may be used to indicate if the data have Weibull or exponential distributions. Describe the corresponding plot for the Gumbel distribution function $F(y) = \exp[-\exp\{-(y - \eta)/\alpha\}]$.

7 Show that the log likelihood (5.26) may be expressed as

$$\ell(\theta) = \int_0^\infty \log h(y;\theta)\, dD(y) - \int_0^\infty R(y)\, dH(y;\theta),$$

where $D(y)$ is a step function with jumps of size one at the values of y that are failures and $R(y)$ is the number of units at risk of failure at time y. Establish that both integrals are over finite ranges. Such expressions are useful in a general treatment of likelihood inference for failure data.

5.5 Missing Data

5.5.1 Types of missingness

Missing observations arise in many applications, but particularly in data from living subjects, for example when frost kills a plant or the laboratory cat kills some experimental mice. They are common in data on humans, who may agree to take part in a

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

two-year study and then drop out after six months, or refuse to answer questions about their salaries or sex-lives. They may occur by accident or by design, for example when lifetimes are censored at the end of a survival study (Section 5.4).

The central problem they pose is obvious: little can be said about unknown data, even if the pattern of missingness suggests its cause and hence indicates to what extent remaining observations can be trusted and lost ones imputed. Loss of data will clearly increase uncertainty, but a more malign effect is that inferences from the data are sharply limited unless we are prepared to make assumptions that the data themselves cannot verify. Thus, if data are missing or might be missing it is essential to consider possible underlying mechanisms and their potential effect on inferences. The discussion below is intended to focus thought about these.

Suppose that our goal is inference for a parameter θ based on data that would ideally consist of n independent pairs (X, Y), but that some values of Y are missing, as shown by an indicator variable, I. Thus the data on an individual have form $(x, y, 1)$ or $(x, ?, 0)$. We suppose that although the missingness mechanism $\Pr(I = 0 \mid x, y)$ may depend on x and y, it does not involve θ. Then the likelihood contribution from an individual with complete data is the joint density of X, Y and I, which we write as

$$\Pr(I = 1 \mid x, y) f(y \mid x; \theta) f(x; \theta),$$

while if Y is unknown we use the marginal density of X and I,

$$\int \Pr(I = 0 \mid x, y) f(y \mid x; \theta) f(x; \theta)\, dy. \tag{5.30}$$

There are now three possibilities:

- data are *missing completely at random*, that is, $\Pr(I = 0 \mid x, y) = \Pr(I = 0)$ is independent both of x and y, and (5.30) reduces to $\Pr(I = 0) f(x; \theta)$;
- data are *missing at random*, that is, $\Pr(I = 0 \mid x, y) = \Pr(I = 0 \mid x)$ depends on x but not on y, and (5.30) equals $\Pr(I = 0 \mid x) f(x; \theta)$; and
- there is *non-ignorable non-response*, meaning that $\Pr(I = 0 \mid x, y)$ depends on y and possibly also on x.

In the first two of these, which are often grouped as *ignorable non-response*, I carries no information about θ and can be omitted for most likelihood inferences. To see why, suppose that we have n independent observations of form $(x_1, y_1, I_1), \ldots, (x_n, y_n, I_n)$, let $\mathcal{M}$ be the set of j for which y_j is unobserved, and suppose that data are missing at random. Then the likelihood is

$$\begin{aligned} L(\theta) &= \prod_{j \in \mathcal{M}} \Pr(I_j = 0 \mid x_j) f(x_j; \theta) \times \prod_{j \notin \mathcal{M}} \Pr(I_j = 1 \mid x_j) f(x_j, y_j; \theta) \\ &\propto \prod_{j \in \mathcal{M}} f(x_j; \theta) \times \prod_{j \notin \mathcal{M}} f(x_j, y_j; \theta), \end{aligned}$$

because the terms involving I_j do not depend on θ. Thus the missing data mechanism does not affect maximum likelihood estimates $\widehat{\theta}$, likelihood ratio statistics or the observed information $J(\widehat{\theta})$. It does affect the expected information, however, so standard errors for $\widehat{\theta}$ should be based on $J(\widehat{\theta})^{-1}$; see the discussion of likelihood

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

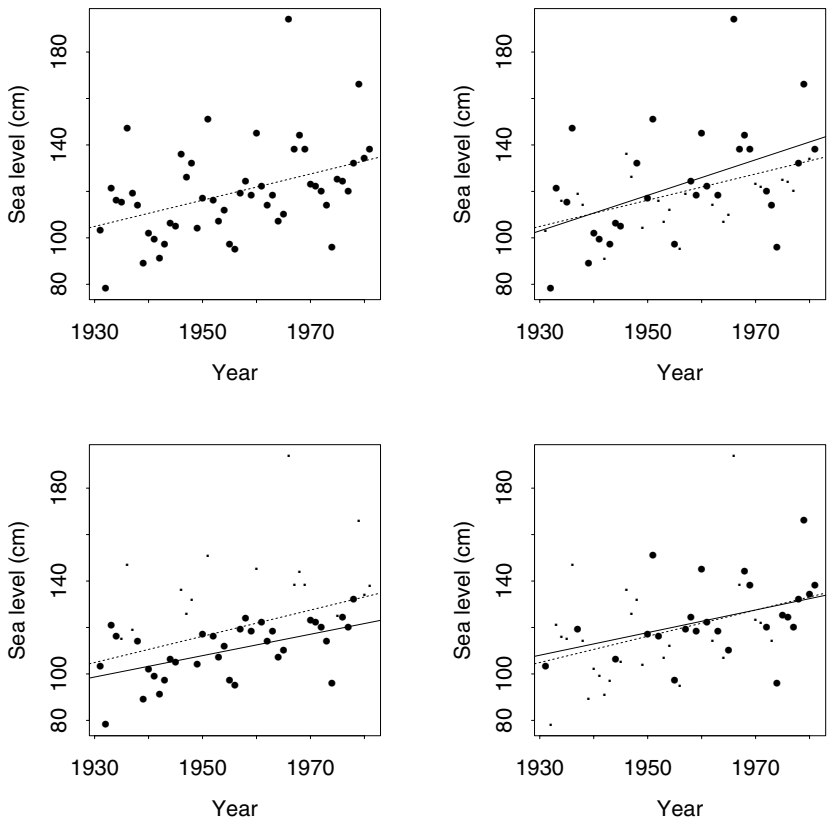

Figure 5.12 Missing data in straight-line regression for Venice sea-level data. Clockwise from top left: original data, data with values missing completely at random, data with values missing at random — missingness depends on x but not on y, and data with non-ignorable non-response — missingness depends on both x and y. Missing values are represented by a small dot. The dotted line is the fit from the full data, the solid lines those from the non-missing data.

inference in Section 5.4 and Problem 5.16. A similar argument applies if data are missing completely at random. If the non-response is non-ignorable, however, the density of I is no longer a constant of integration in (5.30). In that case, knowledge of the observed I_j is informative about θ, and likelihood inference is possible only if $\Pr(I = 0 \mid x, y)$ can be specified.

Example 5.33 (Venice sea level data) The upper left panel of Figure 5.12 shows the data of Example 5.1. Here x represents a year in the range 1931–1981; in the absence of sea level it contains no information about any trend. The annual maximum sea level y is taken to be a normal variable with mean $\beta_0 + \beta_1(x_j - \overline{x})$ and variance σ^2; hence $\theta = (\beta_0, \beta_1, \sigma^2)$ and the full data likelihood has form $f(y \mid x; \theta) f(x)$, of which $f(x)$ is ignored.

The upper right panel of Figure 5.12 shows the effect of data missing completely at random, while in the panel below the probability that a value is unobserved depends on x but not on y; the data are missing at random, with earlier observations missing more often than later ones. The lower left panel shows non-ignorable non-response, because the probability of missingness depends on y and on x; values of y that are larger than their means are more likely to be missing. Here the fitted line differs from those in the other panels due to bias induced by the missingness mechanism.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

	Truth	Average estimate (average standard error)			
		Full	MCAR	MAR	NIN
β_0	120	120 (2.79)	120 (4.02)	120 (4.73)	132 (3.67)
β_1	0.50	0.49 (0.19)	0.48 (0.28)	0.50 (0.32)	0.20 (0.25)

Table 5.8 Average estimates and standard errors for missing value simulation based on Venice data, for full dataset, with data missing completely at random (MCAR), missing at random (MAR) and with non-ignorable non-response (NIN). 1000 samples were taken. Standard errors for the averages for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are at most 0.16 and 0.01; those for their standard errors are at most 0.03 and 0.002.

To assess the extent of this bias, we generated 1000 samples from a model with parameters $\beta_0 = 120$, $\beta_1 = 0.5$ and $\sigma = 20$, close to the estimates for the Venice data and with the same covariate x. We then computed maximum likelihood estimates for the full data and for those observations that remain after applying the non-response mechanisms

$$\Pr(I = 1 \mid x, y) = \begin{cases} 0.5, \\ \Phi\{0.05(x - \overline{x})\}, \\ \Phi\left[0.05(x - \overline{x}) + \{y - \beta_0 - \beta_1(x - \overline{x})\}/\sigma\right], \end{cases}$$

to give data missing completely at random, missing at random, and with non-ignorable non-response. In each case roughly one-half of the observations are missing. Table 5.8 shows that although data loss increases the variability of the estimates, their means are unaffected, provided the probability of non-response does not depend on y. If the probability of missingness depends on the response, however, estimates based on the remaining data become entirely unreliable. ■

The message of this example is bleak: when there is non-ignorable non-response and a non-negligible proportion of the data is missing, the only possible rescue is to specify the missingness mechanism correctly. In practice it is typically hard to tell if missingness is ignorable or not, so fully reliable inference is largely out of reach. Sensitivity analysis to assess how heavily the conclusions depend on plausible mechanisms for non-response is then useful, and we now outline one approach to this.

