MONOGRAFIAS DE MATEMÁTICA Nº 51

THE THEORY OF EXPONENTIAL DISPERSION MODELS AND ANALYSIS OF DEVIANCE

Bent Jørgensen

ISBN

85-244-0067-6

Conselho Nacional de Desenvolvimento Científico e Tecnológico

INSTITUTO DE MATEMÁTICA PURA E APLICADA
Estrada Dona Castorina, 110
22.460 – Rio de Janeiro-RJ

MONOGRAFIAS DE MATEMÁTICA

(títulos já publicados)

01) Azevedo, Alberto / Piccinini, Renzo - INTRODUÇÃO À TEORIA DOS GRUPOS (1970) / reprodução (1984)
02) Santos, Nathan M. - VETORES E MATRIZES (1970) - esgotada
03) Carmo, Manfredo P. do - INTRODUÇÃO À GEOMETRIA DIFERENCIAL GLOBAL (1970) - esgotada
04) Palis Junior, Jacob - SEMINÁRIO DE SISTEMAS DINÂMICOS (1971) - esgotada
05) Carvalho, João Pitombeira de - INTRODUÇÃO À ÁLGEBRA LINEAR (1971) - esgotada
06) Fernandez, Pedro Jesus - INTRODUÇÃO À TEORIA DAS PROBABILIDADES (1971) - esgotada
07) Robinson. R. C. - LECTURES ON HAMILTONIAN SYSTEMS (1972) - esgotada
08) Carmo, Manfredo P. do - NOTAS DE GEOMETRIA RIEMANNIANA (1972) - esgotada
09) Hönig, Chaim S. - ANÁLISE FUNCIONAL E O PROBLEMA DE STURM-LIOUVILLE (1972) - esgotada
10) Melo, Welington de - ESTABILIDADE ESTRUTURAL EM VARIEDADES DE DIMENSÃO 2 (1972) - esgotada
11) Lesmes, Jaime - TEORIA DAS DISTRIBUIÇÕES E EQUAÇÕES DIFERENCIAIS (1972) - esgotada
12) Vilanova, Clóvis - ELEMENTOS DA TEORIA DOS GRUPOS E DA TEORIA DOS ANÉIS (1972) - esgotada
13) Douai, Jean Claude - COHOMOLOGIE DES GROUPES (1973) - esgotada
14) Lawson Jr., H. Blaine - LECTURES ON MINIMAL SUBMANIFOLDS,Vol. I (1973) - esgotada
15) Lima, Elon Lages - VARIEDADES DIFERENCIÁVEIS (1973) - esgotada
16) Mendes, Pedro - TEOREMAS DE Ω-ESTABILIDADE E ESTABILIDADES ESTRUTURAL ÉM VARIEDADES ABERTAS (1973) - esgot.
17) Amann, Herbert - LECTURES ON SOME FIXED POINT THEOREMS (1974) - esgotada
18) - EXERCÍCIOS DE MATEMÁTICA / IMPA (1974) - esgotada
19) Figueiredo, Djairo Guedes de - NÚMEROS IRRACIONAIS E TRANSCEDENTES (1975) - esgotada
20) Zeeman, C. E. - UMA INTRODUÇÃO INFORMAL À TOPOLOGIA DAS SUPERFÍCIES (1975) - esgotada
21) Carmo, Manfredo P. do - NOTAS DE UM CURSO DE GRUPOS DE LIE (1975) - esgotada
22) Prestel, Alexander - LECTURES ON FORMALLY REAL FIELDS (1975) - esgotada
23) Simis, Aron - INTRODUÇÃO À ÁLGEBRA (1976) - esgotada
24) Lesmes, Jaime - SEMINÁRIO DE ANÁLISE FUNCIONAL (1976) - esgotada
25) Brauer, Fred - SOME STABILITY AND PERTURBATION PROBLEM FOR DIFFERENTIAL AND INTEGRAL EQUATIONS (1976) - esg.
26) Rodriguez, Lócio - GEOMETRIA DAS SUBVARIEDADES (1976) - esgotada
27) Miranda, Mário - FRONTIÈRE MINIME (1976)
28) Cardoso, Fernamdo - RESOLUBILIDADE LOCAL DE EQUAÇÕES DIFERENCIAIS PARCIAIS (1977) - esgotada
29) Becker, Eberhard - HEREDITARILY-PYTHAGOREAN FIELDS AND ORDERINGS OF HIGHER LEVEL (1978)
30) Bass, Hyman - PROJECTIVE MODULES AND SYMMETRIC ALGEBRAS (1978)
31) Neyman, J. - PROBABILIDADE FREQÜENTISTA E ESTATÍSTICA FREQÜENTISTA (1978)
32) Dumortier, Freddy - SINGULARITIES OF VECTOR FIELDS (1978)
33) Viswanathan, T. M. - INTRODUÇÃO À ÁLGEBRA E ARITMÉTICA (1979) - esgotada
34) Thayer, F. Javier - NOTES ON PARTIAL DIFFERENTIAL EQUATIONS (1980)
35) Bierstone, Edward - THE STRUCTURE OF ORBIT SPACES AND THE SINGULARITIES OF EQUIVARIANT MAPPINGS (1980)
36) Thayer, F. Javier - THÉORIE SPECTRALE (1982)
37) Carmo, Manfredo P. do - FORMAS DIFERENCIAIS E APLICAÇÕES (1983)
38) Prestel, Alexander / Roquettę, Peter - LECTURES ON FORMALLY p-ADIC FIELDS (1983)
39) Lequain, Yves / Garcia, Arnaldo - ÁLGEBRA: UMA INTRODUÇÃO (1983) - esgotada
40) Barbosa, J. Lucas / Colares, A. Gervásio - MINIMAL SURFACES IN R^3 (1986)
41) Bérard, Pierre H. - SPECTRAL GEOMETRY: DIRECT AND INVERSE PROBLEMS (1986)
42) Bérard, Pierre H. - ANALYSIS ON RIEMANNIAN MANIFOLDS AND GEOMETRIC APPLICATIONS: AN INTRODUCTION (1987)
43) Torres, Felipe Cano - DESINGULARIZATION STRATEGIES FOR THREE-DIMENSIONAL VECTOR FIELDS (1988)
44) Endler, Otto - TEORIA DOS CORPOS (1988)
45) Bruns, Winfried / Vetter, Udo - DETERMINANTAL RINGS (1988)
46) Hefez, Abramo - INTRODUÇÃO À GEOMETRIA PROJETIVA (1990)
47) Gouvêa, Fernando Quadros - FORMAS MODULARES: UMA INTRODUÇÃO (1990)
48) Jørgensen, Bent - EXPONENTIAL DISPERSION MODELS (1991)
49) Bustos, Oscar H. / Frery, Alejandro C. - SIMULAÇÃO ESTOCÁSTICA: TEORIA E ALGORITMOS (Versão Completa) (1992)
50) Létac, Gérard - LECTURES ON NATURAL EXPONENTIAL FAMILIES AND THEIR VARIANCE FUNCTIONS (1992)

INDEX

PREFACE

This book is an introduction to exponential dispersion models and generalized linear models. It has three important features: first, it has an extensive introduction to the mathematical theory of exponential dispersion models, second it has a thorough account of asymptotic theory for generalized linear models, and third it treats some generalizations of exponential dispersion models, which help to put the theory into its proper perspective.

The book is intended for the graduate level, and requires familiarity with basic concepts of statistical inference and linear models. Some prior exposure to discrete data analysis would also be useful.

The book is divided into three chapters. Chapter 1 is an introduction to some of the basic concepts used in the book. Chapter 2 treats the theory of one-dimensional natural exponential families and exponential dispersion models, and includes some theory for variance functions. Chapter 3 concerns analysis of deviance for generalized linear models and other regression models related to dispersion models.

I have tried to make the exposition self-contained, and to include, as far as possible, a derivation of the results stated. The text as such contains rather few references to the literature, but at the end of each chapter there is a section with key references and notes on the origin of some of the ideas. This is not intended as an exhaustive summary of the literature, and doubtless some important references have been left out, as I have cited only the literature that I found most useful as introduction to a given subject.

The book is a preliminary version of the first part of a more comprehensive volume on exponential dispersion models, which I hope to complete in the near future. In the second part of the volume, I intend to treat multivariate exponential dispersion models, and their applications to analysis of correlated data.

I am grateful to the organizers of the 1st School on Linear Models, University of Sao Paulo, for inviting me to present this material at the meeting, and for publishing this preliminary version of the text. I am grateful to Ole E. Barndorff-Nielsen and James K. Lindsey, who have read and commented on part of the manuscript. A special thanks goes to my students Glaura da Conceição Franco, Renata P.L. Jeronymo and Alejandro Frery, who were the first to be exposed to the material, and whose favourable response encouraged me to go on. Finally, I thank Lais Ventura Santos and Rogerio Dias Trindade for typing the manuscript.

Rio de Janeiro, September 1988

Bent Jørgensen

PREFACE TO THE SECOND EDITION

The second edition is identical to the first edition, except for the correction of some typing errors.

Rio de Janeiro, April 1992

Bent Jørgensen

Chapter 1

INTRODUCTION

In this book we consider statistical regression analysis, with special emphasis on regression models based on distributions in the class of exponential dispersion models. In the present chapter we introduce some of the main ideas of the book, and consider some motivating examples.

§1.1 Regression Models

The concept of regression is very important in statistical data analysis. We use the word regression in a rather broad sense, and by a *regression model* we understand a statistical model with the following two ingredients:

(i) A random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ with a distribution involving a vector of unknown parameters $\mu = (\mu_1, \ldots, \mu_r)^T$.

(ii) A relation between μ and the parameter vector $\beta = (\beta_1, \ldots, \beta_k)^T$ of the form $\mu = f(\beta)$, where f is a given smooth, one-to-one function.

We refer to (i) and (ii) as respectively the *random component* and the *systematic component* of the model. When the random component is understood from the context, we refer to (ii) as simply the *model*. The random vector $\mathbf{Y}$ is called the *response variable*, and the distribution of $\mathbf{Y}$ or Y_i is called the *error distribution*. The random component may generally reflect any kind of stochastic variation, including measurement error. The function f is called the *regression function*, and the parameters $\beta_1 \ldots, \beta_k$ are called *regression parameters*. In many cases, a regression model includes an additional parameter σ^2 expressing the dispersion of Y_i in some sense, and we refer to σ^2 as a *dispersion parameter*.

An important class of regression models corresponds to the case where the systematic component has the form

$$g(\mu_i) = \eta_i, \quad i = 1, \ldots, r \tag{1.1}$$

$$\eta_i = \sum_{j=1}^{k} x_{ij}\beta_j, \quad i = 1, \ldots, r. \tag{1.2}$$

The function g, assumed to be one-to-one and smooth, is called the *link function*. The x_{ij}'s, which are known constants, are called *covariates* or *explanatory variables*. The matrix $X = \{x_{ij}\}$ is called the *design matrix* for the model. We call a model of the form (1.1), (1.2) a *linear model*. A regression model which is not of this form is called *nonlinear*.

We assume that the reader is familiar with the idea of a *linear normal model*, in which $Y_1, \ldots, Y_n$ are independent and normally distributed, $Y_i \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots, n$ and the systematic structure is of the form (1.1), (1.2) with $r = n$, and g is the identity function.

Regression models are tools used in the analysis of data from scientific experiments, or more generally from any kind of investigation in which data are gathered in a systematic may. The researcher provides the data in the form of the observed values $y_1, \ldots, y_n$ for $Y_1, \ldots, Y_n$, and values of the explanatory variables x_{ij}.

The objective of the statistical analysis is to find a suitable statistical model, to check the goodness of fit of the model to the data, to estimate the unknown parameters of the model, and to test hypotheses about the parameters.

The researcher could supply the systematic component of the model, which is often a concise mathematical expression of some aspect of the scientific theory under investigation. In more exploratory investigations it may be part of the statistical analysis to build a suitable systematic component. The random component of the regression model should ideally be derived from the scientific theory under investigation, but it too is often to be built as part of the statistical analysis of the data. In the statistical analysis, the systematic component of the model may be investigated, and possibly modified. The results of the analysis, in turn, may be used to confirm or possibly to modify the random component of the model. The following example serves to illustrate the idea of a statistical model as part of the scientific theory under investigation.

Example 1.1.1: *Energy expenditure data 1.* The data in Table 1:1 are from an investigation (Garby et al., 1988), concerning the energy expenditure for human subjects at a given physical activity and for a given time period. We consider the largest set of data from the investigation, consisting of 104 women. The variables measured for the i'th subject were total energy expenditure at rest for a 24 hour period (y_i), mass of fat tissue (x_{i1}) and mass of fat-free tissue (x_{i2}). If we assume that each of the two types of tissue are homogeneous with respect to energy expenditure, we have a model of the form

$$\mu_i = \beta_1 x_{i1} + \beta_2 x_{i2}, \tag{1.3}$$

Table 1.1: *Energy expendidture (y) at rest and mass of fat (x_1) and fat-free (x_2) tissue for 104 female subjects*

y	x_1	x_2	y	x_1	x_2
100.08	48.83	63.42	89.97	41.09	60.11
113.22	51.83	70.12	101.09	34.82	55.38
82.22	22.54	52.36	105.13	40.58	60.37
105.81	41.02	57.56	91.32	34.04	50.86
97.38	48.65	63.20	97.38	47.80	56.35
101.43	37.54	74.86	81.88	46.39	60.26
95.70	54.49	59.51	89.30	46.25	55.85
63.69	33.03	48.72	66.05	20.74	65.31
76.16	31.83	49.57	81.55	33.18	57.22
115.24	51.51	70.84	88.96	49.42	51.43
76.49	29.34	54.96	83.57	41.91	56.69
84.24	23.82	56.38	96.71	40.45	66.85
95.36	39.44	56.76	95.03	49.58	58.92
67.39	25.08	33.92	86.26	33.51	61.69
85.93	36.54	58.96	106.82	56.20	70.10
116.59	39.51	62.59	86.94	41.70	63.60
101.09	41.27	61.13	75.48	39.67	65.83
90.31	32.20	64.20	72.45	6.22	36.08
99.74	49.16	63.59	83.57	28.12	52.68
82.89	51.46	61.39	78.85	9.86	53.34
78.85	23.52	59.88	96.71	36.17	56.33
90.64	37.55	58.25	84.92	50.49	62.46
101.43	62.08	71.42	82.29	34.26	66.79
101.76	36.91	62.59	77.84	31.01	56.84
85.93	28.13	58.17	101.09	25.32	61.38
69.75	36.54	43.76	85.93	34.54	56.36
79.19	32.96	55.64	84.24	31.41	59.89
64.36	9.14	50.96	105.47	13.48	52.62
78.51	6.33	53.97	93.34	39.35	63.65
85.25	41.00	54.80	78.51	38.12	54.18
110.53	48.86	67.74	79.86	29.20	57.70
101.09	55.77	59.93	75.48	25.70	51.70
84.92	46.13	49.57	105.13	35.61	55.69
81.21	33.66	48.64	99.41	31.41	51.69

97.72	50.05	63.95	94.35	41.62	51.48
70.43	21.71	46.79	96.37	37.02	59.88
114.91	66.82	61.43	65.37	30.73	48.67
98.73	68.99	51.45	79.86	53.58	51.72
100.42	52.35	53.62	78.51	40.96	47.70
84.24	61.54	58.90	92.33	37.40	54.50
90.31	45.21	53.29	85.59	34.54	65.86
89.63	45.65	49.25	75.14	37.44	47.46
79.52	34.48	47.62	95.36	62.50	57.00
60.08	17.31	43.22	81.21	61.39	61.64
94.35	47.35	59.78	71.10	22.95	46.18
88.62	33.90	47.40	60.08	34.09	43.74
73.80	7.06	53.84	83.57	6.77	55.33
73.12	5.39	51.91	79.19	5.84	50.86
69.75	2.76	44.84	77.17	30.00	45.00
69.42	6.32	53.88	77.50	3.40	56.30
67.73	5.76	54.84	80.54	8.50	53.10
91.32	36.23	61.17	113.90	47.59	52.81

where $\mu_i = E(Y_i)$, and β_1 and β_2 are the specific energy expenditures for the two types of tissue. The simplest random component for this model is probably the case where $Y_1, \ldots Y_n$ are independent and $Y_i \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots, n$, in other words a linear normal model. Dividing through by $w_i = x_{i1} + x_{i2}$ in (1.3) gives the linear relationship

$$(1.4) \qquad \overline{\mu}_i = \beta_2 + (\beta_1 - \beta_2)\overline{x}_{i1},$$

were $\overline{\mu}_i = \mu_i / w_i$ and $\overline{x}_{i1} = x_{i1}/w_i$. Figure 1.1 shows a plot of $\overline{y}_i = y_i/w_i$ against $\overline{x}_{i1}$, confirming the linearity of the relationship.

The parameters β_1 and β_2 in (1.3) are obviously interesting physiological constants, and their estimates are important in for example nutrition research. The estimates based on the above linear normal model were $\hat{\beta}_1 = 0.306$ and $\hat{\beta}_2 = 1.35$, indicating that the energy expenditure in fat tissue (β_1) is much lower than in fat-free tissue (β_2). Comparable values were obtained in similar experiments involving male subjects. ▮

The view of statistical data analysis taken above, and illustrated by Example 1.1.1, is obviously very simplified and idealized, but it serves as a convenient framework for conveying some basic ideas about data analysis. In this book, we shall pay special attention to the random component of the model, and emphasize the importance of the proper choice of error distribution. The researcher may sometimes be more interested in the systematic component of the

```
   1.8000 |  y
   1.7400 |           *
   1.6800 |
   1.6200 |                *
   1.5600 |
   1.5000 |
   1.4400 |    *
   1.3800 |        *
   1.3200 |    *   2 *
   1.2600 |       *    *
   1.2000 |         *                    *
   1.1400 |       **              *       2* *    *
   1.0800 |            *           2       * 2
   1.0200 |                       * **** **32  *
   0.9600 |                       *   2  **223  * **
   0.9000 |                           24 32 **424 2* *
   0.8400 |                           **  **** 2* *       *
   0.7800 |                   *             * *2 *  * *
   0.7200 |                              *      *    *
   0.6600 |                                         *
   0.6000 |                                                         x
----------:---------:---------:---------:---------:---------:---------:
        0.000     0.120     0.240     0.360     0.480     0.600     0.720
```

Figure 1.1 Plot of $\overline{y}_i$ against $\overline{x}_{i1}$ for the data in Table 1.1.

model, if it expresses his key ideas about the structure of the data, or he may be interested in some aspect of the error distribution, such as for example the tails of the distribution, which are important in reliability or lifetime studies.

It is our duty to provide a form of the random component of the model such that we are allowed to investigate the systematic component as freely and as accurately as possible. The choice of model, in particular the random component, is often one of the most difficult tasks in data analysis, and is probably best described as an art, that requires application of a skilful blend of mathematical insight for the problem at hand, experience and trial and error. In the face of such difficulty one may feel compelled to use nonparametric or semiparametric models. Such models involve fewer assumptions about the data, and hence may avoid to build untenable assumptions into the conslusions. However, we adopt the piont of view that given a wide range of models to choose from, basing the choice on a physical model for the data, and carefully checking the fit of the model, may give a deeper insight than a nonparametric analysis. Nonparametric models obviously have their place in the statistician's toolbag, but here we shall deal exclusively with parametric models.

At this point it is important to remember that there may be many models that explain a given set of data equally well. By the very nature of statistical data analysis, that is the formation of conclusions in the face of uncertainty, it is often wise to be less than definitive in one's statements about the data. This suggests that one should give ample information concerning the data analysis, by for example reporting unsuccessful models etc., allowing the reader (client, listener etc.) to draw his own conclusions. One should also remember that valid models for a given data set may differ from each other in terms of their systematic component, their random component, or both.

§1.2 Exponential Dispersion Models

A very convenient and flexible type of error distribution is provided by the class of exponential dispersion models, the main theme of the book. In dimension one, an exponential dispersion model for a random variable Y is defined by the probability density function

$$p(y;\theta,\lambda) = a(\lambda, y)\exp[\lambda\{y\theta - \kappa(\theta)\}], \quad y \in \mathbb{R}, \tag{2.1}$$

for suitable functions a and κ, where $\lambda > 0$, and θ varies in an interval Θ on the real line. We denote (2.1) by $Y \sim ED(\mu, \sigma^2)$, where $\mu = \kappa'(\theta)$ is the expectation of Y, and $\sigma^2 = 1/\lambda$ is the dispersion parameter.

Example 1.2.1: *The normal distribution.* The normal distribution $N(\mu, \sigma^2)$, with μ and σ^2 unknown, is an exponential dispersion model. This may be seen from the following expression for the density function

$$\begin{aligned} p(y;\mu,\sigma^2) &= (2\pi\sigma^2)^{-1/2}\exp\{(y-\mu)^2/(2\sigma^2)\} \\ &= (2\pi\sigma^2)^{-1/2}\exp\{-y^2/(2\sigma^2)\}\exp\{(y\mu - \frac{1}{2}\mu^2)/\sigma^2\}, \end{aligned}$$

which is of the form (2.1) with $\mu = \theta$, $\kappa(\theta) = \frac{1}{2}\theta^2$, $\lambda = 1/\sigma^2$ and $a(\lambda, y) = (2\pi/\lambda)^{-1/2}\exp(-\frac{1}{2}\lambda y^2)$. ∎

In the following chapters we study in detail regression models based on exponential dispersion models. Particular attention is paid to the class of *generalized linear models*, which are models consisting of the following two ingredients:

(i) $Y_1, \ldots, Y_n$ are independent and $Y_i \sim ED(\mu_i, \sigma^2)$, $i = 1, \ldots, n$, where $ED(\mu, \sigma^2)$ is a given exponential dispersion model.

(ii) The systematic component of the model is given by a link function (1.1) and a linear structure (1.2).

Note that in the terminology introduced in Section 1.1, this model is linear, while the terminology "generalized" linear model refers to the fact that it generalizes the idea of a linear normal model.

In the discrete case, the form (2.1) is not appropriate, and we define a *discrete exponential dispersion model* to be a discrete distribution with probability function of the form

$$p(z;\theta,\lambda) = a(\lambda, z)\exp\{z\theta - \lambda\kappa(\theta)\}, \quad z \in \mathbb{N}_0, \tag{2.2}$$

where $\mathbb{N}_0 = 0, 1, 2, \ldots$. There is a close relation between (2.2) and (2.1), in fact, in terms of the average $y = z/\lambda$, (2.2) is of the same form as (2.1). However, since the support of y depends on λ in the discrete case, it is necessary to use (2.2) for discrete data, whereas (2.1) is preferable for continuous data, because it allows a much more elegant inference. By analogy with the continuous case we call $\sigma^2 = 1/\lambda$ the dispersion parameter of (2.2). The expectation of (2.2) is $m = \lambda\kappa'(\theta)$.

Example 1.2.2: *The binomial distribution.* The binomial distribution $Bi(n, \mu)$ has probability function of the form

$$\begin{aligned} p(z; n, \mu) &= \binom{n}{z}\mu^z(1-\mu)^{n-z} \\ &= \binom{n}{z}\exp[z\log\{\mu/(1-\mu)\} + n\log(1-\mu)], \end{aligned}$$

for $z = 0, 1, \ldots, n$. The probability function is hence of the form (2.2) with $\theta = \log\{\mu/(1-\mu)\}$, $\lambda = n$ and $\kappa(\theta) = \log(1 + e^\theta)$, which shows that the binomial distribution is a discrete exponential dispersion model. ▮

Example 2.3: *Beetle mortality data.* Data involving proportions provide instructive examples of generalized linear models. Table 1.2 shows data from an experiment (Bliss, 1935) in which insects were exposed to gaseous carbon disulphide for a period of 5 hours. The table shows the dose (x_i), the number of insects (n_i) and the number of insects killed (z_i) for 8 experiments. We consider a generalized linear model with $Z_1, \ldots, Z_n$ independent and Z_i binomially distributed, $Z_i \sim Bi(n_i, \mu(x_i))$, where $\mu(x)$ denotes the probability that an insect given the dose x is killed. Given a link function g, the systematic component of the model is assumed to be of the form

$$\mu(x_i) = g^{-1}(\beta_0 + \beta_1 x_i), \quad i = 1, \ldots, n, \tag{2.3}$$

where g is increasing, differentiable and maps the interval $(0,1)$ into $I\!R$.

Figure 1.2 shows the observed proportion of killed insects, z_i/n_i, as a function of x_i for the data in Table 1.2. A typical choice of g that makes (2.3) fit the sigmoid shape seen in Figure 1.2 is the *logit link function* $g(\mu) = \log\{\mu/(1-\mu)\}$, which corresponds to $g^{-1}(\eta) = e^\eta/(1+e^\eta)$. In general, g^{-1} could be any continuous distribution function, for example the standard normal distribution function $\Phi(\eta)$, which gives the so-called *probit link function.* ∎

Table 1.2: *Beetle mortality data*

Dose, x_i ($\log_{10} CS_2 mgl^{-1}$)	Number of insects, n_i	Number killed, z_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

§1.3 Dispersion Models

A *dispersion model* is defined as any class of probability density functions of the form

$$p(y;\mu,\lambda) = a(\lambda,y)\exp\{\lambda t(y,\mu)\}, \quad y \in I\!R, \tag{3.1}$$

where a and t are given functions, $\lambda > 0$ and μ varies in an interval of the real line. We use the notation $Y \sim DM(\mu,\sigma^2)$ to denote (3.1), where $\sigma^2 = 1/\lambda$. Exponential dispersion models are a special case of (3.1), obtained by taking $t(y,\mu) = \theta y - \kappa(\theta)$, where $\mu = \kappa'(\theta)$.

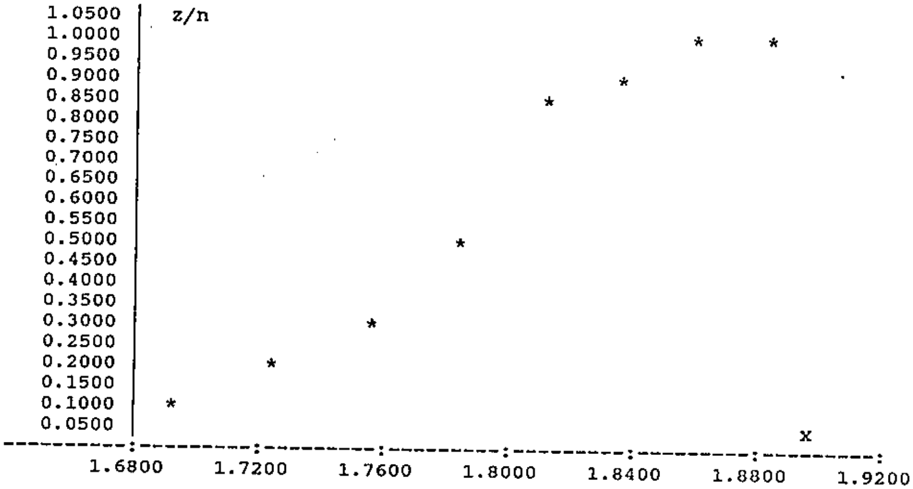

Figure 1.2 Plot of z_i/n_i against x_i for the data in Table 1.2.

By analogy with exponential dispersion models we call $\sigma^2 = 1/\lambda$ the *dispersion parameter*. Similarly, the parameter μ may generally be interpreted as a kind of location parameter, but μ is not necessarily the expectation of the distribution. The analogy with exponential dispersion models extends, as we shall see, to the form of the statistical analysis. Hence, dispersion models help to put exponential dispersion models into their proper perspective, and helps to bring ont the central ideas of analysis of deviance.

Example 1.3.1: *Dispersion models with a location parameter.* Let t be a given function on $\mathbb{R}$ such that

$$a(\lambda)^{-1} = \int \exp\{\lambda t(y)\}dy < \infty$$

for $\lambda > \lambda_0$, where $0 \leq \lambda_0 < \infty$. Then for every $\lambda > \lambda_0$ and $\mu \in \mathbb{R}$ the function

$$p(y; \mu, \lambda) = a(\lambda) \exp\{\lambda t(y - \mu)\} \tag{3.2}$$

is a probability density function on $\mathbb{R}$, and is a special case of (3.1). In this case μ is a location parameter and $\sigma^2 = 1/\lambda$ has a clear interpretation as a dispersion parameter, particularly if the function t is unimodal.

Some interesting special cases of (3.2) correspond to the form $t(y) = -|y|^\delta$ for suitable values of δ. For example $\delta = 2$ gives the normal distribution, and $\delta = 1$ gives the Laplace distribution. Another interesting case of (3.2) is obtained for $t(y) = -\log(1 + y^2)$, corresponding to

$$p(y; \mu, \lambda) = a(\lambda)(1 + (y - \mu)^2)^{-\lambda},$$

which is essentially Student's t-distribution with $2\lambda - 1$ degrees of freedom. ■

Example 1.3.2: *The von Mises-Fisher distribution.* Taking $t(y) = \cos y$ in (3.2) does not yield a distribution, but if we restrict the range of y to the interval $[0, 2\pi)$, we obtain the distribution

$$p(y; \mu, \lambda) = a(\lambda) \exp\{\lambda \cos(y - \mu)\}, \tag{3.3}$$

where y and μ belong to $[0, 2\pi)$. This distribution is known as the ***von Mises-Fisher distribution***, and it is denoted by the symbol $Y \sim vM(\mu, \sigma^2)$, where $\sigma^2 = 1/\lambda$. The observation y for (3.3) may be interpreted as an angle, and hence (3.3) may be useful for observations that lie on a circle, or more generally for directions. For $\lambda > 0$, the mode of (3.3) is μ. ■

Table 1.3: *Wind directions in degrees at Gorleston on Sundays in 1968 according to the four seasons*

Season	Wind directions in degrees									
Winter	50	120	190	210	220	250	260	290	290	320
	320	340								
Spring	10	20	40	60	160	170	200	220	270	290
	340	350								
Summer	10	10	20	20	30	30	40	150	150	150
	170	190	290							
Autumn	30	70	110	170	180	190	240	250	260	260
	290	350								

Example 1.3.3: *Wind directions 1.* Table 1.3 shows wind directions in degrees at Gorleston, England, at 11hr-12hr on Sundays in 1968 (Mardia, 1972). We let y_{ij} denote the j'th observation (in radians) for the i'th season, $i = 1, 2, 3, 4$. We assume that the corresponding random variables Y_{ij} follow the distribution (3.3) with parameters μ_i and $\sigma^2 = 1/\lambda$ where μ_i denotes the modal

wind direction for the i'th season, and σ^2 measures the expected amount of deviation from μ_i, assumed to be the same for all seasons. A statistical analysis of this data would probably include estimation of the parameters, assessment of the goodness of fit of the model, and a test of equality of the four μ_i's. ▮

If we make the transformation to $z = \exp(y)$ in (3.2), we obtain a dispersion model of the form

$$p(z;\rho,\lambda) = z^{-1}a(\lambda)\exp\{\lambda t(\log[z/\rho])\}, \quad z > 0. \tag{3.4}$$

This type of model is suitable for positive data, and in particular the parameter $\rho = \exp(\mu)$ is a scale parameter. The parameter $\sigma^2 = 1/\lambda$ continues to play the role of dispersion parameter for the model.

§1.4 Location and Scale Models

Let $Y_1, \ldots, Y_n$ be independent, and let Y_i have density function

$$p(y;\mu_i,\sigma^2) = f((y-\mu_i)/\sigma)/\sigma, \tag{4.1}$$

where f is a given density function on $\mathbb{R}$. Together with a systematic component for the location parameters $\mu_1, \ldots, \mu_n$, the resulting regression model is called a *location and scale model*. Here we consider the special case where the systematic component is given by the linear specification

$$\mu_i = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j, \quad i = 1, \ldots, n. \tag{4.2}$$

The *scale parameter* σ (or the parameter σ^2) is an example of a dispersion parameter, σ^2 being proportional to the variance of Y_i, if the variance is finite. If f is the standard normal density function, (4.1) and (4.2) reduce to a linear normal model.