Publication bias

Breakthroughs in medical science are regularly reported, offering hope of a new cure or suggesting that some enjoyable activity has dire consequences. It is unwise to take them all at face value, however, as some turn out to be spurious. One reason for this is the publication process to which they are subjected. Once a study is completed, an article describing it is typically submitted to a medical journal for peer review. If the study design and analysis are found to be satisfactory, a decision is taken whether the article should be published. This decision is likely to be positive if the study reports a significant result or if it involved a large number of patients, but will often be negative if no association is found — there is no 'significant finding' — particularly if the study is small and hence deemed unreliable. The end-result of this selection process is *publication bias*, whereby studies finding associations tend to be the ones published, even if in fact there is no effect. Recommendations to change medical practice are usually based not on a single study — unless it is huge, involving many thousands of patients — but on a *meta-analysis* that combines results from all published studies.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

As studies finding no effect are more likely to remain unpublished, however, wrong conclusions can be drawn.

For a simple model of this selection process, suppose that we wish to estimate a parameter μ that represents the effect of a treatment, subject to possible publication bias. A study based on n individuals produces an estimate $\widehat{\mu}$, normally distributed with mean μ and variance σ^2/n. The vagaries of the editorial process are represented by a variable Z, with the study published if Z is positive. We suppose that $\widehat{\mu}$ and Z are related by

$$\widehat{\mu} = \mu + \sigma n^{-1/2} U_1, \quad Z = \gamma_0 + \gamma_1 n^{1/2} + U_2,$$

with U_1 and U_2 standard normal variables with correlation $\rho \geq 0$. One interpretation of U_1 is as the standardized form $n^{1/2}(\widehat{\mu} - \mu)/\sigma$ of $\widehat{\mu}$, which is used to assess significance of the treatment effect. If $\rho > 0$ then publication becomes increasingly likely as U_1 increases, because Z is positively correlated with U_1. In terms of our previous discussion, Y and X correspond to $\widehat{\mu}$ and n, but now neither is observed if the study is unpublished.

The missingness indicator I equals one if $Z > 0$ and zero otherwise, so the marginal probability of publication is

$$\Pr(I = 1) = \Pr(Z > 0) = \Pr\left(U_2 > -\gamma_0 - \gamma_1 n^{1/2}\right) = \Phi\left(\gamma_0 + \gamma_1 n^{1/2}\right). \quad (5.31)$$

If $\gamma_1 > 0$ this increases with n: large studies are then more likely to be published, whatever their outcome. Conditional on the value of $\widehat{\mu}$, (3.21) implies that Z is normal with mean $\gamma_0 + \gamma_1 n^{1/2} + \rho n^{1/2}(\widehat{\mu} - \mu)/\sigma$ and variance $1 - \rho^2$. Hence the conditional probability of publication given $\widehat{\mu}$ is

$$\Pr(I = 1 \mid \widehat{\mu}) = \Pr(Z > 0 \mid \widehat{\mu}) = \Phi\left\{\frac{\gamma_0 + \gamma_1 n^{1/2} + \rho n^{1/2}(\widehat{\mu} - \mu)/\sigma}{(1 - \rho^2)^{1/2}}\right\}. \quad (5.32)$$

If $\rho > 0$, this is increasing in $\widehat{\mu}$: the probability that a study is published increases with the estimated treatment effect, at each study size n. Moreover, as $\widehat{\mu}$ appears in (5.32), non-response — non-publication of a study — is non-ignorable. If $\rho = 0$, (5.32) reduces to (5.31). Unpublished studies are then missing at random: the odds that a study is published depend on its size n but not on its outcome $\widehat{\mu}$.

Conditional on publication, the mean of $\widehat{\mu}$ is

$$\mathrm{E}(\widehat{\mu} \mid Z > 0) = \mu + \rho\sigma n^{-1/2} \zeta\left(\gamma_0 + \gamma_1 n^{1/2}\right), \quad (5.33)$$

where $\zeta(u) = \phi(u)/\Phi(u)$ is the ratio of the standard normal density and distribution functions. If $\gamma_1, \rho > 0$, then $\mathrm{E}(\widehat{\mu} \mid Z > 0) > \mu$, so the mean of a published $\widehat{\mu}$ is always larger than μ, but by an amount that decreases with n. For small γ_1, Taylor expansion gives

$$\mathrm{E}(\widehat{\mu} \mid Z > 0) \doteq \mu + \rho\sigma\gamma_1 \zeta'(\gamma_0) + \rho\sigma\zeta(\gamma_0) n^{-1/2},$$

so the conditional mean of $\widehat{\mu}$ in published studies is roughly linear in $n^{-1/2}$. As just three parameters — intercept, slope and variance — can be estimated from a linear fit, simultaneous estimation of μ, ρ, σ^2, γ_0, and γ_1 is infeasible. In order to assess

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Trial	Magnesium r/m	Control r/m	n	$\widehat{\mu}$	$(v/n)^{1/2}$
1	1/25	3/23	48	1.18	1.05
2	1/40	2/36	76	0.80	0.83
3	2/48	2/46	94	0.04	0.75
4	1/50	9/53	103	2.14	0.72
5	4/56	14/56	112	1.25	0.69
6	3/66	6/66	132	0.69	0.63
7	2/92	7/93	185	1.24	0.53
8	27/135	43/135	270	0.47	0.44
9	10/160	8/156	316	−0.20	0.41
10	90/1159	118/1157	2316	0.27	0.15
Meta-analysis			3652	0.41	0.11
ISIS-4	2216/29011	2103/29039	58050	−0.05	0.03

Table 5.9 Data from 11 clinical trials to compare magnesium treatment for heart attacks with control, with n patients randomly allocated to treatment and control; there are r deaths out of m patients in each group (Copas, 1999). The estimated log treatment effect $\widehat{\mu}$ will be positive if treatment is effective; $(v/n)^{1/2}$ is its standard error. The huge ISIS-4 trial is not included in the meta-analysis.

the impact of selection in the following example, we fix γ_0 and γ_1 to give plausible probabilities of publication for small and large samples, and consider inference for $\theta = (\mu, \rho, \sigma)$.

Now suppose that we wish to estimate μ based on k independent estimates $\widehat{\mu}_1, \ldots, \widehat{\mu}_k$ from published studies of sizes $n_1, \ldots, n_k$. As $\widehat{\mu}_j$ is observed only conditional on its publication, the likelihood contribution from study j is

$$f(\widehat{\mu}_j \mid Z_j > 0; \theta) = \frac{f(\widehat{\mu}_j; \theta)\Pr(Z_j > 0 \mid \widehat{\mu}_j; \theta)}{\Pr(Z_j > 0)}.$$

The marginal density of $\widehat{\mu}_j$ is normal with mean μ and variance σ^2/n_j, and on recalling (5.31) and (5.32), we see that the overall log likelihood is

$$\ell(\mu, \rho, \sigma^2) \equiv -\sum_{j=1}^{k} \left\{ \frac{1}{2}\log\sigma^2 + \frac{n_j}{2\sigma^2}(\widehat{\mu}_j - \mu)^2 + \log\Phi(a_j) - \log\Phi(b_j) \right\}, \tag{5.34}$$

where $a_j = \gamma_0 + \gamma_1 n_j^{1/2}$ and $b_j = (1-\rho^2)^{-1/2}\{a_j + \rho n_j^{1/2}(\widehat{\mu}_j - \mu)/\sigma\}$.

The simplest meta-analysis ignores the possibility of selection bias and amounts to setting $\rho = 0$, presuming the publication of a study to be unrelated to its result. If this is so, then $a_j = b_j$ and the log likelihood is easily maximized, the maximum likelihood estimate of μ being the weighted average

$$\frac{\sum n_j \widehat{\mu}_j}{\sum n_j}. \tag{5.35}$$

When $\rho = 0$, this estimator is normal with mean μ and variance $\sigma^2/\sum n_j$. If in fact $\rho > 0$, then (5.33) implies that $\widehat{\mu}_0$ will tend to exceed μ; the treatment effect will tend to be overstated by the published data.

Example 5.34 (Magnesium data) Table 5.9 shows data from clinical trials on the use of intraveneous magnesium to treat patients with suspected acute myocardial

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

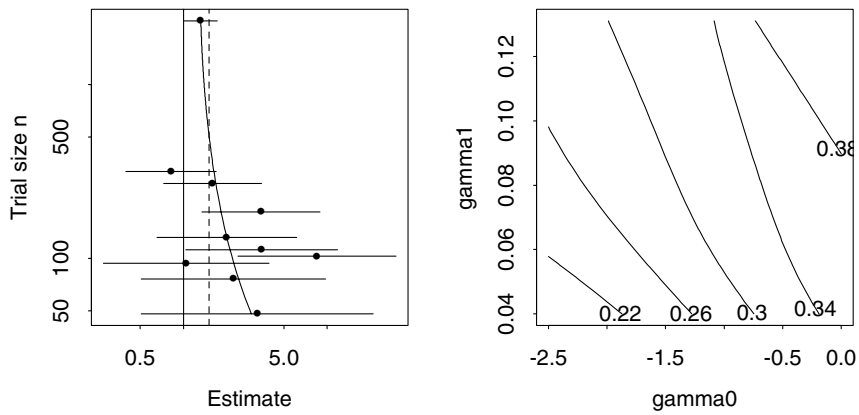

Figure 5.13 Likelihood analysis of magnesium data. Left: funnel plot showing variation of $\widehat{\mu}$ with trial size n, with 95% confidence interval for μ based on each trial. The vertical dotted line is the combined estimate of μ from the ten small trials, ignoring the possibility of publication bias; the vertical solid line shows no treatment effect. The solid line is the estimated conditional mean (5.33). Right: contours of $\widehat{\mu}$ as a function of γ_0 and γ_1.