As suggested by our notation, there are certain analogies between location and scale models, dispersion models, and exponential dispersion models. This is particularly so for the interpretation of the parameters, and this analogy extends to the form of the statistical analysis too.

The prototype application for a location and scale model is in measurements of physical distances. If we express the measurements in a new unit and add a constant, obtaining $Y_i' = sY_i + k$, say, then the new variables follow the same model with σ replaced by σs, β_0 replaced by $s\beta_0 + k$ and β_j replaced by $s\beta_j$. The model is thus able to absorb changes in location and scale of the measurements. This kind of adaptiveness of a statistical model is important because it ensures that the results of the statistical analysis, in particular parameter estimates, have a natural physical interpretation. We are going to show that exponential dispersion models have a different, but equally important type of physical interpretation.

An area where location and scale models have been used extensively is in survival analysis, where the observations are lifetimes of individuals or of components in a machine etc. If $Z_1, \ldots, Z_n$ are the lifetimes of n subjects, the model is defined by assuming that $Y_i = \log Z_i$, $i = 1, \ldots, n$, follow the location and scale model (4.1), (4.2). Hence, Z_i has a density function of the form

$$p(z_i, \delta_i, \sigma^2) = f(\{\log(z_i/\delta_i)\}/\sigma)/(z_i\sigma). \tag{4.3}$$

In this model $\delta_i = \exp(\mu_i)$ is a scale parameter, and σ^2 may be interpreted as a dispersion parameter.

Table 1.4: *Length of remission (weeks) in acute leukemia. A + denotes a censored observation*

Placebo	1 1 2 2 3 4 4 5 5 8 8 8 8 11 11 12 12 15 17 22 23
6-MP	6 6 6 6+ 7 9+ 10 10+ 11+ 13 16 17+ 19+ 20+ 22 23 25+ 32+ 32+ 34+ 35+

Example 1.4.1: *6-MP data.* The data in Table 1.4 shows the outcome of a clinical trial in which 6-mercaptopurine (6-MP) was compared with a placebo in the maintenance of remissions in acute leukemia (Gehan, 1965). Some observations were censored, meaning that only $\min\{Z_i, t_i\}$ is observed, where the t_i's are known censoring times. Ignoring for the moment the question of censoring, we assume that the lifetimes Z_i follow a Weibull distribution, which is equivalent to assuming that $Y_i = \log Z_i$ follows a location and scale model (4.1) corresponding to the extreme-value density

$$f(y) = \exp\{y - \exp(y)\}, \quad y \in \mathbb{R}. \tag{4.4}$$

We let μ_i be the location parameter for the i'th treatment, where $i = 1$ denotes placebo and $i = 2$ denotes 6-MP. Important questions for the analysis of this kind of data are to obtain a reliable estimate for the difference $\mu_1 - \mu_2$, say, between the two treatments, and to test the hypothesis $\mu_1 = \mu_2$. ▮

§1.5 Analysis of Deviance

Analysis of deviance is a technique for making inferences in regression models, analogous to the technique of analysis of variance for linear normal models. Analysis of deviance produces asymptotic versions of the familiar *t*- *F*- and χ^2-tests, with the deviance playing the role of the residual sum of squares.

Consider independent observations $Y_1, \ldots, Y_n$, with $Y_i \sim DM(\mu_i, \sigma^2)$, where $DM(\mu,\sigma^2)$ denotes the dispersion model (3.1). The *deviance* for the parameter $\mu = (\mu_1, \ldots, \mu_n)^T$ is defined by

$$D(\mathbf{y}, \mu) = 2\sum_{i=1}^{n}\{t(y_i, \tilde{\mu}_i) - t(y_i, \mu_i)\}, \tag{5.1}$$

where $\tilde{\mu}_i$ is the value of μ_i that maximizes $t(y_i, \mu_i)$ for the given observation y_i. If D_1 and D_2 are the minimized deviances for two nested hypotheses $H_1 \subseteq H_2$, respectively, the statistic

$$(D_2 - D_1)/(2\sigma^2)$$

is the log likelihood ratio test for H_2 under H_1 when σ^2 is known.

For σ^2 unknown, we are going to show that the statistic

$$F = \frac{(D_2 - D_1)/(f_2 - f_1)}{D_1/f_1} \tag{5.2}$$

is approximately $F(f_2 - f_1, f_1)$-distributed under H_2 for σ^2 small. This statistic, or modified versions of it, may hence be used for a test of H_2 under H_1 when σ^2 is unknown.

Example 1.5.1: *Wind directions 2*. For the von Mises-Fisher distribution we have $t(y, \mu) = \cos(y - \mu)$, which has maximum for $\mu = y$. Hence, the deviance is in this case

$$D(\mathbf{y}, \mu) = 2\{n - \sum_{i=1}^{n}\cos(y_i - \mu_i)\}.$$

For the data in Table 1.3, we let H_2 be the hypothesis that $\mu_1 = \mu_2 = \mu_3 = \mu_4$, and let H_1 be the hypothesis of arbitrary values for μ_1, μ_2, μ_3 and μ_4. The corresponding deviances are $D_1 = 69.4510$ and $D_2 = 86.2458$. Let us collect the results in an analysis of deviance table (Table 1.5). An estimate for σ^2 under H_1 is $D_1/45 = 1.5434$, which indicates that σ^2 may not be small enough for the F-approximation to (5.2) to apply. Keeping this in mind, we find that the value $F = 3.64$ in Table 1.5 corresponds to a p-value of about 0.025 in the $F(3,45)$-distribution, which suggests a difference in wind directions for the four seasons. ▌

Table 1.5: *Analysis of deviance for wind directions.*

Source	d.f.	Deviance	$\hat{\sigma}^2$	F
Between samples	3	$D_2 - D_1 = 16.7958$	5.5986	$F = 3.64$
Within samples	45	$D_1 = 69.4510$	1.5434	
Total	48	$D_2 = 86.2458$		

The parallel between analysis of deviance and the analysis of variance for normal data, which we have illustrated here for a dispersion model, also holds in the class of exponential dispersion models, this class being a special case of the class of dispersion models. A fact that helps to explain this parallel is that the models are parametrized such that the information matrix is block dioagonal, in particular making σ^2 and β orthogonal. One may generalize analysis of deviance to discrete exponential dispersion models and to location and scale models, and the successful generalization of analysis of deviance to these cases depends on obtaining a parametrization of the models' similar to the parametrization mentioned above. However, for the present we shall concentrate on dispersion models and exponential dispersion models.

§1.6 Notes

The development of the theory outlined in this chapter began with the paper by Nelder and Wedderburn (1972), introducing generalized linear models. The development was very much stimulated by the release of the computer program

GLIM (Baker and Nelder, 1978) for handling generalized linear models. The topic gradually became an active area of research, and around 1980, important papers began to appear, such as Whitehead (1980), Aitkin and Clayton (1980), Pregibon (1980, 1981), Morris (1982), McCullagh (1983) and West (1985). Good summaries of the development until 1983 may be found in the books by McCullagh and Nelder (1983), Dobson (1983) and Cordeiro (1986), the last book being in Portuguese.

Analysis of deviance was part of the theory of generalized linear models from the beginning (Nelder and Wedderburn, 1972), and has undergone various generalizations and perfections. Dispersion models and the analysis of deviance for these models were introduced by Sweeting (1981) and Jørgensen (1983), the former author working within a Bayesian framework.

Location and scale models have a long history, which we shall not try to trace here, but we mention the book by Kalbfleisch and Prentice (1981), which considers the use of location and scale models in the analysis of survival data. Analysis of deviance for location and scale models was proposed by Sweeting (1984).

Some general aspects of statistical modelling and data analysis, relevant for the discussion of regression models of the present chapter, and the book as a whole, are discussed by Cox and Snell (1981).

Exercises

Exercise 1.1: Give some examples of regression models that you have met before. If possible, include examples with continuous and discrete data, dependent and independent data, and linear and nonlinear models.

Exercise 1.2: Make a list of the types of statistical tests that you know. Divide the tests according to distribution (normal, χ^2- F- etc.).

Exercise 1.3: Using the methods you already know, analyse the data in Table 1.1. Give estimates of the parameters β_1, $\beta_2$and σ^2, including standard errors. Examine the goodness of fit of the model. Test the hypothesis $\beta_1 = 0$.

Exercise 1.4: Show that the gamma distribution with density

$$p(y; \psi, \lambda) = \psi^{\lambda}\Gamma(\lambda)^{-1}y^{\lambda-1}\exp(-\psi y), \quad y > 0,$$

is an exponential dispersion model, and write down the functions a and κ. Hint: Let $\psi = -\lambda\theta$.

Exercise 1.5: Show that the Poisson distribution is a discrete exponential dispersion model.

Exercise 1.6: Plot $g(z_i/\mu_i)$ as a function of x_i for the data in Table 1.2, using either the logit or the probit link for g. Estimate the parameters β_0 and β_1 from the plots, using for example linear regression.

Exercise 1.7: For $t(y) = -|y|^\delta$, derive the form of the function $a(\lambda)$ in (3.2). Solution: $a(\lambda) = \delta\lambda^{1/\delta}/(2\Gamma(1/\delta))$.

Exercise 1.8: For $t(y) = -|y|$, plot the probability density function (3.4) as a function of y for some values of λ and ρ. Use the expression for $a(\lambda)$ found in Exercise 1.7.

Exercise 1.9: For each season, make a histogram for the data in Table 1.3. If possible, take the circular nature of the observations into account (a "compass rose" type histogram).

Exercise 1.10: Choose a suitable function f, and plot (4.3) as a function of y for some values of δ_i and σ.

Exercise 1.11: Make a plot of the data in Table 1.4, showing the difference between placebo and 6-MP, taking into account the censored observations in an appropriate way.

Exercise 1.12: Plot the extreme-value density function (4.4) as a function of y.

Exercise 1.13: Show that the deviance for the normal distribution is $D(\mathbf{y}, \mu) = \sum(y_i - \mu_i)^2$.

Chapter 2

AN OUTLINE OF EXPONENTIAL DISPERSION MODELS

Exponential dispersion models are important statistical models, because they have a number of important mathematical properties, which are relevant in practice, and because they include a number of important distributions as special cases, giving a convenient general framework which includes a wide range of common statistical techniques. The present chapter outlines the basic properties of one-dimensional exponential dispersion models, suitable for the treatment of generalized linear models.

§2.1 Natural Exponential Families

Consider a family $\mathcal{P}$ of distributions defined by a density function of the form

$$g(z;\theta) = a(z)\exp\{\theta z - \kappa(\theta)\}, \quad z \in \mathbb{R} \tag{1.1}$$

for suitable functions a and κ. We consider the discrete case and the continuous case in parallel, so (1.1) is assumed to be a probability density function on $\mathbb{R}$ in the continuous case, and a probability function in the discrete case. Since the total probability mass is 1 we have

$$\int a(z)e^{\theta z}dz = \exp\{\kappa(\theta)\}, \tag{1.2}$$

where the integral is the Lebesgue integral in the continuous case, and the sum $\sum a(z)e^{\theta z}$ in the discrete case (this convention applies throughout the following). If $a(z)$ is the probability (density) function of a distribution P and κ is defined by (1.2), we say that (1.1) is *generated by* P. It is easy to see (Exercise 2.3) that the support of (1.1) does not depend on the value of θ.

The largest possible domain for the parameter θ is the interval

$$\Theta = \{\theta \in \mathbb{R}\colon \int a(z)e^{\theta z}dz < \infty\},$$

which is called the *canonical parameter domain* for $\mathcal{P}$. The parameter θ is called the *canonical parameter* for (1.1). The family $\mathcal{P}$ with θ varying in Θ is called a *natural exponential family* if the following two conditions are satisfied:

(i) The distribution (1.1) is not degenerate.

(ii) int Θ, the interior of Θ, is non-empty.

Using (1.2) we find that the moment generating function for (1.1) is

$$\begin{aligned} M(s;\theta) &= \int e^{sz} p(z;\theta)dz \\ &= \exp\{\kappa(\theta+s)-\kappa(\theta)\}, \quad s \in \Theta - \theta. \end{aligned} \tag{1.3}$$

Condition (ii) implies that (1.1) is characterized by its moment generating function (1.3).

The cumulant generating function corresponding to (1.3) is

$$\begin{aligned} K(s;\theta) &= \log M(s;\theta) \\ &= \kappa(\theta+s)-\kappa(\theta), \quad s \in \Theta - \theta. \end{aligned}$$

Hence, the ith cumulant $\kappa_i(\theta)$ of (1.1) is

$$\kappa_i(\theta) = K^{(i)}(0,\theta) = \kappa^{(i)}(\theta), \tag{1.4}$$

where $f^{(i)}$ denotes the i-th derivative of the function f. We call κ the *cumulant generator* for (1.1).

The expectation μ of (1.1) is

$$\mu = \kappa'(\theta), \quad \theta \in \text{int}\,\Theta.$$

The function $\tau(\theta) = \kappa'(\theta)$, which gives the relation between the canonical parameter θ and the expectation parameter μ is denoted the *mean value mapping*, and the image $\Omega = \tau(\text{int}\,\Theta)$ of int Θ by τ is called the *mean domain*.

The variance for (1.1) is $Var(Y) = \kappa''(\theta)$. By condition (i) we have $Var(Y) > 0$, and hence $\kappa''(\theta) > 0$ for $\theta \in \text{int}\,\Theta$. Consequently, τ is a strictly increasing function, which has an inverse τ^{-1}. We may thus express the variance of (1.1) in the form

$$\begin{aligned} Var(Y) &= \kappa''(\tau^{-l}(\mu)) \\ &= V(\mu), \quad \mu \in \Omega. \end{aligned}$$

The function V with domain Ω is called the *variance function* for $\mathcal{P}$. Note that V does not depend on the particular parametrization used in (1.1), V simply expresses how the variance behaves as a function of the mean μ.

The variance function plays an important role in the theory of exponential families and exponential dispersion models. A fundamental property of the variance function is that it characterizes the natural exponential family from which it comes, as shown by the following theorem.

Theorem 2.1.1. *The variance function V with domain Ω characterizes $\mathcal{P}$ within the class of all natural exponential families.*

Proof: Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two natural exponential families with the same variance function V with domain Ω. Let τ_i, Θ_i, κ_i, $i = 1, 2$, denote the mean value mapping, canonical parameter domain and cumulant generator, respectively, for $\mathcal{P}_i$, $i = 1, 2$, etc. Then

$$\frac{\partial \tau_i^{-1}}{\partial \mu} = \frac{1}{\tau_i'(\tau_i^{-1}(\mu))} = \frac{1}{V(\mu)}, \quad \mu \in \Omega.$$

Hence, there exists a constant $\theta_0 \in I\!R$ such that

$$\tau_1^{-1}(\mu) = \tau_2^{-1}(\mu) + \theta_0, \quad \mu \in \Omega$$

or, equivalently

$$\tau_1(\theta) = \tau_2(\theta - \theta_0), \quad \theta \in \Theta_1.$$

It follows that $\Theta_2 = \Theta_1 - \theta_0$. Since κ_i satisfies the equation

$$\kappa_i'(\theta) = \tau_i(\theta), \quad \theta \in \Theta_i,$$

we find that

$$\kappa_1(\theta) = k + \kappa_2(\theta - \theta_0), \quad \theta \in \Theta_1,$$

for some constant $k \in I\!R$. Hence, the cumulant generating function for a member of $\mathcal{P}_1$ is

$$\begin{aligned} K_1(s;\theta) &= \kappa_1(\theta + s) - \kappa_1(\theta) \\ &= \kappa_2(\theta - \theta_0 + s) - \kappa_2(\theta - \theta_0) \\ &= K_2(s; \theta - \theta_0), \quad s \in \Theta_2 - (\theta - \theta_0). \end{aligned}$$

Hence, the members of $\mathcal{P}_2$ have the same moment generating functions as the members of $\mathcal{P}_1$. By condition (ii), the moment generating functions $K_1(s;\theta)$ and $K_2(s;\theta - \theta_0)$ characterize their respective distributions. This implies $\mathcal{P}_1 = \mathcal{P}_2$. ∎

§2.2 Exponential Dispersion Models

2.2.1 The Discrete Case.

Let us recall the definition of an exponential dispersion model from Section 1.2. A discrete exponential dispersion model $\mathcal{P}$ is defined as a class of probability functions of the form

$$p^*(z;\theta,\lambda) = a^*(\lambda,z)\exp\{\theta z - \lambda\kappa(\theta)\}, \quad z \in \mathbb{N}_0. \tag{2.1}$$

This distribution is denoted $ED^*(\theta,\lambda)$. We observe that (2.1) has the form (1.1) for any given value of λ. We assume in the following that (2.1) is a natural exponential family for any given value of λ, that is, satisfies condition (i) and (ii) in section 2.1. Each of these natural exponential families have the same canonical parameter domain, denoted Θ, because the canonical parameter domain for (2.1) is the domain for the function $\lambda\kappa(\cdot)$, which is independent of the value of λ. Hence, the domain of variation for the parameter (θ,λ) in (2.1) is $\Theta \times \Lambda$, where Λ, called the *index set*, is the domain for λ.