Myocardial infarction is the medical term for heart attack — death of part of the heart muscle because of lack of oxygen and other nutrients.

infarction. For each trial, we consider the difference in log proportion of deaths between control and treated groups, the estimated treatment effect $\widehat{\mu} = \log(r_2/m_2) - \log(r_1/m_1)$. Now $m_1 \doteq m_2$ for each trial and the proportion of deaths is small, so the delta method suggests that an approximate variance for $\widehat{\mu}$ is $4/(\widehat{\lambda}n)$, where $\widehat{\lambda} = 0.097$ is the death rate estimated from all the trials and $n = m_1 + m_2$ is the size of each trial. The combined sample is large enough to treat $\widehat{\lambda}$ and hence $\sigma^2 = 4/\widehat{\lambda}$ as constant. Although the estimated treatment effects $\widehat{\mu}$ from the ten small trials are individually inconclusive, the meta-analysis estimate (5.35) is 0.41 with standard error 0.11; this gives an estimated reduction in the probability of death by a factor $\exp(0.41) = 1.51$ with 0.95 confidence interval (1.22,1.86). A similar published meta-analysis concluded that the magnesium treatment was 'effective, safe and simple'.

For a more skeptical view, consider the *funnel plot* of n and $\exp(\widehat{\mu})$ in the left panel of Figure 5.13; note the logarithmic axes. Symmetry about the overall weighted average (5.35) would show lack of publication bias, but the visible asymmetry suggests that small studies tend to be published only if $\widehat{\mu}$ is sufficiently positive.

The right panel shows how the maximum likelihood estimate of μ from (5.34) depends on γ_0 and γ_1. The contours are very roughly parallel with slope -0.05, suggesting that the maximum likelihood estimate varies mainly as a function of $\gamma_0 + 400^{1/2}\gamma_1$, or equivalently the probability $\Phi(\gamma_0 + 400^{1/2}\gamma_1)$ that a study of size $n = 400$ is published. For example, if the selection probabilities are 0.9 and 0.1 for the largest and smallest studies in Table 5.9, then this probability is 0.32, $\widehat{\rho} = 0.5$ and the estimated treatment effect is 0.27 with standard error 0.12 from observed information. This estimate is substantially less than the value 0.41 obtained when $\rho = 0$, and the significance of the estimated treatment effect is much reduced. The estimated conditional mean (5.33) in the left panel shows how the selection due to having $\rho > 0$ affects the mean of published studies.

The sensitivity of the estimated effect to potential publication bias suggests that treatment policy conclusions cannot be based on Table 5.9. Indeed, a subsequent much larger trial — ISIS-4 — found no evidence that magnesium is effective. ■

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Publication bias is an example of selection bias, where the mechanism underlying the choice of data introduces an uncontrolled bias into the sample. This is endemic in observational studies, for example in epidemiology and the social sciences, and it can greatly weaken what conclusions may be drawn.

5.5.2 EM algorithm

The fitting of certain models is simplified by treating the observed data as an incomplete version of an ideal dataset whose analysis would have been easy. The key idea is to estimate the log likelihood contribution from the missing data by its conditional value given the observed data. This yields a very general and widely used *estimation-maximization* or EM algorithm for maximum likelihood estimation.

Let Y denote the observed data and U the unobserved variables. Our goal is to use the observed value y of Y for inference on a parameter θ, in models where we cannot easily calculate the density

$$f(y;\theta) = \int f(y \mid u;\theta) f(u;\theta)\, du$$

and hence cannot readily compute the likelihood for θ based only on y. We write the *complete-data log likelihood* based on both y and the value u of U as

$$\log f(y,u;\theta) = \log f(y;\theta) + \log f(u \mid y;\theta), \tag{5.36}$$

where the first term on the right is the *observed-data log likelihood* $\ell(\theta)$. As the value of U is unobserved, the best we can do is to remove it by taking expectation of (5.36) with respect to the conditional density $f(u \mid y;\theta')$ of U given that $Y = y$; for reasons that will become apparent we use θ' rather than θ for this expectation. This yields

$$\mathrm{E}\{\log f(Y,U;\theta) \mid Y = y;\theta'\} = \ell(\theta) + \mathrm{E}\{\log f(U \mid Y;\theta) \mid Y = y;\theta'\}, \tag{5.37}$$

which we express as

$$Q(\theta;\theta') = \ell(\theta) + C(\theta;\theta'). \tag{5.38}$$

We now fix θ' and treat $Q(\theta;\theta')$ and $C(\theta;\theta')$ as functions of θ. If the conditional distribution of U given $Y = y$ is non-degenerate and no two values of θ give the same model, then the argument at (4.31) applied to $f(y \mid u;\theta)$ shows that $C(\theta';\theta') \geq C(\theta;\theta')$, with equality only when $\theta = \theta'$. Hence

$$Q(\theta;\theta') \geq Q(\theta';\theta') \text{ implies } \ell(\theta) - \ell(\theta') \geq C(\theta';\theta') - C(\theta;\theta') \geq 0. \tag{5.39}$$

Moreover under mild smoothness conditions, $C(\theta;\theta')$ has a stationary point at $\theta = \theta'$. Hence if $Q(\theta;\theta')$ is stationary at $\theta = \theta'$, so too is $\ell(\theta)$.

This leads to the *EM algorithm*: starting from an initial value θ' of θ,

1. compute $Q(\theta;\theta') = \mathrm{E}\left\{\log f(Y,U;\theta) \mid Y = y;\theta'\right\}$; then
2. with θ' fixed, maximize $Q(\theta;\theta')$ over θ, giving $\theta^\dagger$, say; and
3. check if the algorithm has converged, using $\ell(\theta^\dagger) - \ell(\theta')$ if available, or $|\theta^\dagger - \theta'|$, or both. If not, set $\theta' = \theta^\dagger$ and go to 1.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Steps 1 and 2 are the expectation (E) and maximization (M) steps of the algorithm. As the M-step ensures that $Q(\theta^\dagger;\theta') \geq Q(\theta';\theta')$, we see from (5.39) that $\ell(\theta^\dagger) \geq \ell(\theta')$: the log likelihood never decreases. Moreover, if $\ell(\theta)$ has just one stationary point, and if $Q(\theta;\theta')$ eventually reaches a stationary value at $\widehat{\theta}$, then $\widehat{\theta}$ must maximize $\ell(\theta)$. If $\ell(\theta)$ has more than one stationary point the algorithm may converge to a local maximum of the log likelihood or to a turning point. As the EM algorithm never decreases the log likelihood it is more stable than Newton–Raphson-type algorithms, which do not have this desirable property.

As one might expect, the convergence rate of the algorithm depends on the amount of missing information. If knowledge of Y tells us little about U, then $Q(\theta;\theta')$ and $\ell(\theta)$ will be very different and the algorithm slow. This may be quantified by differentiating (5.36) and taking expectations with respect to the conditional distribution of U given Y, to give

$$-\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}} = \mathrm{E}\left\{ -\frac{\partial^2 \log f(y,U;\theta)}{\partial\theta\partial\theta^{\mathrm{T}}} \,\middle|\, Y=y;\theta \right\} - \mathrm{E}\left\{ -\frac{\partial^2 \log f(U \mid y;\theta)}{\partial\theta\partial\theta^{\mathrm{T}}} \,\middle|\, Y=y;\theta \right\},$$

or $J(\theta) = I_c(\theta;y) - I_m(\theta;y)$, interpreted as meaning that the observed information equals the complete-data information minus the missing information; this is sometimes called the *missing information principle*. If U is determined by Y, then the conditional density $f(u \mid y;\theta)$ is degenerate and under mild conditions the missing information will be zero. It turns out that the rate of convergence of the algorithm equals the largest eigenvalue of the matrix $I_c(\theta;y)^{-1} I_m(\theta;y)$; values of this eigenvalue close to one imply slow convergence and occur if the missing information is a high proportion of the total.

When the EM algorithm is slow it may be worth trying to accelerate it by replacing the M-step with direct maximization, assuming of course that $\ell(\theta)$ is unavailable. It turns out that (Exercise 5.5.5)

$$\frac{\partial \ell(\theta)}{\partial\theta} = \left.\frac{\partial Q(\theta;\theta')}{\partial\theta}\right|_{\theta'=\theta}, \quad \frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}} = \left\{ \frac{\partial^2 Q(\theta;\theta')}{\partial\theta\partial\theta^{\mathrm{T}}} + \frac{\partial^2 Q(\theta;\theta')}{\partial\theta\partial\theta'^{\mathrm{T}}} \right\}\Bigg|_{\theta'=\theta}. \tag{5.40}$$

Thus even if $\ell(\theta)$ is inaccessible, its derivatives may be obtained from those of $Q(\theta;\theta')$ and used in a generic maximization algorithm. The second of these formulae also provides standard errors for the maximum likelihood estimate $\widehat{\theta}$ when $Q(\theta;\theta')$ is known but $\ell(\theta)$ is not.