One may show that either $\lambda \geq 0$ for all members of $\mathcal{P}$ or $\lambda \leq 0$ for all members of $\mathcal{P}$. Hence we adopt the convention that $\Lambda \subseteq \mathbb{R}_+$, where in particular the value $\lambda = 0$ has been excluded from Λ. We shall also adopt the convention that $1 \in \Lambda$. Apart from the exclusion of 0, these conventions imply no loss of generality, because they may be achieved by a change of sign of λ, followed by a scale transformation of λ.

All of the exponential dispersion models considered in the following have either $\Lambda = \mathbb{R}_+$ or $\Lambda = \mathbb{N}$. We have already seen an example of the latter case, namely the binomial distribution. In the case $\Lambda = \mathbb{R}_+$, the distribution is called *infinitely divisible*. We shall see in Section 2.5.2 that the negative binomial distribution is an example of an infinitely divisible discrete exponential dispersion model.

The moment generating function of (2.1) is

$$M^*(s;\theta,\lambda) = \exp[\lambda\{\kappa(\theta+s) - \kappa(\theta)\}], \quad s \in \Theta - \theta. \tag{2.2}$$

Note that the limiting form of (2.2) for $\lambda \to 0$ is $M \equiv 1$, which is the moment generating function of the degenerated distribution at 0. By (2.2), the ith cumulant for (2.1) is

$$\kappa_i(\theta,\lambda) = \lambda\kappa^{(i)}(\theta), \quad \theta \in \operatorname{int}\Theta.$$

Extending the terminology for natural exponential families, we call κ the *cumulant generator* for (2.1), and similarly, we use the notation

$$\begin{aligned} \mu &= \tau(\theta) = \kappa'(\theta), \\ V(\mu) &= \kappa''(\tau^{-1}(\mu)), \\ \Omega &= \tau(\operatorname{int}\Theta). \end{aligned}$$

In this notation, the expectation for (2.1) is

$$m = \lambda\mu = \lambda\tau(\theta), \quad \theta \in \operatorname{int}\Theta,$$

and the domain for m is $\lambda\Omega$. In most cases we have $\Omega = \mathbb{R}_+$ in the discrete case, which implies $\lambda\Omega = \mathbb{R}_+$ for any λ in Λ, in which case the domain for m does not depend on the value for λ.

The variance function for (2.1), as a natural exponential family with λ known, is $m \longmapsto \lambda V(m/\lambda)$, defined on $\lambda\Omega$. Hence, given the function V, we may derive the variance functions $\lambda V(m/\lambda)$, and hence by Theorem 2.1.1. we may reconstruct the natural exponential family (2.1) for any given value of λ. In this sense, the function V characterizes (2.1) and we shall call V the *variance function* of (2.1). Note, however, that V is determined from $\mathcal{P}$ only up to a scale transformation of λ, so a more precise statement would be that the family of functions $\lambda V(m/\lambda)$ characterizes (2.1) among all discrete exponential dispersion models.

2.2.2 The Continuous Case.

In Section 1.2 we defined a continuous exponential dispersion model to be a family of distributions with probability density functions of the form

$$p(y;\theta,\lambda) = a(\lambda, y)\exp[\lambda\{\theta y - \kappa(\theta)\}], \quad y \in \mathbb{R}. \tag{2.3}$$

Like in the discrete case, (2.3) is of the form (1.1) for λ known, and we assume that (2.3) satisfies condition (i) and (ii) in Section 2.1, so that (2.3) is a natural exponential family for any given known value of λ. Note that the canonical parameter for (2.3), viewed as a natural exponential family with λ known, is $\overline{\theta} = \theta\lambda$, whereas (2.1) has exponential family canonical parameter θ. By arguments similar to the discrete case, we find that the parameter (θ, λ) varies in a set of the form $\Theta \times \Lambda$, where, without loss of generality, we may assume $\Lambda \subseteq \mathbb{R}_+$ and $1 \in \Lambda$. Analogously to the discrete case, we call Λ the *index set* for (2.3), and we say that (2.3) is *infinitely divisible* if $\Lambda = \mathbb{R}_+$.

If we transform from y to $z = \lambda y$ in (2.3), we get a probability density function of the form

$$p^*(z;\theta,\lambda) = a^*(\lambda,z)\exp\{\theta z - \lambda\kappa(\theta)\}, \quad z \in \mathbb{R}, \tag{2.4}$$

where $a^*(\lambda,z) = \lambda^{-1}a(\lambda,z/\lambda)$. The density function (2.4) is essentially of the same form as the discrete exponential dispersion model (2.1), and for this reason, results for the discrete case may often be translated to the continuous case and vice versa. To stress this analogy, we use the notation $ED^*(\theta,\lambda)$ to denote either the discrete exponential dispersion model (2.1) or the continuous model (2.4). We shall use the common name *convolution family* about either (2.1) or (2.4). The reason for this terminology will become clear in Section 2.3.

Let us pause for a moment to consider the question why it is necessary to have different definitions of an exponential dispersion model in the discrete and the continuous case. The simplest answer is that in the discrete case, distributions like the binomial and the negative binomial are of the form (2.1), whereas in the continuous case, the normal distribution is of the form (2.3), and hence this distinction between the continuous and the discrete case is a useful one if we want a theory that encompasses the most useful standard discrete and continuous distributions. The second aspect of the question is that the form (2.3) leads to a much more elegant form of inference than does (2.1), so from this point of view (2.3) is to be preferred. However, there exist no nontrivial discrete models of the form (2.3), and hence one is forced to work with (2.1) in the discrete case. A simple argument for this statement is that if Z is a discrete variable with probability function of the form (2.1), then, although $Y = Z/\lambda$ has probability function essentially of the form (2.3), it is not a practically relevant model, because it has support $\{0, 1/\lambda, 2/\lambda, \dots\}$ which depends on the value for λ.

The main parallel between the discrete case and the continuous case is that the moment generating function of (2.4) has the form (2.2). By the transformation $y = z/\lambda$, we find that the moment generating function of (2.3) is

$$M(s;\theta,\lambda) = \exp[\lambda\{\kappa(\theta + s/\lambda) - \kappa(\theta)\}], \quad s \in \lambda(\Theta - \theta).$$

Hence the ith cumulant for (2.3) is

$$\kappa_i(\theta,\lambda) = \kappa^{(i)}(\theta)\lambda^{1-i}, \quad \theta \in \operatorname{int}\Theta.$$

Defining $\tau(\theta) = \kappa'(\theta)$, $V(\mu) = \kappa''(\tau^{-1}(\mu))$ and $\Omega = \tau(\operatorname{int}\Theta)$ by analogy with the discrete case, we may write the expectation for (2.3) as

$$\mu = \tau(\theta) \in \Omega, \quad \theta \in \operatorname{int}\Theta$$

and the variance as

$$\sigma^2 V(\mu), \quad \mu \in \Omega,$$

where V is called the *variance function* for (2.3) and $\sigma^2 = 1/\lambda$ is the dispersion parameter. Like in the discrete case, we find that V characterizes (2.3) among all continuous exponential dispersion models.

We shall denote the continuous exponential dispersion model (2.3) by the symbol $ED(\mu, \sigma^2)$, where $\mu = \tau(\theta)$ and $\sigma^2 = 1/\lambda$ are the parameters defined above. Note here that the expectation μ for (2.3) does not necessarily exist when θ is on the boundary of Θ. For most practical purposes this is not important, but formally speaking, the parameter μ is not defined for θ on the boundary of Θ. The relation between the notation $ED^*(\theta, \lambda)$ for (2.4) and $ED(\mu, \sigma^2)$ for (2.3) is that $Z \sim ED^*(\theta, \lambda)$ if and only if $Z/\lambda \sim ED(\mu, \sigma^2)$ where $\mu = \tau(\theta)$ and $\sigma^2 = 1/\lambda$.

§2.3 Convolution and Asymptotic Normality

Exponential dispersion models enjoy a remarkable convolution property, which generalizes the convolution property of the normal distribution. Consider a given convolution family $ED^*(\theta, \lambda)$ and assume that $Z_1, \ldots, Z_n$ are independent and

$$Z_i \sim ED^*(\theta, \lambda_i), \quad i = 1, \ldots, n,$$

for $(\theta, \lambda_i) \in \Theta \times \Lambda, \quad i = 1, \ldots, n$. By (2.2), the moment generating function of $Z. = Z_1 + \cdots + Z_n$ is

$$M^*_{Z.}(s; \theta, \lambda_1, \ldots, \lambda_n) = \exp[\sum_{i=1}^{n} \lambda_i \{\kappa(\theta + s) - \kappa(\theta)\}], \quad s \in \Theta - \theta.$$

Hence, we find

$$Z. \sim ED^*(\theta, \lambda_1 + \cdots + \lambda_n). \tag{3.1}$$

Example 2.3.1: *The binomial distribution.* If $Z_1, \ldots, Z_n$ are independent and $Z_i \sim Bi(\lambda_i, \mu)$, $i = 1, \ldots, n$, then by (3.1)

$$Z_1 + \cdots + Z_n \sim Bi(\lambda_1 + \cdots + \lambda_n, \mu),$$

because $\mu = \tau(\theta)$, and hence the θs are identical for $Z_1, \ldots, Z_n$, as required for (3.1). This is the standard convolution formula for the binomial distribution. ∎

As the example illustrates, formula (3.1) applies in particular to discrete exponential dispersion models. In the continuous case, let us assume that $Y_1, \ldots, Y_n$ are independent and

$$Y_i \sim ED(\mu, \sigma^2/w_i), \; i = 1, \ldots, n,$$

for given numbers $w_1, \ldots, w_n$, such that $w_i/\sigma^2 \in \Lambda$.Then

$$Z_i = \sigma^{-2} w_i Y_i \sim ED^*(\theta, \sigma^{-2} w_i),$$

where $\theta = \tau^{-1}(\mu)$. By (3.1) we get

$$Z. = Z_1 + \cdots + Z_n \sim ED^*(\theta, \sigma^{-2} w.),$$

where $w. = w_1 + \cdots + w_n$. Hence we have the convolution formula

$$\sum_{i=1}^{n} w_i Y_i / w. \sim ED(\mu, \sigma^2/w.), \tag{3.2}$$

because $Z.\sigma^2/w. \sim ED(\mu, \sigma^2/w.)$. As a special case of (3.2) we see that for $Y_1, \ldots, Y_n$ independent and identically distributed $ED(\mu, \sigma^2)$, we have

$$\frac{1}{n} \sum_{i=1}^{n} Y_i \sim ED(\mu, \sigma^2/n).$$

Hence a continuous exponential dispersion model is closed with respect to averaging identically distributed random variables, and more generally, as in (3.2), with respect to weighted averaging, where the weights are the reciprocals of the dispersion parameters.

Example 2.3.2: *The normal distribution.* If $Y_1, \ldots, Y_n$ are independent and $Y_i \sim N(\mu, \sigma^2/w_i)$, then by (3.2)

$$\sum_{i=1}^{n} w_i Y_i / w. \sim N(\mu, \sigma^2/w.),$$

where $w. = w_1 + \cdots + w_n$. This result is a special case of the more general convolution result for the normal distribution, which may be written as follows. If $Y_1, \ldots, Y_n$ are independent and $Y_i \sim N(\mu_i, \sigma^2/w_i)$, $i = 1, \ldots, n$, then

$$\sum_{i=1}^{n} w_i Y_i / w. \sim N(\sum_{i=1}^{n} w_i \mu_i / w., \sigma^2/w.). \quad \blacksquare \tag{3.3}$$

The result (3.3) raises the question whether a result parallel to (3.3) holds for the weighted average of variables with distribution $ED(\mu_i, \sigma^2/w_i)$. However, the result (3.3) relies on the fact that μ_i is a location parameter for the normal distribution. Thus, if we write

$$Y_i = \mu_i + E_i, \ i = 1, \ldots, n,$$

where $E_i \sim N(0, \sigma^2/w_i)$, we find that (3.3) may be obtained by applying (3.2) to $E_1, \ldots, E_n$, and using the linearity of the averaging operation. The normal distribution is not the only exponential dispersion model which is closed under translation, as we shall see in Section 2.8, but it is probably the only case in which a result like (3.3) holds.

The convolution results (3.1) and (3.2) have a number of important practical and theoretical consequences. For example, (3.1) implies that the index set Λ is closed with respect to addition. Since $1 \in \Lambda$ we conclude in particular that $\mathbb{N} \subseteq \Lambda$, in other words Λ contains all positive integers.

A second conclusion from (3.1) is that by the central limit theorem, an exponential dispersion model is approximately normal for λ large. Thus, for $Z \sim ED^*(\theta, \lambda)$ we have, for any fixed value of θ,

$$(3.4) \qquad (Z - m)/\lambda^{1/2} \xrightarrow{d} N(0, V(\mu)) \quad \text{for} \quad \lambda \to \infty,$$

where $m = \lambda\mu$ and $\xrightarrow{d}$ denotes convergence in distribution. This result applies in particular for a discrete exponential dispersion model. For the binomial distribution, (3.4) gives de Moivre-Laplace's Theorem. In the continuous case, we may write (3.4) in the form

$$(3.5) \qquad (Y - \mu)/\sigma \xrightarrow{d} N(0, V(\mu)) \quad \text{for} \quad \sigma^2 \to 0,$$

for $Y \sim ED(\mu, \sigma^2)$. For the normal distribution, this result is trivial.

The convolution results (3.1) and (3.2) allow an interpretation of an exponential dispersion model in terms of an underlying stochastic process. We illustrate this by two examples.

Example 2.3.3: *Energy expenditure data 2.* In Example 1.1.1 we considered the energy expenditure for human subjects. Suppose we divide the tissues of subject i into k compartments, homogeneous with respect to energy expenditure. Let $w_{i1}, \ldots, w_{ik}$ denote the masses of the k compartments, and let $Y_{i1}, \ldots, Y_{ik}$ be the corresponding energy expenditures, such that $w_i = w_{i1} + \cdots + w_{ik}$ is the total body mass of the subject and $Y_i = Y_{i1} + \cdots + Y_{ik}$ is the total energy expenditure of the subject. Inspired by (3.2), we assume that the average energy expenditure per unit of body mass follows an exponential dispersion model

$$(3.6) \qquad \overline{Y}_{ij} = Y_{ij}/w_{ij} \sim ED(\overline{\mu}_i, \sigma^2/w_{ij}),$$

where $\overline{\mu}_i$ is the theoretical energy expenditure per unit of body mass for the subject. If $Y_{i1}, \ldots, Y_{ik}$ are independent, then by (3.2)

$$(3.7) \qquad \overline{Y}_i = Y_i/w_i = \sum_{j=1}^{k} w_{ij}\overline{Y}_{ij}/w_i \sim ED(\overline{\mu}_i, \sigma^2/w_i).$$

The argument leading to (3.7) shows that, according to the model, the same form of statistical model may be assumed for an average energy expenditure, no matter whether the average is based on an entire individual, on parts of the tissues of an individual (if such a measurement were feasible) or on an aggregation of several individuals with the same body composition. This kind of adaptiveness of an exponential dispersion model ensures that the parameters have a physical interpretation, in particular for the interpretation of $\overline{\mu}_i$ as the average energy expenditure per unit of body mass.