Example 5.35 (Negative binomial model) For a toy example, suppose that conditional on $U = u$, Y is a Poisson variable with mean u, and that U is gamma with mean θ and variance θ^2/ν. Inference is required for θ with the shape parameter $\nu > 0$ supposed known. Here (5.36) equals

$$y \log u - u - \log y! + \nu \log \nu - \nu \log \theta + (\nu - 1)\log u - \nu u/\theta - \log \Gamma(\nu),$$

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

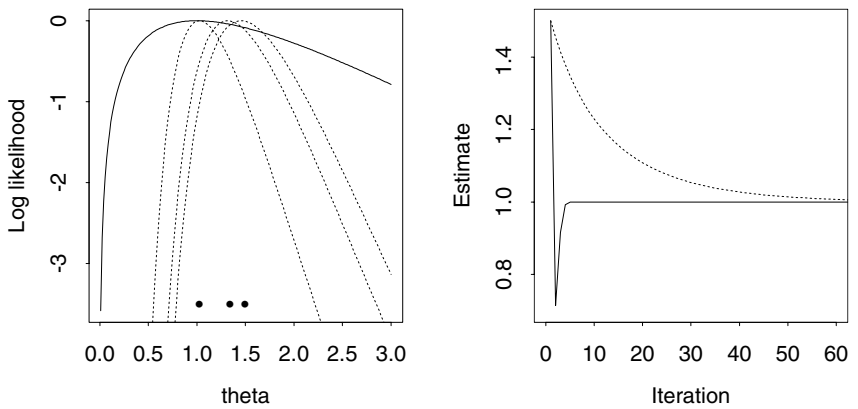

Figure 5.14 EM algorithm for negative binomial example. Left panel: observed-data log likelihood $\ell(\theta)$ (solid) and functions $Q(\theta;\theta')$ for $\theta' = 1.5$, 1.347 and 1.028 (dots, from right). The blobs show the values of θ that maximize these functions, which correspond to the first, fifth and fortieth iterations of the EM algorithm. Right: convergence of EM algorithm (dots) and Newton–Raphson algorithm (solid). The panel shows how successive EM iterations update θ' and $\widehat{\theta}$. Notice that the EM iterates always increase $\ell(\theta)$, while the Newton–Raphson steps do not.

and hence (5.37) equals

$$Q(\theta;\theta') = (y+\nu-1)\mathrm{E}(\log U \mid Y=y;\theta') - (1+\nu/\theta)\mathrm{E}(U \mid Y=y;\theta') - \nu\log\theta$$

plus terms that depend neither on U nor on θ.

The E-step, computation of $Q(\theta;\theta')$, involves two expectations, but fortunately $\mathrm{E}(\log U \mid Y=y;\theta')$ does not appear in terms that involve θ and so is not required. To compute $\mathrm{E}(U \mid Y=y;\theta')$, note that Y and U have joint density

$$f(y \mid u)f(u;\theta) = \frac{u^y}{y!}e^{-u} \times \frac{\nu^\nu u^{\nu-1}}{\theta^\nu\Gamma(\nu)}e^{-\nu u/\theta}, \quad y=0,1,\ldots, \quad u>0, \quad \theta>0,$$

so the marginal density of Y is

$$f(y;\theta) = \int_0^\infty f(y \mid u)f(u;\theta,\nu)\,du = \frac{\Gamma(y+\nu)\nu^\nu}{\Gamma(\nu)y!}\frac{\theta^y}{(\theta+\nu)^{y+\nu}}, \quad y=0,1,\ldots$$

Hence the conditional density $f(u \mid y;\theta')$ is gamma with shape parameter $y+\nu$ and mean $\mathrm{E}(U \mid Y=y;\theta') = (y+\nu)/(1+\nu/\theta')$, and we can take

$$Q(\theta;\theta') \equiv -(1+\nu/\theta)(y+\nu)/(1+\nu/\theta') - \nu\log\theta,$$

where we have ignored terms independent of both θ and θ'.

The M-step involves maximization of $Q(\theta;\theta')$ over θ for fixed θ', so we differentiate with respect to θ and find that the maximizing value is

$$\theta^\dagger = \theta'(y+\nu)/(\theta'+\nu). \tag{5.41}$$

In this example, therefore, the EM algorithm boils down to choosing an initial θ', updating it to $\theta^\dagger$ using (5.41), setting $\theta' = \theta^\dagger$ and iterating to convergence.

The log likelihood based only on the observed data y is

$$\ell(\theta) = \log f(y;\theta) \equiv y\log\theta - (y+\nu)\log(\theta+\nu), \quad \theta>0.$$

This is shown in the left panel of Figure 5.14 for $y=1$ and $\nu=15$. The panel also shows the functions $Q(\theta;\theta')$ on the first, fifth and fourtieth iterations starting at $\theta' = 1.5$, which gives the sequence $\theta' = 1.5, 1.45, 1.41, \ldots$. The functions $Q(\theta;\theta')$ are

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

much more concentrated than is $\ell(\theta)$, showing that the amount of missing information is large. The difference in curvature corresponds to the information lost through not observing U.

Here the unmodified EM algorithm converges slowly. The right panel of Figure 5.14 illustrates this, as successive values of $\theta^\dagger$ descend gently towards the limiting value $\theta = 1$: convergence has still not been achieved after 100 iterations, at which point $\theta^\dagger = 1.00056$. The ratio of missing to complete-data information, 15/16, indicates slow convergence. The Newton–Raphson algorithm (4.25) using the derivatives (5.40) converges much faster, with $\widehat{\theta} = 1$ to seven decimal places after only five iterations, so here it pays handsomely to use the derivative information in (5.40). ■

Example 5.36 (Mixture density) Mixture models arise when an observation Y is taken from a population composed of distinct subpopulations, but it is unknown from which of these Y is taken. If the number p of subpopulations is finite, Y has a p-component mixture density

$$f(y;\theta) = \sum_{r=1}^{p} \pi_r f_r(y;\theta), \quad 0 \le \pi_r \le 1, \sum_{r=1}^{p} \pi_r = 1,$$

where π_r is the probability that Y comes from the rth subpopulation and $f_r(y;\theta)$ is its density conditional on this event. An indicator U of the subpopulation from which Y arises takes values $1, \ldots, p$ with probabilities $\pi_1, \ldots, \pi_p$. In many applications the components have a physical meaning, but sometimes a mixture is used simply as a flexible class of densities. For simplicity of notation below, let θ contain all unknown parameters including the π_r.

If the value u of U were known, the likelihood contribution from (y, u) would be $\prod_r \{f_r(y;\theta)\pi_r\}^{I(u=r)}$, giving contribution

$$\log f(y, u;\theta) = \sum_{r=1}^{p} I(u = r)\{\log \pi_r + \log f_r(y;\theta)\}$$

to the complete-data log likelihood. In order to apply the EM algorithm we must compute the expectation of $\log f(y, u;\theta)$ over the conditional distribution

$$\Pr(U = r \mid Y = y;\theta') = \frac{\pi_r' f_r(y;\theta')}{\sum_{s=1}^{p} \pi_s' f_s(y;\theta')}, \quad r = 1, \ldots, p. \tag{5.42}$$

This probability can be regarded as the weight attributable to component r if y has been observed; for compactness below we denote it by $w_r(y;\theta')$. The expected value of $I(U = r)$ with respect to (5.42) is $w_r(y;\theta')$, so the expected value of the log likelihood based on a random sample $(y_1, u_1), \ldots, (y_n, u_n)$ is

$$\begin{aligned} Q(\theta;\theta') &= \sum_{j=1}^{n}\sum_{r=1}^{p} w_r(y_j;\theta')\{\log \pi_r + \log f_r(y_j;\theta)\} \\ &= \sum_{r=1}^{p}\left\{\sum_{j=1}^{n} w_r(y_j;\theta')\right\}\log \pi_r + \sum_{r=1}^{p}\sum_{j=1}^{n} w_r(y_j;\theta')\log f_r(y_j;\theta). \end{aligned}$$

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

9172	9350	9483	9558	9775	10227	10406	16084	16170	18419
18552	18600	18927	19052	19070	19330	19343	19349	19440	19473
19529	19541	19547	19663	19846	19856	19863	19914	19918	19973
19989	20166	20175	20179	20196	20215	20221	20415	20629	20795
20821	20846	20875	20986	21137	21492	21701	21814	21921	21960
22185	22209	22242	22249	22314	22374	22495	22746	22747	22888
22914	23206	23241	23263	23484	23538	23542	23666	23706	23711
24129	24285	24289	24366	24717	24990	25633	26960	26995	32065
32789	34279								

Table 5.10 Velocities (km/second) of 82 galaxies in a survey of the Corona Borealis region (Roeder, 1990). The error is thought to be less than 50 km/second.

The M step of the algorithm entails maximizing $Q(\theta;\theta')$ over θ for fixed θ'. As the π_r do not usually appear in the component density f_r, the maximizing values $\pi_r^\dagger$ are obtained from the first term of Q, which corresponds to a multinomial log likelihood; see (4.45). Thus $\pi_r^\dagger = n^{-1}\sum_j w_r(y_j;\theta')$, the average weight for component r.

Estimates of the parameters of the f_r are obtained from the weighted log likelihoods that form the second term of $Q(\theta;\theta')$. For example, if f_r is normal with mean μ_r and variance σ_r^2, simple calculations give the weighted estimates

$$\mu_r^\dagger = \frac{\sum_{j=1}^n w_r(y_j;\theta')y_j}{\sum_{j=1}^n w_r(y_j;\theta')} \quad \sigma_r^{2\dagger} = \frac{\sum_{j=1}^n w_r(y_j;\theta')(y_j-\mu_r^\dagger)^2}{\sum_{j=1}^n w_r(y_j;\theta')}, \quad r = 1,\ldots,p.$$

Given initial values of $(\pi_r, \mu_r, \sigma_r^2) \equiv \theta'$, the EM algorithm simply involves computing the weights $w_r(y_j;\theta')$ for these initial values, updating to obtain $(\pi_r^\dagger, \mu_r^\dagger, \sigma_r^{2\dagger}) \equiv \theta^\dagger$, and checking convergence using the log likelihood, $|\theta^\dagger - \theta'|$, or both. If convergence is not yet attained, θ' is replaced by $\theta^\dagger$ and the cycle repeated.

We illustrate these calculations using the data in Table 5.10, which gives the velocities at which 82 galaxies in the Corona Borealis region are moving away from our own galaxy. It is thought that after the Big Bang the universe expanded very fast, and that as it did so galaxies formed because of the local attraction of matter. Owing to the action of gravity they tend to cluster together, but there seem also to be 'superclusters' of galaxies surrounded by voids. If galaxies are indeed super-clustered the distribution of their velocities estimated from the red-shift in their light-spectra would be multimodal, and unimodal otherwise. The data given are from sections of the northern sky carefully sampled to settle whether there are superclusters.

Cursory examination of the data strongly suggests clustering. In order to estimate the number of clusters we fit mixtures of normal densities by the EM algorithm with initial values chosen by eye. The maximized log likelihood for $p = 2$ is -220.19, found after 26 iterations. In fact this is the highest of several local maxima; the global maximum of $+\infty$ is found by centering one component of the mixture at any of the y_j and letting the corresponding $\sigma_r^2 \to \infty$; see Example 4.42. Only the local maxima yield sensible fits, the best of which is found using randomly chosen initial values. The number of iterations needed depends on these and on the number of components, but is typically less than 40. This procedure gives maximized log likelihoods -240.42, -203.48, -202.52 and -192.42 for fits with $p = 1$, 3, 4 and 5. The latter gives a single component to the two observations around 16,000 and so does not seem very

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

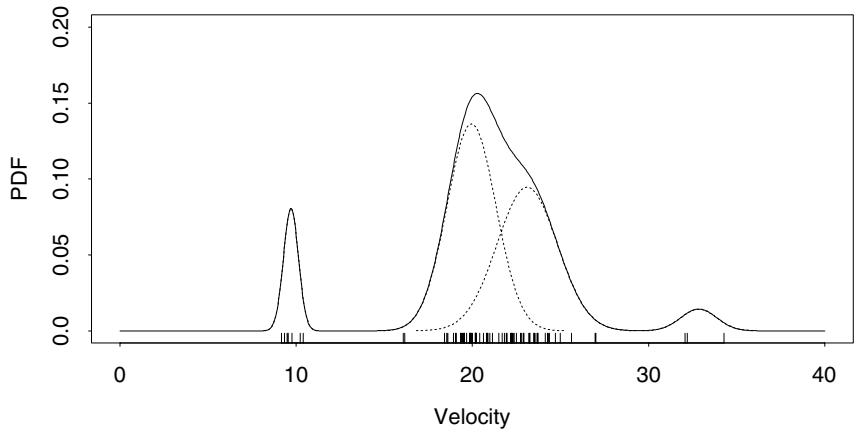

Figure 5.15 Fit of a 4-component mixture of normal densities to the data in Table 5.10 (10^3 km/second). Individual components $\widehat{\pi}_r f_r(y;\widehat{\theta})$ are shown by dotted lines.

sensible. Standard likelihood asymptotics do not apply here, but evidently there is little difference between the 3- and 4-component fits, the second of which is shown in Figure 5.15. Both fits have three modes, and the evidence for clustering is very strong.

An alternative is to apply a Newton–Raphson algorithm directly to the log likelihood $\ell(\theta)$ based on the mixture density, but if this is to be reliable the model must be reparametrized so that the parameter space is unconstrained, using $\log \sigma_r^2$ and expressing $\pi_1, \ldots, \pi_p$ in terms of $\theta_1, \ldots, \theta_{p-1}$ of Example 5.12. As mentioned in Example 4.42, the effect of the spikes in $\ell(\theta)$ can be reduced by replacing $f_r(y;\theta)$ by $F_r(y+h;\theta) - F_r(y-h;\theta)$, where h is the degree of rounding of the data, here 50 km/second. ■

Exponential family models

The EM algorithm has a particularly simple form when the complete-data log likelihood stems from an exponential family, giving

$$\log f(y, u; \theta) = s(y, u)^{\mathrm{T}}\theta - \kappa(\theta) + c(y, u).$$

The expected value of this is needed with respect to the conditional density $f(u \mid y; \theta')$. Evidently the final term will not depend on θ and can be ignored, so the M-step will involve maximizing

$$Q(\theta; \theta') = \mathrm{E}\{s(y, U)^{\mathrm{T}}\theta \mid Y = y; \theta'\} - \kappa(\theta),$$

or equivalently solving for θ the equation

$$\mathrm{E}\{s(y, U) \mid Y = y; \theta'\} = \frac{d\kappa(\theta)}{d\theta}.$$

The likelihood equation for θ based on the complete data would be $s(y, u) = d\kappa(\theta)/d\theta$, so the EM algorithm simply involves replacing $s(y, u)$ by its conditional expectation $\mathrm{E}\{s(y, U) \mid Y = y; \theta'\}$ and solving the likelihood equation. Thus a routine to fit the complete-data model can readily be adapted for missing data if the conditional expectations are available.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Example 5.37 (Positron emission tomography) Positron emission tomography is performed by introducing a radioactive tracer into an animal or human subject. Radioactive emissions are then used to assess levels of metabolic activity and blood flow in organs of interest. Positrons emitted by the tracer annihilate with nearby electrons, giving pairs of photons that fly off in opposite directions. Some of these are counted by bands of gamma detectors placed around the subject's body, but others miss the detectors. The detected counts are used to form an image of the level of metabolic activity in the organs based on the estimated spatial concentration of isotope.

Pixels and voxels are picture and volume elements, in 2 and 3 dimensions respectively.

For a statistical model, the region of interest is divided into n pixels or voxels and it is assumed that the number of emissions U_{ij} from the jth pixel detected at the ith detector is a Poisson variable with mean $p_{ij}\lambda_j$; here λ_j is the intensity of emissions from that pixel and p_{ij} the probability that a single emission is detected at the ith detector. The p_{ij} depend on the geometry of the detection system, the isotope and other factors, but can be taken to be known. The U_{ij} are unknown but can plausibly be assumed independent. The counts Y_i at the d detectors are observed and have independent Poisson distributions with means $\sum_{j=1}^{n} p_{ij}\lambda_j$.

The complete-data log likelihood,

$$\sum_{i=1}^{d}\sum_{j=1}^{n}\{u_{ij}\log(p_{ij}\lambda_j) - p_{ij}\lambda_j\},$$

is an exponential family in which the maximum likelihood estimates of the unknown λ_j have the simple form $\widehat{\lambda}_j = \sum_i u_{ij} / \sum_i p_{ij}$. The E-step requires only the conditional expectations $\mathrm{E}(U_{ij} \mid Y; \lambda')$. As $Y_i = U_{i1} + \cdots + U_{in}$, the conditional density of U_{ij} given $Y_i = y_i$ is binomial with denominator y_i and probability $p_{ij}\lambda'_j / \sum_h p_{ih}\lambda'_h$. Thus the M-step yields

$$\lambda_j^{\dagger} = \frac{\sum_{i=1}^{d}\mathrm{E}(U_{ij} \mid Y_j = y_j; \lambda')}{\sum_{i=1}^{d} p_{ij}} = \frac{\sum_{i=1}^{d} y_j p_{ij}\lambda'_j / \sum_{h=1}^{n} p_{ih}\lambda'_h}{\sum_{i=1}^{d} p_{ij}}$$
$$= \lambda'_j \frac{1}{\sum_{i=1}^{d} p_{ij}} \sum_{i=1}^{d} \frac{y_i p_{ij}}{\sum_{h=1}^{n}\lambda'_h p_{ih}}, \quad j = 1, \ldots, n.$$

The algorithm converges to a unique global maximum of the observed-data log likelihood provided that $d > n$, with the positivity constraints on the λ_j satisfied at each step.

Though simple, this algorithm has the undesirable property that the resulting images are too rough if it is iterated to full convergence. The difficulty is that although we would anticipate that adjacent pixels would be similar, the model places no constraint on the λ_j and so the final image is too close to the data. Some modification is required, such as adding a smoothing step to the algorithm or introducing a roughness penalty (Section 10.7.2). ■

The EM algorithm is particularly attractive in exponential family problems, but is used much more widely. In more general situations both E- and M-steps may

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

be complicated, and it often pays to break them into smaller components, perhaps involving Monte Carlo simulation to compute the conditional expectations required for the E-step. Discussion of this here would take us too far afield, but some of the recent research devoted to this is mentioned in the bibliographic notes.