The probabilistic interpretation of (3.6) and (3.7) is that we may view energy consumption as generated by a stochastic process, with the mass w playing the role of "time" for the stochastic process. Since weight is, in principle, a continuous variable, this requires the exponential dispersion model to have $\Lambda = I\!R_+$, in other words be infinitely divisible. ▮

Table 2.1: *Number of accidents on straight highways in Denmark 1963*

Season	1	2	3	4
two-lane	35	56	53	50
four-lane	199	184	227	270

Example 2.3.4: Table 2.1 shows the number of accidents on straight two-lane and four-lane highways in Denmark in 1963. The data are classified according to road type and season. Let Z_{ij} denote the number of accidents for road type i and season j, where $i = 1$ denotes two-lane and $i = 2$ denotes four-lane highway, and assume that the Z_{ij}s are independent. A possible model for this kind of data is a discrete exponential dispersion model of the form

$$(3.8) \qquad Z_{ij} \sim ED^*(\theta_{ij}, \lambda t_j),$$

where t_j is the length of the jth season of the year (hence the t_js are nearly equal). The expectation of Z_{ij} is $t_j m_{ij}$, where $m_{ij} = \lambda\tau(\theta_{ij})$ is the expected number of accidents per time unit for road type i and season j. The systematic component of the model could for example be the generalized linear model given by

$$\log m_{ij} = \mu + \alpha_i + \beta_j + \delta_{ij}. \tag{3.9}$$

A test of $\delta_{ij} = 0$ in (3.9) could be of interest, in order to be able to separate the effects of road type and season. For the present data, however, such a test would require λ to be known, because λ can not be estimated from (3.9), there being only one observation for each combination of i and j.

The convolution result (3.1) again allows an interpretation of the model (3.8) in terms of an underlying stochastic process. Specifically, consider a subdivision of the jth season into intervals of length t_{jk}, say for the kth interval, and let Z_{ijk} denote the number of accidents for the kth interval of the jth season for road type i. Assume that the Z_{ijk}s are independent and

$$Z_{ijk} \sim ED^*(\theta_{ij}, \lambda t_{ijk}),$$

such that in particular the expected number of accidents per time unit, m_{ij}, is constant within each season for each road type. Then formula (3.1) implies that $Z_{ij} = \sum Z_{ijk}$ has distribution (3.8), so that the model has the same kind of adaptiveness as we saw in the continuous case. Note that the model must again be infinitely divisible, because time is a continuous variable. ∎

§2.4 Continuous Exponential Dispersion Models

2.4.1 The Normal Distribution.

In this and the following two sections we summarize the basic properties of the three most important continuous exponential dispersion models, the normal, the gamma, and the inverse Gaussian distributions.

In Section 1.2 we saw that the probability density function of the normal distribution $N(\mu, \sigma^2)$ may be written in the form

$$p(y; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\{-y^2/(2\sigma^2) + \sigma^{-2}(y\mu - \mu^2/2)\},$$

which shows that we have an exponential dispersion model with parameter $(\theta, \lambda) = (\mu, \sigma^{-2})$ and cumulant generator $\kappa(\theta) = \theta^2/2$, $\theta \in I\!R$. As we saw earlier, the convolution formula (3.2) takes the form

$$\sum_{i=1}^{n} w_i Y_i / w. \sim N(\mu, \sigma^2/w.),$$

where $Y_1, \ldots, Y_n$ are independent, $Y_i \sim N(\mu, \sigma^2/w_i)$ and $w. = w_1 + \cdots + w_n$.

Since $\kappa'(\theta) = \theta$ and $\kappa''(\theta) = 1$, the variance function is

$$V(\mu) = 1, \ \mu \in I\!R,$$

which gives the well-known result $Var(Y) = \sigma^2$ for $Y \sim N(\mu, \sigma^2)$.

2.4.2 The Gamma Distribution.

The probability density function of the gamma distribution may be written in the form

$$\text{(4.1)} \qquad p^*(z; \psi, \lambda) = \Gamma(\lambda)^{-1} z^{\lambda-1} \exp(-\psi z + \lambda \log \psi), \quad z > 0,$$

where $\lambda, \psi > 0$. It follows that the gamma distribution is a convolution family of the form (2.4) with canonical parameter $\theta = -\psi < 0$ and cumulant generator $\kappa(\theta) = -\log(-\theta)$. We denote (4.1) by the symbol $Ga^*(\theta, \lambda)$.

Applying the convolution formula for convolution families, (3.1), to (4.1) we obtain the formula

$$\text{(4.2)} \qquad Ga^*(\theta, \lambda_1) * \cdots * Ga^*(\theta, \lambda_n) = Ga^*(\theta, \lambda_1 + \cdots + \lambda_n),$$

where $*$ denotes convolution. This is the standard convolution formula for the gamma distribution. In particular we have the relation

$$\chi^2(f) = Ga^*(-1/2, f/2)$$

with the χ^2-distribution, so that (4.2) gives the standard convolution formula for the χ^2-distribution.

An alternative way to write (4.1) is to take $-\psi = \lambda\theta$, which gives

$$\text{(4.3)} \qquad p(y; \theta, \lambda) = \frac{\lambda^\lambda}{\Gamma(\lambda)} y^{\lambda-1} \exp[\lambda\{\theta y + \log(-\theta)\}], \ y > 0,$$

where $\theta < 0$. Hence, we may also view the gamma distribution as a continuous exponential dispersion model, again with cumulant generator $\kappa(\theta) = -\log(-\theta)$.

We have $\kappa'(\theta) = -1/\theta$ and $\kappa''(\theta) = 1/\theta^2$. We denote (4.3) by the symbol $Ga(\mu, \sigma^2)$, where $\mu = -1/\theta$ is the expectation and $\sigma^2 = 1/\lambda$ is the dispersion parameter. The variance function for (4.3) is

$$V(\mu) = \mu^2, \quad \mu > 0, \tag{4.4}$$

which gives the variance $\sigma^2\mu^2$ for (4.3). In particular, σ is the coefficient of variation for (4.3). By (3.5) we have, letting $Y \sim Ga(\mu, \sigma^2)$,

$$(Y - \mu)/\sigma \xrightarrow{d} N(0, \mu^2) \quad \text{for} \quad \sigma^2 \to 0.$$

The gamma distribution may, as we have seen, be viewed as either a convolution family or an exponential dispersion model. This happens because the distribution is closed with respect to scale transformations. Thus, for $Y \sim Ga(\mu, \sigma^2)$ and $c > 0$ we have

$$cY \sim Ga(c\mu, \sigma^2). \tag{4.5}$$

In particular $\lambda Y \sim Ga(\lambda\mu, \sigma^2)$, which shows that the convolution family (2.4) corresponding to the exponential dispersion model (4.3) is again the family of gamma distributions. The connection between the two notations $Ga^*(\theta, \lambda)$ and $Ga(\mu, \sigma^2)$ is

$$Ga^*(\theta, \lambda) = Ga(-\lambda/\theta, 1/\lambda), \tag{4.6}$$

which may be obtained by comparing (4.1) and (4.3) Let us finally show how the convolution formula (3.2) looks for the gamma distribution. If $Y_1, \ldots, Y_n$ are independent and $Y_i \sim Ga(\mu, \sigma^2/w_i)$, (3.2) gives

$$\sum_{i=1}^{n} w_i Y_i / w. \sim Ga(\mu, \sigma^2/w.), \tag{4.7}$$

where $w. = w_1 + \cdots + w_n$. Obviously, (4.7) is equivalent to (4.2), which also follows directly from (4.5) and (4.6), because

$$w_i Y_i / w. \sim Ga(w_i\mu/w., \sigma^2/w_i) = Ga^*(-w./(\sigma^2\mu), w_i/\sigma^2).$$

2.4.3 The Inverse Gaussian Distribution.

The probability density function of the inverse Gaussian distribution has the form

$$p(y; \theta, \lambda) = (\frac{\lambda}{2\pi y^3})^{1/2} \exp[-\frac{\lambda}{2y} + \lambda\{\theta y + (-2\theta)^{1/2}\}], \quad y > 0, \tag{4.8}$$

where $\lambda > 0$ and $\theta \leq 0$. The fact that (4.8) is a probability density function is shown in Exercise 2.21. The form of (4.8) shows that we have a continuous exponential dispersion model with cumulant generator $\kappa(\theta) = -(-2\theta)^{1/2}$ and expectation $\mu = \kappa'(\theta) = (-2\theta)^{-1/2}$. We denote (4.8) by the symbol $IG(\mu, \sigma^2)$, where $\sigma^2 = 1/\lambda$ is the dispersion parameter.

Since $\kappa''(\theta) = (-2\theta)^{-3/2}$, the variance function of (4.8) is

$$V(\mu) = \mu^3, \ \mu > 0,$$

and the variance of (4.8) is $\sigma^2\mu^3$. By (3.2) we have, for $Y_1, \ldots, Y_n$ independent and $Y_i \sim IG(\mu, \sigma^2/w_i)$,

$$\sum_{i=1}^{n} w_i Y_i / w. \sim IG(\mu, \sigma^2/w.), \tag{4.9}$$

where $w. = w_1 + \cdots + w_n$. The distribution is asymptotically normal for σ^2 small,

$$(Y - \mu)/\sigma \xrightarrow{d} N(0, \mu^3) \quad \text{for} \quad \sigma^2 \to 0,$$

where $Y \sim IG(\mu, \sigma^2)$.

Like the gamma, the inverse Gaussian distribution is closed with respect to scale transformations. Thus, if $Y \sim IG(\mu, \sigma^2)$, then

$$cY \sim IG(c\mu, \sigma^2/c).$$

In particular, the inverse Gaussian distribution is also a convolution family. To explore this fact, let us take $\psi = \lambda\theta$ in (4.8), which gives

$$\begin{aligned} &p^*(z; \psi, \lambda) \\ &= (\frac{\lambda}{2\pi z^3})^{1/2} \exp\{-\frac{\lambda}{2z} + \psi z + \lambda^{1/2}(-2\psi)^{1/2}\}, \quad y > 0. \end{aligned} \tag{4.10}$$

Hence we have a convolution family of the form (2.4) with index parameter $\lambda^{1/2}$. We denote (4.10) by the symbol $IG^*(\psi, \lambda^{1/2})$. By (3.1) we obtain the alternative version of the convolution formula

$$IG^*(\psi, \lambda_1^{1/2}) * \cdots * IG^*(\psi, \lambda_n^{1/2}) = IG^*(\psi, \lambda_1^{1/2} + \cdots + \lambda_n^{1/2}). \tag{4.11}$$

For $\psi = 0$ in (4.10) we get the probability density function

$$p^*(z; 0, \lambda) = (\frac{\lambda}{2\pi z^3})^{1/2} \exp\{-\lambda/(2z)\}, \ z > 0. \tag{4.12}$$

This distribution is known as the positive stable distribution with stability index $\frac{1}{2}$. In fact, taking $\psi = 0$ and $\lambda_1 = \cdots = \lambda_n = \lambda$ in (4.11), we obtain the formula

$$IG^*(0, \lambda^{1/2}) + \cdots + IG^*(0, \lambda^{1/2}) = IG^*(0, n\lambda^{1/2}), \tag{4.13}$$

where the distribution $IG^*(0, n\lambda^{1/2})$ is n^2 times the distribution $IG^*(0, \lambda^{1/2})$, by the scale transformation property of the inverse Gaussian distribution. Hence, (4.13) is essentially the defining property of a stable distribution. See Exercise 2.25 for the definition af a stable distribution.

Table 2.2 summarizes the three continuous exponential dispersion models considered above and two models to be considered in Sections 2.6 and 2.8. The table assumes an exponential dispersion model of the form

$$p(y; \theta, \lambda) = a(\lambda, y) \exp[\lambda\{\theta y - \kappa(\theta)\}], \ y \in \mathbb{R}.$$

Table 2.2: *Some Continuous Exponential Dispersion Models*

$ED(\mu, \sigma^2)$	$a(\lambda, y)$	$\kappa(\theta)$	$\tau(\theta)$	$V(\mu)$	Θ	Ω
$N(\mu, \sigma^2)$	$(\frac{\lambda}{2\pi})^{1/2} e^{-\lambda y^2/2}$	$\frac{1}{2}\theta^2$	θ	1	$\mathbb{R}$	$\mathbb{R}$
$Ga(\mu, \sigma^2)$	$\frac{\lambda^\lambda y^{\lambda-1}}{\Gamma(\lambda)}$	$-\log(-\theta)$	$-\frac{1}{\theta}$	μ^2	$\mathbb{R}_-$	$\mathbb{R}_+$
$IG(\mu, \sigma^2)$	$(\frac{\lambda}{2\pi y^3})^{1/2} e^{-\lambda/(2y)}$	$-(-2\theta)^{1/2}$	$(-2\theta)^{-1/2}$	μ^3	$(-\infty, 0]$	$\mathbb{R}_+$
$GHS(\mu, \sigma^2)$	$\frac{\lambda 2^{\lambda-2} \lvert(\Gamma(\frac{\lambda}{2} + \frac{i\lambda y}{2})\rvert^2}{\Gamma(\lambda)\Gamma(\lambda/2)^2}$	$-\log\cos\theta$	$\tan(\theta)$	$1 + \mu^2$	$(-\frac{\pi}{2}, \frac{\pi}{2})$	$\mathbb{R}$
Stable, $\alpha = 1$	$-$	$\theta \log \theta - 1$	$\log \theta$	$e^{-\mu}$	$[0, \infty)$	$\mathbb{R}$

§2.5 Discrete Exponential Dispersion Models

2.5.1 The Binomial Distribution.

As we saw in Section 1.2, the binomial distribution is a discrete exponential dispersion model with probability function

$$\text{(5.1)} \qquad p^*(z;\theta,\lambda) = \binom{\lambda}{z} \exp\{\theta z - \lambda \log(1+e^\theta)\}, \; z = 0, 1, \ldots, \lambda,$$

where $\theta \in I\!R$ and $\lambda \in I\!N$. The cumulant generator is $\kappa(\theta) = \log(1+e^\theta)$, which gives $\mu = \kappa'(\theta) = e^\theta/(1+e^\theta)$ and $\kappa''(\theta) = e^\theta/(1+e^\theta)^2$. We use the notation $Bi(\lambda,\mu)$ for the binomial distribution (5.1), where, in the usual interpretation of the binomial distribution, λ is the number of trials, (5.1) gives the probability of exactly z successful trials, and μ is the probability of success for each trial.

For $Z \sim Bi(\lambda,\mu)$ we have

$$E(Z) = m = \lambda\mu$$

and, since the variance function is $V(\mu) = \mu(1-\mu)$,

$$Var(Z) = \lambda V(\mu) = \lambda\mu(1-\mu).$$

By (3.4) we have

$$(Z-m)/\lambda^{1/2} \xrightarrow{d} N(0, \mu(1-\mu)) \quad \text{for} \quad \lambda \to \infty,$$

which, as mentioned earlier, is de Moivre-Laplace's theorem. Finally, as noted above, the convolution formula (3.1) takes the form

$$Bi(\lambda_1,\mu) * \cdots * Bi(\lambda_n,\mu) = Bi(\lambda_1 + \cdots + \lambda_n, \mu)$$

for the binomial distribution.