Exercises 5.5

1 Data are *observed at random* if $\Pr(I = 0 \mid x, y) = \Pr(I = 0 \mid y)$, where I is the indicator that y is missing. Show that if data are observed at random and missing data are missing at random, then data are missing completely at random.

2 Show that Bayesian inference for θ is unaffected by the model for non-response if data are missing completely at random or missing at random, but not if there is non-ignorable non-response. What happens when $\Pr(I \mid x, y)$ depends on θ?

3 In Example 5.33, suppose that y is normal with mean $\beta_0 + \beta_1 x$ and variance σ^2, and that it is missing with probability $\Phi(a + by + cx)$, where a, b and c are unknown. Use (3.25) to find the likelihood contributions from pairs (x, y) and $(x, ?)$, and discuss whether the parameters are estimable.

4 When $\rho = 0$, show that (5.35) is the maximum likelihood estimate of μ and find its variance.

5 Use the fact that $\int f(u \mid y; \theta)\, du = 1$ for all y and θ to show that

$$0 = \mathrm{E}\left\{ \frac{\partial \log f(U \mid Y; \theta)}{\partial \theta} \,\middle|\, Y = y; \theta \right\},$$
$$0 = \mathrm{E}\left\{ \frac{\partial^2 \log f(U \mid Y; \theta)}{\partial \theta \partial \theta^{\mathrm{T}}} + \frac{\partial \log f(U \mid Y; \theta)}{\partial \theta} \frac{\partial \log f(U \mid Y; \theta)}{\partial \theta^{\mathrm{T}}} \,\middle|\, Y = y; \theta \right\}.$$

Now use (5.38) to establish (5.40).
Check this in the special case of Example 5.35, and hence give the Newton–Raphson step for maximization of the observed-data log likelihood, even though $\ell(\theta)$ itself is unknown. Write a program to compare the convergence of the EM and Newton–Raphson algorithms in that example.
(Oakes, 1999)

6 Check the forms of $\pi_r^\dagger$, $\mu_r^\dagger$ and $\sigma_r^{2\dagger}$ in Example 5.36, and verify that they respect the constraints $\sigma_r^2 > 0$, $0 \le \pi_r \le 1$ and $\sum \pi_r = 1$ on the parameter values.

7 Check the details of Example 5.37.

8 (a) To apply the EM algorithm to data censored at a constant c, let U denote the underlying failure time and suppose that $Y = \min(U, c)$ and $D = I(U \le c)$ are observed. Thus the complete-data log likelihood is $\log f(u; \theta)$. Show that

$\delta(\cdot)$ is the Dirac delta function.

$$f(u \mid y, d; \theta) = \begin{cases} \delta(u - y), & d = 1, \\ \frac{f(u;\theta)}{1 - F(c;\theta)}, & u > c, d = 0. \end{cases}$$

(b) If $f(u; \theta) = \theta e^{-\theta u}$, show that $\mathrm{E}(U \mid Y = y, D = d; \theta') = dy + (1 - d)(c + 1/\theta')$, and deduce that the iteration for a random sample $(y_1, d_1), \ldots, (y_n, d_n)$ is

$$\theta^\dagger = \frac{n}{\sum_{j=1}^n \{ d_j y_j + (1 - d_j)(c + 1/\theta') \}}.$$

Show that the missing information is $\sum (1 - d_j)/\theta^2$ and find the rate of convergence of the algorithm. Discuss briefly.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

5.6 Bibliographic Notes

Linear regression is discussed in more depth in Chapter 8, and references to the enormous literature on the topic can be found in Section 8.8. Exponential family models date to work of Fisher and others in the 1930s, are widely used in applications and have been intensively studied. Chapter 5 of Pace and Salvan (1997) is a good reference, while longer more mathematical accounts are Barndorff-Nielsen (1978) and Brown (1986). The term natural exponential family was introduced by Morris (1982, 1983), who highlighted the importance of the variance function.

The roots of group transformation models go back to Pitman (1938, 1939), but owe much of their modern development to D. A. S. Fraser, summarized in Fraser (1968, 1979).

Survival analysis is a huge field with inter-related literatures on industrial and medical problems, though time-to-event data arise in many other fields also. The early literature is mostly concerned with reliability, of which Crowder *et al.* (1991) is an elementary account, while the literature on biostatistical and medical applications has grown enormously over the last 30 years. Cox and Oakes (1984), Miller (1981), Kalbfleisch and Prentice (1980), and Collett (1995) are standard accounts at about this level; see also Klein and Moeschberger (1997). Competing risks are surveyed by Tsiatis (1998); a helpful earlier account is Prentice *et al.* (1978). Their nonidentifiability was first pointed out by Cox (1959). Aalen (1994) gives an elementary account of frailty models, with further references. Keiding (1990) describes inference using the Lexis diagram.

The formal study of missing data began with Rubin (1976), though *ad hoc* procedures for dealing with missing observations in standard models were widely used much earlier. A standard reference is Little and Rubin (1987). More recently the related notion of data coarsening, which encompasses censoring, truncation and grouping as well as missingness, has been discussed by Heitjan (1994).

Although data in areas such as epidemiology and the social and economic sciences are often analyzed as if they were selected randomly from some well-defined population, the possibility that bias has entered the selection process is ever-present; publication bias is just one example of this. There is a large literature on selection bias from many points of view, much of which is mentioned by Copas and Li (1997) and its discussants. Example 5.34 is taken from Copas (1999). Molenberghs *et al.* (2001) give an example of analysis of sensitivity to missing data in contingency tables, with references to related literature.

Special cases of the EM algorithm were used well before it was crystallized and named by Dempster *et al.* (1977), who gave numerous applications and pointed the way for the substantial further work largely summarized in McLachlan and Krishnan (1997). A useful shorter account is Chapter 4 of Tanner (1996). One common criticism of the algorithm is its slowness, and Meng and van Dyk (1997) and Jamshidian and Jennrich (1997) describe some of the many approaches to speeding it up; they also contain further references. Oakes (1999) gives references to the literature on computing standard errors for EM estimates. Modern applications go far beyond the

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

simple exponential family models used initially and may require complex E- and M-steps including Monte Carlo simulation; see for example McCulloch (1997).

Mixture models and their generalizations are widely used in applications, particularly for classification and discrimination problems; see Titterington *et al.* (1985) and Lindsay (1995). The thorny problem of selecting the number of components is given an airing by Richardson and Green (1997) and their discussants, using methods discussed in Section 11.3.3.

5.7 Problems

1 In the linear model (5.3), suppose that $n = 2r$ is an even integer and define $W_j = Y_{n-j+1} - Y_j$ for $j = 1, \ldots, r$. Find the joint distribution of the W_j and hence show that

$$\tilde{\gamma}_1 = \frac{\sum_{j=1}^{r}(x_{n-j+1} - x_j)W_j}{\sum_{j=1}^{r}(x_{n-j+1} - x_j)^2}$$

satisfies $\mathrm{E}(\tilde{\gamma}_1) = \gamma_1$. Show that

$$\mathrm{var}(\tilde{\gamma}_1) = \sigma^2 \left\{ \sum_{j=1}^{n}(x_j - \overline{x})^2 - \frac{1}{2}\sum_{j=1}^{r}(x_{n-j+1} + x_j - 2\overline{x})^2 \right\}^{-1}.$$

Deduce that $\mathrm{var}(\tilde{\gamma}_1) \geq \mathrm{var}(\widehat{\gamma}_1)$ with equality if and only if $x_{n-j+1} + x_j = c$ for some c and all $j = 1 \ldots, r$.

2 Show that the scaled chi-squared density with known degrees of freedom ν,

$$f(v; \sigma^2) = \frac{v^{\nu/2-1}}{(2\sigma^2)^{\nu/2}\Gamma\left(\frac{1}{2}\nu\right)} \exp\left(-\frac{v}{2\sigma^2}\right), \quad v > 0, \sigma^2 > 0, \nu = 1, 2, \ldots,$$

is an exponential family, and find its canonical parameter and observation and cumulant-generating function.

3 Show that the geometric density

$$f(y; \pi) = \pi(1 - \pi)^y, \quad y = 0, 1, \ldots, \ 0 < \pi < 1,$$

is an exponential family, and give its cumulant-generating function.
Show that $S = Y_1 + \cdots + Y_n$ has negative binomial density

$$\binom{n + s - 1}{n - 1}\pi^n(1 - \pi)^s, \quad s = 0, 1, \ldots,$$

and that this is also an exponential family.

4 (a) Suppose that Y_1 and Y_2 have gamma densities (2.7) with parameters λ, κ_1 and λ, κ_2. Show that the conditional density of Y_1 given $Y_1 + Y_2 = s$ is

$$\frac{\Gamma(\kappa_1 + \kappa_2)}{s^{\kappa_1+\kappa_2-1}\Gamma(\kappa_1)\Gamma(\kappa_2)} u^{\kappa_1-1}(s - u)^{\kappa_2-1}, \quad 0 < u < s, \kappa_1, \kappa_2 > 0,$$

and establish that this is an exponential family. Give its mean and variance.
(b) Show that $Y_1/(Y_1 + Y_2)$ has the beta density.
(c) Discuss how you would use samples of form $y_1/(y_1 + y_2)$ to check the fit of this model with known ν_1 and ν_2.

5 If Y has density (5.7) and $\mathcal{Y}_1$ is a proper subset of $\mathcal{Y}$, show the the conditional density of Y given that $Y \notin \mathcal{Y}_1$ is also a natural exponential family.
Find the cumulant-generating function for the truncated Poisson density given by $f_0(y) \propto 1/y!$, $y = 1, 2, \ldots$, and give the likelihood equation and information quantities.
Compare with Practical 4.3.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

6 Show that the two-locus multinomial model in Example 4.38 is a natural exponential family of order 2 with natural observation and parameter $s(Y) = (Y_A + Y_{AB}, Y_B + Y_{AB})^{\mathrm{T}}$ and $(\theta_A, \theta_B)^{\mathrm{T}} = (\log\{\alpha/(1-\alpha)\}, \log\{\beta/(1-\beta)\})$ and cumulant-generating function $m\log(1+e^{\theta_A}) + m\log(1+e^{\theta_B})$. Deduce that the elements of $s(Y)$ are independent. Under what circumstances will maximum likelihood estimation of θ_A, θ_B give infinite estimates?

7 Suppose that $Y_1, \ldots, Y_n$ follow (5.2). Show that the joint density of the Y_j is a linear exponential family of order three, and give the canonical statistics and parameters and the cumulant-generating function. Find the minimal representations in the cases where the x_j (i) are, and (ii) are not, all equal.
Is the model an exponential family when $\mathrm{E}(Y_j) = \beta_0 \exp(x_j\beta_1)$?

8 Show that the multivariate normal distribution $N_p(\mu, \Omega)$ is a group transformation model under the map $Y \mapsto a + BY$, where a is a $p \times 1$ vector and B an invertible $p \times p$ matrix. Given a random sample $Y_1, \ldots, Y_n$ from this distribution, show that

$$\overline{Y} = n^{-1}\sum_{j=1}^{n} Y_j, \quad \sum_{j=1}^{n}(Y_j - \overline{Y})(Y_j - \overline{Y})^{\mathrm{T}}$$

is a minimal sufficient statistic for μ and Ω, and give equivariant estimators of them. Use these estimators to find the maximal invariant.

9 Show that the model in Example 4.5 is an exponential family. Is it steep? What happens when $R_j = 0$ whenever $x_j < a$ and $R_j = m_j$ otherwise?
Find its minimal representation when all the x_j are equal.

10 Independent observations $y_1, \ldots, y_n$ from the exponential density $\lambda\exp(-\lambda y)$, $y > 0$, $\lambda > 0$, are subject to Type II censoring stopping at the rth failure. Show that a minimal sufficient statistic for λ is $S = Y_{(1)} + \cdots + Y_{(r)} + (n-r)Y_{(r)}$, where $0 < Y_{(1)} < Y_{(2)} < \cdots$ are order statistics of the Y_j, and that $2\lambda S$ has a chi-squared distribution on $2r$ degrees of freedom.
A Type II censored sample was 0.2, 0.8, 1.1, 1.4, 2.1, 2.4, 2.4+, 2.4+, 2.4+, where + denotes censoring. On the assumption that the sample is from the exponential distribution, find a 90% confidence interval for λ. How would you check whether the data are exponential?

11 Let $X_1, \ldots, X_n$ be an exponential random sample with density $\lambda\exp(-\lambda x)$, $x > 0, \lambda > 0$. For simplicity suppose that $n = mr$. Let Y_1 be the total time at risk from time zero to the rth failure, Y_2 be the total time at risk between the rth and the $2r$th failure, Y_3 the total time at risk between the $2r$th and $3r$th failures, and so forth.
(a) Let $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ be the ordered values of the X_j. Show that the joint density of the order statistics is

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = n!\, f(x_1)f(x_2)\cdots f(x_n), \quad x_1 < x_2 < \cdots < x_n,$$

and by writing $X_{(1)} = Z_1$, $X_{(2)} = Z_1 + Z_2$, …, $X_{(n)} = Z_1 + \cdots + Z_n$, where the Z_j are the *spacings* between the order statistics $X_{(j)}$, show that the Z_j are independent exponential random variables with hazard rates $(n+1-j)\lambda$.
(b) Hence show that the Y_j have independent gamma distributions with means r/λ and variances r/λ^2. Deduce that the variables $\log Y_j$ are independently distributed with constant variance.
(c) Now suppose that the hazard rate is not constant, but is a slowly-varying smooth function of time, $\lambda(t)$. Explain how a plot of $\log Y_j$ against the midpoint of the time interval between the $(r-1)j$th and the rjth failures can be used to estimate $\log\lambda(t)$.
(Cox, 1979)

12 Let $Y_1, \ldots, Y_n$ be independent exponential variables with hazard λ subject to Type I censoring at time c. Show that the observed information for λ is D/λ^2, where D is the number of the Y_j that are uncensored, and deduce that the expected information is $i(\lambda \mid c) = n\{1 - \exp(-\lambda c)\}/\lambda^2$ conditional on c.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Now suppose that the censoring time c is a realization of a random variable C, whose density is gamma with index ν and parameter $\lambda\alpha$:

$$f(c) = \frac{(\lambda\alpha)^\nu c^{\nu-1}}{\Gamma(\nu)} \exp(-c\lambda\alpha), \qquad c > 0, \alpha, \nu > 0.$$

Show that the expected information for λ after averaging over C is

$$i(\lambda) = n\{1 - (1 + 1/\alpha)^{-\nu}\}/\lambda^2.$$

Consider what happens when (i) $\alpha \to 0$, (ii) $\alpha \to \infty$, (iii) $\alpha = 1$, $\nu = 1$, (iv) $\nu \to \infty$ but $\mu = \nu/\alpha$ is held fixed. In each case explain qualitatively the behaviour of $i(\lambda)$.

13 In a competing risks model with $k = 2$, write

$$\begin{aligned}\Pr(Y \le y) &= \Pr(Y \le y \mid I = 1)\Pr(I = 1) + \Pr(Y \le y \mid I = 2)\Pr(I = 2) \\ &= pF_1(y) + (1 - p)F_2(y),\end{aligned}$$

say. Hence find the cause-specific hazard functions h_1 and h_2, and express F_1, F_2 and p in terms of them.
Show that the likelihood for an uncensored sample may be written

$$p^r(1 - p)^{n-r} \prod_{j=1}^{r} f_1(y_j) \prod_{j=r+1}^{n} f_2(y_j)$$

and find the likelihood when there is censoring.
If $f(y_1 \mid y_2)$ and $f(y_2 \mid y_1)$ be arbitrary densities with support $[y_2, \infty)$ and $[y_1, \infty)$, then show that the joint density

$$f(y_1, y_2) = \begin{cases} pf_1(y_1)f(y_2 \mid y_1), & y_1 \le y_2, \\ (1 - p)f_2(y_2)f(y_1 \mid y_2), & y_1 > y_2, \end{cases}$$

produces the same likelihoods. Deduce that the joint density is not identifiable.

14 Find the cause-specific hazard functions for the bivariate survivor functions

$$\begin{aligned}\mathcal{F}(y_1, y_2) &= \exp[1 - \theta_1 y_1 - \theta_2 y_2 - \exp\{\beta(\theta_1 y_1 + \theta_2 y_2)\}], \\ \mathcal{F}^*(y_1, y_2) &= \exp\left[1 - \theta_1 y_1 - \theta_2 y_2 - \sum_{i=1}^{2} \frac{\theta_i}{\theta_1 + \theta_2} \exp\{\beta(\theta_1 + \theta_2)y_i\}\right],\end{aligned}$$

where $y_1, y_2 > 0, \theta_1, \theta_2 > 0$ and $\beta > -1$. Under what condition does $\mathcal{F}$ yield independent variables?
Write down the likelihoods based on random samples $(y_1, i_1, d_1), \ldots, (y_n, i_n, d_n)$ from these two models. Discuss the interpretation of $\widehat{\beta} \gg 0$ in the absence of external evidence for $\mathcal{F}$ over $\mathcal{F}^*$.
(Prentice *et al.*, 1978)

15 (a) Let $Z = X_1 + \cdots + X_N$, where N is Poisson with mean μ and the X_i are independent identically distributed variables with moment-generating function $M(t)$. Show that the cumulant-generating function of Z is $K_Z(t) = \mu\{M(t) - 1\}$ and that $\Pr(Z = 0) = e^{-\mu}$.
If the X_i are gamma variables, show that $K_Z(t)$ may be written as

$$\frac{\alpha}{(\alpha - 1)\delta}[\{1 - \alpha t/(\gamma\delta)\}^{1-\alpha} - 1], \qquad \gamma, \delta > 0, \tag{5.43}$$

where $\alpha > 1$, show that $\mathrm{E}(Z) = \gamma$ and $\mathrm{var}(Z)/\mathrm{E}(Z)^2 = \delta$, and find $\Pr(Z = 0)$ in terms of α, δ and γ. Show that as $\alpha \to 1$ the limiting distribution of Z is gamma, and explain why.
(b) For a frailty model, set $\gamma = 1$ and suppose that an individual has hazard $Zh(y)$, $y > 0$. Compute the population cumulative hazard $H_Y(y)$ and show that if $\alpha > 1$ then

$$\lim_{y \to \infty} H_Y(y) < \infty.$$

Z is a continuous variable for $0 < \alpha < 1$, but you need not show this.