2.5.2 The Negative Binomial Distribution.

Consider the negative binomial distribution, whose probability function is

$$\begin{aligned} p^*(z;\phi,\lambda) &= \binom{\lambda+z-1}{z}\phi^z(1-\phi)^\lambda \\ &= \binom{\lambda+z-1}{z}\exp\{\theta z+\lambda\log(1-e^\theta)\},\ z\in I\!N_0, \end{aligned} \tag{5.2}$$

where $\theta=\log\phi<0$. Hence we have a discrete exponential dispersion model with cumulant generator $\kappa(\theta)=-\log(1-e^\theta)$, $\mu=\kappa'(\theta)=e^\theta/(1-e^\theta)$ and $\kappa''(\theta)=e^\theta/(1-e^\theta)^2$. We denote the negative binomial distribution (5.2) by $Nb(\lambda,\mu)$. For $Z\sim Nb(\lambda,\mu)$ we have

$$m=E(Z)=\lambda\mu$$

and, since $V(\mu)=\mu(1+\mu)$,

$$Var(Z)=\lambda V(\mu)=\lambda\mu(1+\mu).$$

The convolution formula (3.1) takes the form

$$Nb(\lambda_1,\mu)*\cdots*Nb(\lambda_n,\mu)=Nb(\lambda_1+\cdots+\lambda_n,\mu),$$

which is the standard convolution formula for the negative binomial distribution. Finally, for $Z\sim Nb(\lambda,\mu)$ we have convergence to normality,

$$(Z-m)/\lambda^{1/2}\xrightarrow{d}N(0,\mu(1+\mu))\quad\text{for}\quad\lambda\to\infty.$$

2.5.3 The Poisson Distribution.

The Poisson distribution is a very special example of a discrete exponential dispersion model. First it has a dual interpretation as a natural exponential family and as an exponential dispersion model. Second, as we show in the next section, it appears as the limiting distribution in a general limit theorem for discrete exponential dispersion models.

The Poisson distribution $Po(m)$ with mean m has probability function

$$\begin{aligned} p(z;m) &= \frac{m^z}{z!}e^{-m} \\ &= \frac{1}{z!}\exp\{\phi z-e^\phi\},\quad z\in I\!N_0, \end{aligned} \tag{5.3}$$

where $\phi = \log m \in I\!R$. Hence the Poisson distribution is a natural exponential family of the form (1.1) with cumulant generator $\kappa(\phi) = e^{\phi}$. We have $m = \kappa'(\phi) = e^{\phi}$ and $\kappa''(\phi) = e^{\phi}$, and hence the variance function is $V(m) = m, \ m > 0$.

It is useful to consider the Poisson distribution from an alternative point of view, namely as a discrete exponential dispersion model, which will allow us to draw on the general theory for exponential dispersion models. Thus, letting $m = \lambda e^{\theta}$ in (5.3), we obtain

$$p^*(z;\theta,\lambda) = \frac{\lambda^z}{z!}\exp\{\theta z - \lambda e^{\theta}\}, \quad z \in I\!N_0, \tag{5.4}$$

which is a discrete exponential dispersion model with cumulant generator $\kappa(\theta) = e^{\theta}$. Thus we have $\mu = \kappa'(\theta) = e^{\theta}$ and $\kappa''(\theta) = e^{\theta}$, which gives, for $Z \sim Po(m)$,

$$E(Z) = m = \lambda\mu,$$

and since $V(\mu) = \mu$

$$Var(Z) = \lambda V(\mu) = \lambda\mu.$$

The convolution formula (3.1) takes the form

$$Po(\lambda_1 e^{\theta}) * \cdots * Po(\lambda_n e^{\theta}) = Po((\lambda_1 + \cdots + \lambda_n)e^{\theta}),$$

which is nothing more than the standard convolution formula $Po(m_1) * \cdots * Po(m_n) = Po(m_1 + \cdots + m_n)$. The convergence formula (3.4) takes the form

$$(Z - m)/\lambda^{1/2} \xrightarrow{d} N(0,\mu) \quad \text{for} \quad \lambda \to \infty,$$

for $Z \sim Po(m)$, $m = \lambda e^{\theta}$, which is the well - known convergence of the Poisson to normality for large values of the expectation.

Even though the Poisson distribution is formally a discrete exponential dispersion model, as in (5.4), we note that the parameter (θ, λ) is not identifiable; only the expectation λe^{θ} is identifiable. The next theorem shows that this property essentially characterizes the Poisson distribution.

Theorem 2.5.1. *Consider a discrete exponential dispersion model*

$$p^*(z;\theta,\lambda) = a^*(\lambda,z)\exp\{\theta z - \lambda\kappa(\theta)\}, \quad z \in I\!N_0. \tag{5.5}$$

Assume that $p^*(1;\theta,\lambda) > 0$ for some $(\theta,\lambda) \in \Theta \times \Lambda$, and assume that the index set is $\Lambda = I\!R_+$. If (5.5) yields the same natural exponential family for every $\lambda \in \Lambda$, then (5.5) is the Poisson distribution.

Proof: For λ known, (5.5) is a natural exponential family with variance function

$$m \longmapsto \lambda V(m/\lambda), \quad m \in \lambda\Omega. \tag{5.6}$$

If all the natural exponential families in (5.5) are identical, they have the same variance function, and in particular the domains of the variance function are the same. If the domain $\lambda\Omega$ does not depend on λ, Ω must be either $I\!R_-$, $I\!R_+$ or $I\!R$, but only $\Omega = I\!R_+$ is possible, because Ω is a subset of the convex support for (5.5), which is a subset of $I\!R_+$. Equating the values of (5.6) for $\lambda = 1$ and λ general we get

$$V(m) = \lambda V(m/\lambda), \quad m \in \Omega,$$

which for $m = 1$ and $\mu = 1/\lambda$ gives

$$V(\mu) = \mu V(1), \;\; \mu > 0,$$

where we have used the assumption $\Lambda = I\!R_+$. This is the variance function for the natural exponential family defined by $V(1)Y$, where $Y \sim Po(\mu)$ (see Exercise 2.27). By Theorem 2.1.1, one of these scaled Poisson distributions must be equal to (5.5). However, since we have assumed that $p^*(1;\theta,\lambda) > 0$ for some value of (θ,λ), the only possibility is $V(1) = 1$, so that (5.5) is the Poisson distribution. ▮

Table 2.3: *Some Discrete Exponential Dispersion Models*

$ED^*(\theta,\lambda)$	$a^*(\lambda,z)$	$\kappa(\theta)$	$\tau(\theta)$	$V(\mu)$	Θ	Ω
$Bi(\lambda,\mu)$	$\binom{\lambda}{z}$	$\log(1+e^\theta)$	$e^\theta/(1+e^\theta)$	$\mu(1-\mu)$	$I\!R$	$(0,1)$
$Nb(\lambda,\mu)$	$\binom{\lambda+z-1}{z}$	$-\log(1-e^\theta)$	$e^\theta/(1-e^\theta)$	$\mu(1+\mu)$	$I\!R_-$	$I\!R_+$
$Po(\lambda\mu)$	$\lambda^z/z!$	e^θ	e^θ	μ	$I\!R$	$I\!R_+$

To illustrate the practical aspects of the interpretation of the Poisson distribution as a discrete exponential dispersion model, consider the data in Example 2.3.4, where we assumed a discrete exponential dispersion model of the form

$$Z_{ij} \sim ED^*(\theta_{ij}, \lambda t_j).$$

In the special case of the Poisson distribution we get

$$Z_{ij} \sim Po(\lambda t_j e^{\theta_{ij}}) = Po(t_j m_{ij}),$$

where m_{ij} is the expected number of accident per time unit for the ith road type and the jth season. Hence, we have obtained the standard Poisson model, which corresponds to an underlying Poisson process with rate m_{ij}.

Table 2.3 summarizes the three discrete exponential dispersion models considered here. The table assumes a probability function of the form

$$p^*(z;\theta,\lambda) = \exp\{\theta z - \lambda\kappa(\theta)\}, \quad z \in I\!N_0.$$

2.5.4 Convergence of Discrete Exponential Dispersion Models to The Poisson Distribution.

A basic result in probability theory states that the Poisson distribution is a limiting case of the binomial distribution. Thus, if $Z \sim Bi(\lambda,\mu)$ we have

$$(5.7) \qquad Z \xrightarrow{d} Po(m) \quad \text{for} \quad \lambda \to \infty$$

for any fixed value of $m = \lambda\mu$. For the negative binomial distribution, $Z \sim Nb(\lambda,\mu)$, (5.7) also holds for any fixed value of $m = \lambda\mu$. Both of these results are special cases of the following result.

Theorem 2.5.2. *Consider a discrete exponential dispersion model*

$$(5.8) \qquad p^*(z;\theta,\lambda) = a^*(\lambda,z)\exp\{\theta z - \lambda\kappa(\theta)\}, \; z \in I\!N_0,$$

and assume that $p^*(z;\theta_0,\lambda_0) > 0$ for $z = 0,1$, for some $(\theta_0,\lambda_0) \in \Theta \times \Lambda$. If Z has distribution (5.8), then (5.7) holds for any fixed value of $m = \lambda\mu$, where $\mu = \kappa'(\theta)$.

Proof: We pretend to show that the moment generating function of (5.8) converges to the moment generating function of the Poisson distribution, which implies convergence in distribution. After a reparametrization, we may take $\theta_0 = 0$ and $\lambda_0 = 1$. We have

$$(5.9) \qquad \kappa(\theta) = \log\{\sum_{z=0}^{\infty} a_z \exp(\theta z)\},$$

where $a_z = a(1,z), z \in I\!N_0$. Since $\theta_0 = 0$ and $\lambda_0 = 1$ we have, by assumption, $a_0 > 0$ and $a_1 > 0$. Since the support of (5.8) is bounded to the left, we have $\inf \Theta = -\infty$. Now

$$\tau(\theta) = \kappa'(\theta) = e^{-\kappa(\theta)} \sum_{z=0}^{\infty} z a_z \exp(\theta z),$$

and since $e^{\kappa(\theta)} \to a_0 \neq 0$ for $\theta \to -\infty$, we obtain $\tau(\theta) \to 0$ for $\theta \to -\infty$. Hence, since τ is an increasing function, we conclude that $\inf \Omega = 0$.

Let $a \simeq b$ denote $\lim(a - b) = 0$. From (5.9) we have, using a Taylor approximation for log,

$$\begin{aligned} \kappa(\theta) &\simeq \log(a_0 + a_1 e^{\theta}) \\ &\simeq \log a_0 + a_1 a_0^{-1} e^{\theta} \quad \text{for } \theta \to -\infty. \end{aligned} \tag{5.10}$$

Hence

$$\tau(\theta) \simeq a_1 a_0^{-1} e^{\theta} \quad \text{for } \theta \to -\infty. \tag{5.11}$$

For $m = \lambda\tau(\theta)$ fixed we have

$$\theta = \tau^{-1}(m/\lambda) \to -\infty \quad \text{for } \lambda \to \infty. \tag{5.12}$$

Hence, the moment generating function for Z is, for $m = \lambda\tau(\theta)$ fixed and λ large,

$$\begin{aligned} M(s;\theta,\lambda) &= \exp[\lambda\{\kappa(\theta+s) - \kappa(\theta)\}] \\ &\simeq \exp\{\lambda a_1 a_0^{-1} e^{\theta}(e^s - 1)\} \\ &\simeq \exp\{m(e^s - 1)\}, \end{aligned}$$

where we have used (5.12), (5.10) and (5.11). Hence $M(s;\theta,\lambda)$ converges to the moment generating function of the Poisson distribution, which implies (5.7). ∎

§2.6 Quadratic Variance Functions

2.6.1 Linear Transformations and Variance Functions.

The variance function is an important tool for handling exponential dispersion models, as we saw for example in the proof of Theorem 2.5.1, due to the fact that an exponential dispersion model is characterized by its variance function.

The variance function furthermore seems to have a mathematically tractable form in most cases of interest, as illustrated by the examples in Sections 2.4 and 2.5. Encouraged by these facts, we proceed to study three classes of potential variance functions, namely quadratic variance functions and, in Section 2.7 and 2.8, power variance functions and exponential variance functions.

In order to study variance functions, we need to study linear transformations of exponential dispersion models. Consider the linear transformation

$$f(y) = \alpha + \beta y, \tag{6.1}$$

where $\beta \neq 0$. It is easy to see that if Y follows a continuous exponential dispersion model, then so does $U = f(Y)$, see Exercise 2.17. If Y has variance function V with domain Ω, then the variance of U is

$$\beta^2\sigma^2 V((\mu - \alpha)/\beta), \quad \mu \in \alpha + \beta\Omega, \tag{6.2}$$

where $\mu = E(U)$. Hence, the variance function of U is $\beta^2 V((\mu - \alpha)/\beta)$. In particular, the linear transformation (6.1) changes the domain of the variance function from Ω to $\alpha + \beta\Omega$. Furthermore, a reparametrization from σ^2 to $\sigma^2 c$, say, where $c > 0$, changes the variance function V to $c^{-1}V$. By combining a linear transformation and a reparametrization, we can hence translate, reflect and re-scale the variance function arbitrarily.

In the discrete case, we shall explore the relation between the convolution family (2.4) and the continuous exponential dispersion model (2.3). Let

$$p^*(z; \theta, \lambda) = a^*(\lambda, z)\exp\{\theta z - \lambda\kappa(\theta)\}, \quad z \in \mathbb{N}_0 \tag{6.3}$$

be a discrete exponential dispersion model. By transforming to $y = z/\lambda$, we obtain

$$p^*(\lambda y; \theta, \lambda) = a^*(\lambda, y\lambda)\exp[\lambda\{\theta y - \kappa(\theta)\}], \quad y \in \lambda^{-1}\mathbb{N}_0. \tag{6.4}$$

Note that the domain for y is $\lambda^{-1}\mathbb{N}_0$, which depends on λ. By the analogy between (6.4) and the continuous exponential dispersion model (2.3), we find that the form of (6.4) is preserved under linear transformations, except that the domain for y becomes a linear transformation of $\lambda^{-1}\mathbb{N}_0$. In particular, the variance function of (6.4) is $\sigma^2 V(\mu)$, for $\sigma^2 = 1/\lambda$ fixed, which is transformed into (6.2) by the transformation (6.1).

The transformed value of $z = \lambda y$ by the transformation (6.1) is $\lambda(\alpha + \beta y)$, and transforming back from (6.4) to (6.3) by the transformation $y \longmapsto z = \lambda y$, we obtain the transformation

$$z \longmapsto \lambda(\alpha + \beta z/\lambda) = \alpha\lambda + \beta z. \tag{6.5}$$

Hence, the transformation (6.5) is the parallel for the discrete case (6.3) to the transformation (6.1) for the continous case.

The reader should not become confused because of the fact that the transformation (6.5) depends on the parameter λ, because the results are to be used with Theorem 2.1.1. This theorem concerns natural exponential families, and in the present connection, an exponential dispersion model is simply a way to handle certain classes of natural exponential families indexed by λ.

The advantage of the transformation (6.5) is that the variance function V is transformed in the same way as in the continuous case. Thus, if (6.3) has variance function $\lambda V(m/\lambda)$, as a natural exponential family for λ known, then the transformation (6.5) produces the new variance function

$$m \mapsto \lambda\beta^2 V((m/\lambda - \alpha)/\beta).$$

In other words, $V(\mu)$ has been replaced by $\beta^2 V((\mu - \alpha)/\beta)$, exactly as in the continuous case.

These results allow us to make statements like: "up to a linear transformation, the only exponential dispersion model with a specified variance function is ...". In the discrete case it is understood here that any model obtained by applying the transformation (6.5) to (6.3) is called a discrete exponential dispersion model.