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

Give an interpretation of this in terms of the distribution of the lifetime Y. (Are all the individuals in the population liable to fail?)
(c) Obtain the population hazard rate $h_Y(y)$, take $h(y) = y^2$, and graph $h_Y(y)$ for $\delta = 0, 0.5, 1, 2.5$. Discuss this in relation to the divorce rate example on page 201.
(d) Now suppose that there are two groups of individuals, the first with individual hazards $h(y)$ and the second with individual hazards $rh(y)$, where $r > 1$. Thus the effect of transferring an individual from group 1 to group 2, if this were possible, would be to increase his hazard by a factor r. If frailties in the two groups have the same cumulant-generating function (5.43), show that the ratio of group hazard functions is

$$\frac{h_2(y)}{h_1(y)} = r\left\{\frac{1+\alpha^{-1}\delta H(y)}{1+r\alpha^{-1}\delta H(y)}\right\}^{\alpha}.$$

Establish that this is a decreasing function of y, and explain why its limiting value is less than one, that is, the risk is eventually lower in group 2, if $\alpha > 1$. What difficulties does this pose for the interpretation of group differences in survival?
(Aalen, 1994; Hougaard, 1984)

16 (a) Show that when data (X, Y) are available, but with values of Y missing at random, the log likelihood contribution can be written

$$\ell(\theta) \equiv I \log f(Y \mid X;\theta) + \log f(X;\theta),$$

and deduce that the expected information for θ depends on the missingness mechanism but that the observed information does not.
(b) Consider binary pairs (X, Y) with indicator I equal to zero when Y is missing; X is always seen. Their joint distribution is given by

$$\Pr(Y=1 \mid X=0) = \theta_0, \quad \Pr(Y=1 \mid X=1) = \theta_1, \quad \Pr(X=1) = \lambda,$$

while the missingness mechanism is

$$\Pr(I=1 \mid X=0) = \eta_0, \quad \Pr(I=1 \mid X=1) = \eta_1.$$

(i) Show that the likelihood contribution from (X, Y, I) is

$$\left[\left\{\theta_1^Y(1-\theta_1)^{1-Y}\right\}^X\left\{\theta_0^Y(1-\theta_0)^{1-Y}\right\}^{1-X}\right]^I \times \left\{\eta_0^I(1-\eta_0)^{1-I}\right\}^{1-X}\left\{\eta_1^I(1-\eta_1)^{1-I}\right\}^X \times \lambda^X(1-\lambda)^{1-X}.$$

Deduce that the observed information for θ_1 based on a random sample of size n is

$$-\frac{\partial^2\ell(\theta_0,\theta_1)}{\partial\theta_1^2} = \sum_{j=1}^{n} I_j X_j\left\{\frac{Y_j}{\theta_1^2} + \frac{1-Y_j}{(1-\theta_1)^2}\right\}.$$

Give corresponding expressions for $\partial^2\ell(\theta_0,\theta_1)/\partial\theta_0^2$ and $\partial^2\ell(\theta_0,\theta_1)/\partial\theta_0\partial\theta_1$.
(ii) Statistician A calculates the expected information treating $I_1, \ldots, I_n$ as fixed and thereby ignores the missing data mechanism. Show that he gets $i_A(\theta_1,\theta_1) = M\lambda/\{\theta_1(1-\theta_1)\}$, where $M = \sum I_j$, and find the corresponding quantities $i_A(\theta_0,\theta_1)$ and $i_A(\theta_0,\theta_0)$. If he uses this procedure for many sets of data, deduce that on average M is replaced by $n\Pr(I=1) = n\{\lambda\eta_1 + (1-\lambda)\eta_0\}$.
(iii) Statistician B calculates the expected information taking into account the missingness mechanism. Show that she gets $i_B(\theta_1,\theta_1) = n\lambda\eta_1/\{\theta_1(1-\theta_1)\}$, and obtain $i_B(\theta_0,\theta_1)$ and $i_B(\theta_0,\theta_0)$.
(iv) Show that A and B get the same expected information matrices only if Y is missing completely at random. Does this accord with the discussion above?
(c) Statistician C argues that expected information should never be used in data analysis: even if the data actually observed are complete, unless it can be guaranteed that data

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

could not be missing at random for any reason, every expected information calculation should involve every potential missingness mechanism. Such a guarantee is impossible in practice, so no expected information calculation is ever correct. Do you agree?
(Kenward and Molenberghs, 1998)

17 (a) In Example 5.34, suppose that n patients are divided randomly into control and treatment groups of equal sizes $n_C = n_T = n/2$, with death rates λ_C and λ_T. If the numbers of deaths R_C and R_T are small, use a Poisson approximation to the binomial to show that the difference in log rates is roughly $\widehat{\mu} = \log R_C - \log R_T$. What would you conclude if $\widehat{\mu} \doteq 0$?
(b) Show that if $\lambda_C \doteq \lambda_T = \lambda$, then $\text{var}(\widehat{\mu}) \doteq 4/(n\lambda)$, and use the estimates $\widehat{\lambda}_C = R_C/n_C$, $\widehat{\lambda}_T = R_T/n_T$ and $\widehat{\lambda} = (R_C + R_T)/(n_C + n_T)$ to check a few values of $\widehat{\mu}$ and the standard errors in Table 5.9.
(c) In practice the variance in (b) is typically too small, because it does not allow for inter-trial variability. Different studies are performed with different populations, in which the treatment may have different effects. We can imagine two stages: we first choose a population in which the treatment effect is $\mu + \eta$, where η is random with mean zero and variance σ^2; then we perform a trial with n subjects and produce an estimator $\widehat{\mu}$ of $\mu + \eta$ with variance ν/n. Show that $\widehat{\mu}$ may be written $\mu + \eta + \varepsilon$, give the variance of ε, and deduce that when both stages of the trial are taken into account, $\widehat{\mu}$ has mean μ and variance $\sigma^2 + \nu/n$.
How would this affect the calculations in Example 5.34?

18 (a) Show that the t density of Example 4.39 may be obtained by supposing that the conditional density of Y given $U = u$ is $N(\mu, \nu\sigma^2/u)$ and that $U \sim \chi^2_\nu$. Show that $U \overset{D}{=} V/\{\nu + (y-\mu)^2/\sigma^2\}$ conditional on Y, where $V \sim \chi^2_{\nu+1}$, and with $\theta = (\mu, \sigma^2)$ deduce that

$$\text{E}(U \mid Y; \theta) = \frac{\nu + 1}{\nu + (y-\mu)^2/\sigma^2}.$$

(b) Consider the EM algorithm for estimation of θ when ν is known. Show that the complete-data log likelihood contribution from (y, u) may be written

$$-\frac{1}{2}\sigma^2 - \frac{1}{2}u(y-\mu)^2/2(\nu\sigma^2),$$

and hence give the M-step. Write down the algorithm in detail.
(c) Show that the result of the EM algorithm satisfies the self-consistency relation $\theta = g(\theta)$, and given the form of g when σ^2 is both known and unknown.
(d) The Cauchy log likelihood shown in the right panel of Figure 4.2 corresponds to setting $\nu = \sigma^2 = 1$. In this case explain why $\mu^\dagger$ converges to a local or a global maximum or a local minimum, depending on the initial value for μ.

19 Suppose that $U_1, \ldots, U_q$ have a multinomial distribution with denominator m and probabilities $\pi_1, \ldots, \pi_q$ that depend on a parameter θ, and that the maximum likelihood estimator of θ based on the U_s has a simple form. Some of the categories are indistinguishable, however, so the observed data are $Y_1, \ldots, Y_p$, where $Y_r = \sum_{s \in \mathcal{A}_r} U_s$; $\mathcal{A}_1, \ldots, \mathcal{A}_p$ partition $\{1, \ldots, q\}$ and none is empty.
(a) Show that the E-step of the EM algorithm for estimation of θ involves

$$\text{E}(U_s \mid Y = y; \theta') = \frac{y_r \pi'_s}{\sum_{t \in \mathcal{A}_r} \pi'_t}, \quad s \in \mathcal{A}_r,$$

and say how the M-step is performed.
(b) Let $(\pi_1, \ldots, \pi_5) = (1/2, \theta/4, (1-\theta)/4, (1-\theta)/4, \theta/4)$, and suppose that $y_1 = u_1 + u_2 = 125$, $y_2 = u_3 = 18$, $y_3 = u_4 = 20$ and $y_4 = u_5 = 34$. These data arose in a genetic linkage problem and are often used to illustrate the EM algorithm. Show that

$$\theta^\dagger = \frac{y_4 + y_1\theta'/(2+\theta')}{m - 2y_1/(2+\theta')},$$

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006

and find the maximum likelihood estimate starting with $\theta' = 0.5$.
(c) Show that the maximum likelihood estimator of $\widehat{\lambda}_A$ in the single-locus model of Example 4.38 may be written $\widehat{\lambda}_A = (2u_1 + u_2 + u_5)/m$ and establish that

$$\mathrm{E}(U_1 \mid Y; \lambda') = y_1 \lambda'_A / (2 - 2\lambda'_B - \lambda'_A).$$

Give the corresponding expressions for $\widehat{\lambda}_B$ and $\mathrm{E}(U_2 \mid Y; \lambda')$. Hence give the M-step for this model. Apply the EM algorithm to the data in Table 4.3, using starting-values obtained from categories with probabilities $2\lambda_A\lambda_B$ and λ_O^2.
(d) Compute standard errors for your estimates in (b) and (c).
(Rao, 1973, p. 369)

Downloaded from https://www.cambridge.org/core. University of Toronto, on 25 Aug 2020 at 18:27:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511815850.006