2.6.2 Classification of Quadratic Variance Functions.

By a quadratic variance function we mean a variance function

$$V(\mu) = a\mu^2 + b\mu + c, \quad \mu \in \Omega, \tag{6.6}$$

which is a polynomial of degree at most two. For simplicity, we let $V(\mu)$ be defined by (6.6) for any $\mu \in I\!R$. We saw in Sections 2.4 and 2.5 that the normal, Poisson, binomial, negative binomial and gamma distributions all have quadratic variance functions, cf. Tables 2.2 and 2.3. A sixth exponential dispersion model with quadratic variance function is the *generalized hyperbolic secant distribution*, which we describe in Section 2.6.3. We now show that up to a linear transformation, these six distributions are the only exponential dispersion models with quadratic variance functions.

In the following, c_1 and c_2 denote the two real roots of (6.6), if they exist, $d = -b/(2a)$ denotes the stationary point of (6.6), for $a \neq 0$, and $V(d)$ denotes the value of V at the stationary point.

(i) *The normal distribution* $(a = b = 0, c > 0)$

For $V(\mu) = c > 0$ constant, the corresponding exponential dispersion model is the normal distribution, which, up to a reparametrization, has variance function $V(\mu) = 1$, $\mu \in \mathbb{R}$.

(ii) *The Poisson distribution* $(a = 0, b \neq 0)$

The Poisson distribution has a linear variance function $V(\mu) = \mu$, $\mu \in \mathbb{R}_+$. By a linear transformation and a reparametrization, any variance function of the form $b\mu + c$ with $b \neq 0$ may be obtained. Hence, up to a linear transformation, the only exponential dispersion model with a linear variance function is the Poisson distribution.

(iii) *The binomial distribution* $(a < 0)$

For $a < 0$ we must have $V(d) > 0$ in order to obtain $V(\mu) > 0$ on $\Omega = (c_1, c_2)$. By a linear transformation we may obtain $c_1 = 0$ and $c_2 = 1$, which gives $V(\mu) = -a\mu(1-\mu)$. For $-a \in \mathbb{N}$ this is the variance function of the binomial distribution. One may show that $-a \notin \mathbb{N}$ is not possible.

(iv) *The negative binomial distribution* $(a > 0, V(d) < 0)$

In the case $a > 0$ there are three possibilities, $V(d) > 0$, $V(d) = 0$ and $V(d) < 0$. For $V(d) < 0$, V has two real roots $c_1 < c_2$. By a reflection we may obtain $\Omega = (c_2, \infty)$, and by a linear transformation we may obtain $c_1 = -1$ and $c_2 = 0$, which gives $\Omega = (0, \infty)$. Hence we have $V(\mu) = a\mu(1+\mu)$, and by a reparametrization, we obtain the variance function of the negative binomial distribution, $V(\mu) = \mu(1+\mu)$.

(v) *The gamma distribution* $(a > 0, V(d) = 0)$

For $a > 0$ and $V(d) = 0$, we have $c_1 = c_2 = d$. By a reflection we may obtain $\Omega = (d, \infty)$, and by a linear transformation we may obtain $V(\mu) = \mu^2, \quad \mu > 0$, which is the variance function of the gamma distribution.

(vi) *The generalized hyperbolic secant distribution* $(a > 0,\ V(d) > 0$

For $a > 0$ and $V(d) > 0$ we have $\Omega = \mathbb{R}$. By a translation we may obtain $d = 0$, and by a combination of a scale transformation and a reparametrization, we may obtain

$$V(\mu) = 1 + \mu^2, \quad \mu \in \mathbb{R}. \tag{6.7}$$

This is the variance function of the generalized hyperbolic secant distribution, which we study in the next section. With this distribution we have exhausted all possible exponential dispersion models with quadratic variance functions.

2.6.3 The Generalized Hyperbolic Secant Distribution.

To find the exponential dispersion model corresponding to (6.7), we follow the steps of the proof of Theorem 2.1.1. Thus, for V given by (6.7) we obtain the equation

$$\frac{\partial \tau^{-1}}{\partial \mu} = \frac{1}{1+\mu^2}, \quad \mu \in \mathbb{R},$$

which has solution

$$\tau(\theta) = \tan(\theta), \quad |\theta| < \pi/2,$$

where we have ignored the arbitrary constant, which, according to the proof of Theorem 2.1.1, does not affect the result. Next, we solve

$$\kappa'(\theta) = \tan(\theta), \quad |\theta| < \pi/2,$$

which, ignoring again the arbitrary constant, gives

$$\kappa(\theta) = -\log\{\cos(\theta)\}, \quad |\theta| < \pi/2.$$

The corresponding exponential dispersion model is continuous, has support $\mathbb{R}$, and on its convolution family form has probability density function

$$(6.8) \qquad p^*(z;\theta,\lambda) = a^*(\lambda,z)\exp[\theta z + \lambda \log\{\cos(\theta)\}], \quad z \in \mathbb{R},$$

where

$$\begin{aligned} a^*(\lambda,z) &= \frac{2^{\lambda-2}\,|\Gamma(\lambda/2+iz/2)|^2}{\Gamma(\lambda)\Gamma(\lambda/2)^2} \\ &= \frac{2^{\lambda-2}}{\Gamma(\lambda)} \prod_{j=0}^{\infty} \left\{ 1 + \left[\frac{z}{\lambda+2j}\right]^2 \right\}^{-1}, \quad z \in \mathbb{R}. \end{aligned}$$

The distribution is infinitely divisible, with index set $\Lambda = \mathbb{R}_+$.

In the special case $\lambda = 1$, (6.8) is related to the beta distribution. In fact, from Abramowitz and Stegun (1965, p. 256)

$$|\Gamma(1/2+iz/2)|^2 = \pi/\cosh(\pi z/2),$$

and hence

$$a^*(1,z) = 1/\{2\cosh(\pi z/2)\},$$

which implies

$$(6.9) \qquad p^*(z;\theta,1) = e^{\theta z}\cos(\theta)/\{2\cosh(\pi z/2)\}, \quad z \in \mathbb{R}.$$

By the transformation $z = \log\{u/(1-u)\}/\pi$, (6.9) turns into

$$f(u;\theta) = \pi^{-1}\cos(\theta)u^{-1/2+\theta/\pi}(1-u)^{-1/2-\theta/\pi}, \quad 0 < u < 1. \tag{6.10}$$

This is the probability density function of the beta distribution with parameters $1/2 + \theta\pi$ and $1/2 - \theta/\pi$. In fact, by the reflection formula (Abramowitz and Stegun, 1965, p.256).

$$\begin{aligned}\frac{\Gamma(1)}{\Gamma(1/2-\theta/\pi)\Gamma(1/2+\theta\pi)} &= \pi^{-1}\sin\{\pi(1/2+\theta/\pi)\}\\ &= \pi^{-1}\cos(\theta),\end{aligned} \tag{6.11}$$

which gives the relation between the cosine function and the beta function.

The probability density function (6.8) is symmetric and bell-shaped for $\theta = 0$ and skewed for $\theta \neq 0$. The tails of (6.9) decrease exponentially,

$$p^*(z;\theta,1) \simeq c\exp\{(\theta \mp \pi/2)z\} \quad \text{for } z \to \pm\infty. \tag{6.12}$$

Since $\Theta = (-\pi/2, \pi/2)$ is the domain of θ for any $\lambda > 0$, the tail behaviour of (6.8) for general λ is similar to (6.12). See Exercise 2.35 for an examination of the coefficients of skewness and kurtosis for (6.8).

For practical purposes, we shall often work with the exponential dispersion model corresponding to (6.8), obtained by the transformation $y = z/\lambda$. This distribution is denoted $GHS(\mu, \sigma^2)$, with μ and σ^2 denoting respectively the expectation and the dispersion parameter.

§2.7 Power Variance Functions

2.7.1 The Moment Generating Function.

A class of variance functions of considerable importance are the power variance functions, defined by

$$V(\mu) = \mu^p, \quad \mu \in \Omega, \tag{7.1}$$

where $\Omega = I\!R$ for $p = 0$ and $\Omega = I\!R_+$ otherwise. We have already seen a number of examples of exponential dispersion models with power variance functions, namely the normal ($p = 0$), Poisson ($p = 1$), gamma ($p = 2$) and inverse Gaussian distributions ($p = 3$), of which the first three variance function are also quadratic.

It is convenient to introduce the notation

$$\alpha = (p-2)/(p-1), \tag{7.2}$$

and in the following we denote quantities related to V in (7.1) by κ_α, τ_α, Θ_α etc., where α and p are related by (7.2), so that $p = (\alpha-2)/(\alpha-1)$. In particular we let $\alpha = -\infty$ denote the case $p = 1$. In this way (7.2) defines a one-to-one relation between $p \in I\!R$ and $\alpha \in (1,\infty) \cup [-\infty, 1)$.

To find the exponential dispersion model corresponding to (7.1), if it exists, we again follow the steps of the proof of Theorem 2.1.1. Thus, for the variance function (7.1) we find

$$\frac{\partial \tau_\alpha^{-1}}{\partial \mu} = \mu^{-p}, \quad \mu \in \Omega_\alpha \tag{7.3}$$

and

$$\kappa_\alpha'(\theta) = \tau_\alpha(\theta), \quad \theta \in \Theta_\alpha. \tag{7.4}$$

From (7.3) we find, for $\theta \in \Theta_\alpha$,

$$\tau_\alpha(\theta) = \begin{cases} (\frac{\theta}{\alpha-1})^{\alpha-1} & \text{for } \alpha \neq 1, -\infty \\ e^\theta & \text{for } \alpha = -\infty. \end{cases}$$

From τ_α we find κ_α by solving (7.4), which gives, for $\theta \in \Theta_\alpha$,

$$\kappa_\alpha(\theta) = \begin{cases} \frac{\alpha-1}{\alpha}(\frac{\theta}{\alpha-1})^\alpha & \text{for } \alpha \neq 0, 1, -\infty \\ -\log(-\theta) & \text{for } \alpha = 0 \\ e^\theta & \text{for } \alpha = -\infty, \end{cases} \tag{7.5}$$

where, in solving (7.3) and (7.4), we have ignored the arbitrary constants in the solutions.

The canonical parameter domain Θ_α is the largest interval for which κ_α is finite, whence

$$\Theta_\alpha = \begin{cases} I\!R & \text{for } \alpha = 2 \text{ or } \alpha = -\infty \\ [0,\infty) & \text{for } 1 < \alpha < 2 \text{ or } \alpha > 2 \\ (-\infty, 0) & \text{for } -\infty < \alpha \leq 0 \\ (-\infty, 0] & \text{for } 0 < \alpha < 1. \end{cases} \tag{7.6}$$

If an exponential dispersion model corresponding to (7.5) exists, the moment generating function of the corresponding convolution family is, for $s \in \Theta_\alpha - \theta$,

$$(7.7) \quad M_\alpha^*(s;\theta,\lambda) = \begin{cases} \exp[\frac{\lambda}{2-p}(\frac{\theta}{\alpha-1})^\alpha\{(1+s/\theta)^\alpha - 1\}] & \text{for } \alpha \neq 0, 1, -\infty \\ (1+s/\theta)^{-\lambda} & \text{for } \alpha = 0 \\ \exp\{\lambda e^\theta(e^s - 1)\} & \text{for } \alpha = -\infty, \end{cases}$$

where $\theta \neq 0$. Among the potential cumulant generators (7.5), we recognize for $p = 0$, 1, 2 and 3 the cumulant generators corresponding to respectively the normal ($\alpha = 2$), Poisson ($\alpha = -\infty$), gamma ($\alpha = 0$) and inverse Gaussian ($\alpha = 1/2$) distributions, and the corresponding moment generating functions are given by (7.7). In the next two sections we show that for $\alpha < 0$ ($1 < p < 2$), (7.7) is the moment generating function of a compound Poisson distribution, which is continuous with support $\mathbb{R}_+$, except for an atom at zero. For $0 < \alpha < 1$ ($2 < p < \infty$) and $1 < \alpha \leq 2$ ($p \leq 0$), we show that (7.7) is generated by an extreme stable distribution, which is continuous and has support $\mathbb{R}_+$, ($0 < \alpha < 1$) or $\mathbb{R}$, ($1 < \alpha \leq 2$). In Section 2.7.2 we show that the remaining cases, $\alpha > 2$, corresponding to $0 < p < 1$, do *not* correspond to exponential dispersion models. Table 2.4 summarizes all exponential dispersion models with power variance functions.

Before we proceed to the special cases, we derive convolution and scale transformation properties using the moment generating function. Let $ED_\alpha^*(\theta,\lambda)$ denote the distribution with moment generating function (7.7), if it exists. Then for $Z \sim ED_\alpha^*(\theta,\lambda)$ and $c > 0$ we have

$$(7.8) \qquad cZ \sim ED_\alpha^*(\theta/c, \lambda c^\alpha), \quad \alpha \neq -\infty, 1.$$

As we saw for the gamma and inverse Gaussian distributions, a result like (7.8) implies that the model may be viewed as both a convolution family and as an exponential dispersion model. The exponential dispersion model given by the moment generating function

$$M_\alpha(s;\theta,\lambda) = M_\alpha^*(s/\lambda;\theta,\lambda)$$

will be denoted $ED_\alpha(\mu,\sigma^2)$, where, as usual, $\mu = \tau_\alpha(\theta)$ and $\sigma^2 = 1/\lambda$. For $Y \sim ED_\alpha(\mu,\sigma^2)$, the equivalent of (7.8) is

$$(7.9) \qquad cY \sim ED_\alpha(c\mu, \sigma^2 c^{2-p}), \quad \alpha \neq 1,$$

where $c > 0$. The results (7.8) and (7.9) easily follow from the form of $M_\alpha^*(s;\theta,\lambda)$ and $M_\alpha(s;\theta,\lambda)$, respectively. Since (7.9) holds for any $c > 0$, any positive value for the dispersion parameter $\sigma^2 c^{2-p}$ is possible in (7.9) if $p \neq 2$. Hence, the domain of variation for $\lambda = 1/\sigma^2$ is $\Lambda = \mathbb{R}_+$, and the distributions are infinitely divisible. For $p = 2$, the gamma distribution, this argument is not valid, but we already know that the gamma distribution is infinitely divisible.

Table 2.4: *Exponential dispersion models with power variance functions*

Distribution	p	α	Support	Ω	Θ
Generated by extreme stable distributions	$p<0$	$1<\alpha<2$	$\mathbb{R}$	$\mathbb{R}_+$	$[0,\infty)$
Normal distribution	$p=0$	$\alpha=2$	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$
Not exponential dispersion models	$0<p<1$	$\alpha>2$	–	$\mathbb{R}_+$	$[0,\infty)$
Poisson distribution	$p=1$	$\alpha=-\infty$	$\mathbb{N}_0$	$\mathbb{R}_+$	$\mathbb{R}$
Compound Poisson distributions	$1<p<2$	$\alpha<0$	$\mathbb{R}_+^{1)}$	$\mathbb{R}_+$	$\mathbb{R}_-$
Gamma distribution	$p=2$	$\alpha=0$	$\mathbb{R}_+$	$\mathbb{R}_+$	$\mathbb{R}_-$
Generated by positive stable distributions	$2<p<3$	$0<\alpha<\frac{1}{2}$	$\mathbb{R}_+$	$\mathbb{R}_+$	$(-\infty,0]$
Inverse Gaussian distribution	$p=3$	$\alpha=\frac{1}{2}$	$\mathbb{R}_+$	$\mathbb{R}_+$	$(-\infty,0]$
Generated by positive stable distributions	$p>3$	$\frac{1}{2}<\alpha<1$	$\mathbb{R}_+$	$\mathbb{R}_+$	$(-\infty,0]$

1) Continuous for $y > 0$, with an atom at $y = 0$.

The distributions $ED_\alpha^*(\theta, \lambda)$ and $ED_\alpha(\mu, \sigma^2)$ obviously satisfy the convolution formulas (3.1) and (3.2), respectively. Let us recapitulate (3.1), which we need in the following, writing it in the form

$$ED_\alpha^*(\theta,\lambda_1) * \cdots * ED_\alpha^*(\theta,\lambda_n) = ED_\alpha^*(\theta,\lambda_1 + \cdots + \lambda_n). \tag{7.10}$$

2.7.2 Exponential Dispersion Models Generated by Extreme Stable Distributions.

We shall now investigate the case $\alpha > 0$, $\alpha \neq 1$, and show that for $0 < \alpha < 1$ and $1 < \alpha \leq 2$, (7.7) is generated by extreme stable distributions, and that for $\alpha > 2$, (7.7) is not a moment generating function. The reader may wonder at this point whether there exists an exponential dispersion model for $\alpha = 1$, in some sense. Actually there exist extreme stable distributions with stability index $\alpha = 1$, and in Section 2.8 we consider this case and show that it corresponds to exponential dispersion models with exponential variance functions.

Note first that by (7.6) we have $0 \in \Theta_\alpha$ for $\alpha > 0, \alpha \neq 1$. Hence, if (7.7) is a moment generating function, then in particular the distribution satisfies (7.10) with $\theta = 0$. For $\theta = 0$ and $\lambda_1 = \cdots = \lambda_n = \lambda$ in (7.10) we obtain

$$ED^*_\alpha(0, \lambda) * \cdots * ED^*_\alpha(0, \lambda) = ED^*_\alpha(0, n\lambda),$$

and by (7.8) the distribution $ED^*_\alpha(0, n\lambda)$ is the same as $n^{1/\alpha}\ ED^*_\alpha(0, \lambda)$. Hence, $ED^*_\alpha(0, \lambda)$ is a stable distribution with stability index α. However, from the theory of stable distributions it is known that α must belong to the interval $0 < \alpha \leq 2$. Hence, for $\alpha > 2$ $(0 < p < 1)$ there exist no exponential dispersion models with variance function μ^p. We state the result in the form of a theorem, indicating a proof that is relevant in our context.

Theorem 2.7.1. *There exist no exponential dispersion models with variance function $V(\mu) = \mu^p$ for $0 < p < 1$.*

Proof (sketch): Let $\alpha > 2$ be given, and assume that $M^*_\alpha(s; \theta, \lambda)$ is the moment generating function of a distribution, which we denote $ED^*_\alpha(\theta, \lambda)$. The variance of this distribution is

$$\lambda\kappa''_\alpha(\theta) = \lambda\{\theta/(\alpha - 1)\}^{\alpha-2}$$

which is zero for $\theta = 0$, where $0 \in \Theta_\alpha$, as we saw in Section 2.7.1. Hence $ED^*_\alpha(0, \lambda)$ is a degenerate distribution with moment generating function e^{sc} for some $c \in I\!R$. However, (7.7) is not the moment generating function of a degenerate distribution. Hence we have reached a contradiction, and we conclude that $M^*_\alpha(s; \theta, \lambda)$ is not a moment generating function for $\alpha > 2$. ▮

Turning now to the remaining cases $0 < \alpha < 1$ and $1 < \alpha \leq 2$, it may be shown that $M^*_\alpha(s; 0, \lambda)$ is the moment generating function of the extreme stable distribution with stability index α, as predicted by our previous analysis. This implies that $M^*_\alpha(s; \theta, \lambda)$ is a moment generating function for any $\theta \in \Theta_\alpha$, and hence we have justified the existence of exponential dispersion models with variance function $V(\mu) = \mu^p$ for $p > 2$ $(0 < \alpha < 1)$ and $p \leq 0$ $(1 < \alpha \leq 2)$.

There is apparently no closed-form expression for the probability density functions of the stable distributions, but series expansions of the densities exist. We distinguish between the cases $0 < \alpha < 1$ and $1 < \alpha < 2$ ($\alpha = 2$ is the normal distribution, which we have already considered).

$0 < \alpha < 1$

In this case the stable distributions are positive, and we shall refer to the distributions as positive stable, rather than extreme stable distributions. The continuous exponential dispersion model $ED_\alpha(\mu, \sigma^2)$ has probability density function

$$p(y; \theta, \lambda) = a_\alpha^*(\lambda, \lambda y)\lambda \exp[\lambda\{\theta y - \kappa_\alpha(\theta)\}] \tag{7.11}$$

for $y > 0$, where

$$a_\alpha^*(\lambda, y) = \frac{1}{\pi y}\sum_{k=1}^{\infty}\frac{\Gamma(1+\alpha k)}{k!}\lambda^k\kappa_\alpha^k(-y^{-1})\sin(-k\pi\alpha), \quad y > 0, \tag{7.12}$$

is the probability density function of the positive stable distribution with moment generating function $M_\alpha^*(s; 0, \lambda)$.

$1 < \alpha < 2$

In this case the distributions have support $\mathbb{R}$. The exponential dispersion model $ED_\alpha(\mu, \sigma^2)$ has probability density function (7.11) with

$$a_\alpha^*(\lambda, y) = \frac{1}{\pi y}\sum_{k=1}^{\infty}\frac{\Gamma(1+k/\alpha)}{k!}\{\frac{-y}{(\lambda\kappa_\alpha(1))^{1/\alpha}}\}^k\sin(-k\pi/\alpha), \quad y \in \mathbb{R}, \tag{7.13}$$

being the probability density function of the extreme stable distribution with moment generating function $M_\alpha^*(s; 0, \lambda)$. The results (7.12) and (7.13) may be proved by Fourier inversion, cf Feller (1971, p. 581).

2.7.3 Compound Poisson Distributions.

We shall now investigate the case $\alpha < 0$, corresponding to $1 < p < 2$, and show that it corresponds to a class of compound Poisson distributions.

Let $N, X_1, X_2, \ldots$ be a sequence of independent random variables, such that N is Poisson distributed $Po(m)$ and the X_is are identically distributed. Define

$$Z = \sum_{i=1}^{N} X_i \tag{7.14}$$

where Z is defined as 0 for $N = 0$. The distribution (7.14) is called a *compound Poisson distribution*. Now assume that

$$X_i \sim ED^*(\theta, \lambda), \quad i = 1, 2, \ldots,$$

for a given convolution family $ED^*(\theta,\lambda)$. The moment generating function of Z is

$$\begin{aligned} E(e^{sZ}) &= E\{E(e^{sZ}|N)\} \\ &= E\{M^*(s;\theta,\lambda)^N\} \\ &= \exp[m\{M^*(s;\theta,\lambda)-1\}], \end{aligned} \tag{7.15}$$

where $M^*(s;\theta,\lambda)$ is the moment generating function of $ED^*(\theta,\lambda)$. Note that, by (3.1), we have

$$Z|N=n \sim ED^*(\theta,n\lambda), \tag{7.16}$$

for $n \geq 1$.

By (7.7), the moment generating function $M^*_\alpha(s;\theta,\lambda)$ is, for $\alpha<0$,

$$M^*_\alpha(s;\theta,\lambda) = \exp[m\{(1+s/\theta)^\alpha - 1\}], \tag{7.17}$$

where

$$m = \frac{\lambda}{2-p}(\frac{\theta}{\alpha-1})^\alpha.$$

Taking

$$M^*(s;\theta,\lambda) = (1+s/\theta)^\alpha, \quad s<-\theta,$$

which by (7.7) is the moment generating function of the gamma distribution $Ga^*(\theta,-\alpha)$, in the notation of section 2.4.2, we find that (7.17) has the compound Poisson form (7.15). Hence we have shown that for $\alpha<0$, the distribution $ED^*_\alpha(\theta,\lambda)$ is a compound Poisson distribution (7.14), given by $X_i \sim Ga^*(\theta,\lambda)$ with $\lambda=-\alpha$.

By (7.16) we have, for $n \geq 1$

$$Z \mid N=n \sim Ga^*(\theta,-n\alpha), \tag{7.18}$$

and

$$P(Z=0) = \exp\{-\frac{\lambda}{2-p}(\frac{\theta}{\alpha-1})^\alpha\}. \tag{7.19}$$

Hence, the distribution $ED^*_\alpha(\theta,\lambda)$ has a positive probability, given by (7.19), for the outcome $Z=0$, whereas (7.18) shows that conditionally on the event

$\{Z > 0\}$ the distribution is continuous. The continuous part of the probability density function is, using (7.18)

$$\begin{aligned} p_\alpha^*(z;\theta,\lambda) &= \sum_{k=1}^{\infty} p^*(z;\theta,-k\alpha)\frac{m^k}{k!}e^{-m} \\ &= \sum_{k=1}^{\infty} \frac{(-\theta)^{-k\alpha}m^k z^{-k\alpha-1}}{\Gamma(-k\alpha)k!}\exp\{\theta z - m\} \\ &= \frac{1}{z}\sum_{k=1}^{\infty}\frac{\lambda^k\kappa_\alpha^k(-z^{-1})}{\Gamma(-k\alpha)k!}\exp\{\theta z - \lambda\kappa_\alpha(\theta)\}, \end{aligned} \tag{7.20}$$

where $p^*(z;\theta,\lambda)$ denotes the probability density function of the gamma distribution $Ga^*(\theta,\lambda)$, and κ_α is defined in (7.5). Defining

$$p_\alpha^*(0;\theta,\lambda) = \exp\{-\lambda\kappa_\alpha(\theta)\}, \tag{7.21}$$

the probability (7.19), we thus have a probability (density) function of the required form (2.4) for a convolution family, and it follows that the distribution $ED_\alpha^*(\theta,\lambda)$ satisfies all the requirements for a convolution family, and the general theory applies also in this case.

The convolution family (7.20), (7.21) may be transformed to an exponential dispersion model $ED_\alpha(\mu,\sigma^2)$, defined as the distribution of $Y = Z/\lambda$ where $Z \sim ED_\alpha^*(\theta,\lambda)$. The corresponding probability density function is given by

$$p_\alpha(y;\theta,\lambda) = a_\alpha(\lambda,y)\exp[\lambda\{\theta y - \kappa_\alpha(\theta)\}], \quad y \geq 0, \tag{7.22}$$

where

$$a_\alpha(\lambda,y) = \begin{cases} \frac{1}{y}\sum_{k=1}^{\infty}\frac{\lambda^k\kappa_\alpha^k(-(\lambda y)^{-1})}{\Gamma(-k\alpha)k!} & \text{for } y > 0 \\ 1 & \text{for } y = 0. \end{cases}$$

The reader will have noticed a certain similarity between the compound Poisson density function (7.20) and the probability density function (7.11), (7.12) for the exponential dispersion model generated by positive stable distributions. In fact, by the reflection formula $\Gamma(u)\Gamma(1-u) = \pi/\sin(\pi u)$, we find that (7.12) may be written in the form

$$a_\alpha^*(\lambda,y) = \frac{1}{y}\sum_{k=1}^{\infty}\frac{\lambda^k\kappa_\alpha^k(-y^{-1})}{\Gamma(-\alpha k)k!}, \quad y > 0,$$

which is analogous to (7.20) with $\theta = 0$.

The compound Poisson model (7.22) is a useful statistical model, because there are many practical examples of the kind of data it requires: positive and continuous, but with positive probability of a zero outcome. Theoretically, the distribution arises for data generated by an underlying compound Poisson process, corresponding to (7.14). Two examples will suffice to illustrate the idea. In insurance, we may think of Z, the yearly claim for an individual insurance holder, as given by (7.14), where N is the number of claims for the year and the X_is are the individual claims. In meteorology, the daily rainfall at a given site takes the form (7.14), where N is the number of rainfall periods during the day, and X_i is the amount of rainfall for the ith period. In cases like these, the form of (7.14) suggests that a compound Poisson model may provide an adequate model for the variable Z. If Z itself is observed, the convolution family form of the model (7.20) is appropriate, whereas if the average amount per unit of time is observed, the exponential dispersion model (7.22) is appropriate.

§2.8 Exponential Variance Functions

2.8.1 The Extreme Stable Distribution With Index 1.

In Section 2.7.2 we showed that extreme stable distributions with stability index $0 < \alpha < 1$ or $1 < \alpha \leq 2$ generate exponential dispersion models with power variance functions. We now examine the case $\alpha = 1$ and show that the corresponding extreme stable distribution generates an exponential dispersion model with variance function

$$V(\mu) = e^{-\mu}, \quad \mu \in \mathbb{R}. \tag{8.1}$$

In Section 2.8.2 we characterize this class of exponential dispersion models via its translation properties.

Let us consider the most general exponential variance function,

$$V(\mu) = \exp(a + b\mu), \quad \mu \in \mathbb{R}. \tag{8.2}$$

Introducing the dispersion parameter σ^2 we obtain $\sigma^2 V(\mu) = (\sigma^2 e^a)e^{b\mu}$, and hence we may take $a = 0$ without loss of generality. From (8.2) we find that $-\log(\sigma^2)/b$ is a location parameter for $b \neq 0$, in the sense that the exponential dispersion model corresponding to (8.2), if it exists, is closed with respect to translations, and the translation $y \to y + c$ changes σ^2 into $\sigma^2 e^{-bc}$. In particular, up to a scale transformation, (8.1) is the only exponential variance function. The reason for the negative sign in (8.1) is to simplify some formulas in the following.

Following the steps of the proof of Theorem 2.1.1, we find that if an exponential dispersion model corresponding to (8.1) exists, it has mean value mapping

$$\tau(\theta) = \log(\theta), \quad \theta \in \Theta$$

and cumulant generator

$$\kappa(\theta) = \theta\{\log(\theta) - 1\}, \quad \theta \in \Theta, \tag{8.3}$$

where $\Theta = [0, \infty)$. Note that κ is defined for $\theta = 0$ by continuity. If (8.3) corresponds to a convolution family $ED^*(\theta, \lambda)$, the corresponding moment generating function is

$$M^*(s; \theta, \lambda) = \exp[\lambda\{(\theta+s)\{\log(\theta+s)-1\} - \theta\{\log(\theta)-1\}\}], \quad s \in \Theta - \theta. \tag{8.4}$$

As for power variance functions, the connection with stable distribution is established via the scale transformation property of the moment generating function. For $Z \sim ED^*(\theta, \lambda)$, we find from (8.4) for $c > 0$

$$cZ \sim ED^*(\theta/c, c\lambda) + \lambda c \log c. \tag{8.5}$$

Applying (3.1) to $ED^*(\theta, \lambda)$, taking $\theta = 0$ and $\lambda_1 = \cdots = \lambda_n$, we find from (8.5)

$$ED^*(0, \lambda) * \cdots * ED^*(0, \lambda) = ED^*(0, n\lambda) = nED^*(0, \lambda) - \lambda n \log n. \tag{8.6}$$

Since the right-hand side of (8.6) is a linear transformation of $ED^*(0, \lambda)$, we find that this distribution, if it exists, is stable, with moment generating function

$$M^*(s; 0, \lambda) = \exp[\lambda\{s\{\log(s) - 1\}\}], \quad s \in \Theta.$$

One may show that this stable distribution exists, is continuous, and has support $I\!R$. More precisely, it is known as the extreme stable distribution with stability index $\alpha = 1$. By an argument similar to the argument used for the pover variance function models in Section 2.7.1, we find that (8.5) implies that the distribution is infinitely divisible.

Hence, we have shown that the variance function (8.1) corresponds to a continuous exponential dispersion model $ED(\mu,\sigma^2)$ with support $I\!R$. The corresponding moment generating function is $M(s;\theta,\lambda) = M^*(s/\lambda;\theta,\lambda)$, where $M^*(s;\theta,\lambda)$ is defined by (8.4). We claimed above that this exponential dispersion model is closed with respect to translation. In fact, if $Y \sim ED(\mu,\sigma^2)$ and $c \in I\!R$, then

$$c+Y \sim ED(c+\mu,\sigma^2 e^c). \tag{8.7}$$

This result may be obtained directly from the moment generating function, and as we saw above, the form of the result (8.7) follows directly from the form of the variance function. To be specific, with c and Y as above, we have

$$Var(Y) = \sigma^2 e^{-\mu},$$

and if we assume $c+Y \sim ED(c+\mu,\overline{\sigma}^2)$, then

$$Var(Y) = Var(c+Y) = \overline{\sigma}^2 e^{-c} e^{-\mu}.$$

Comparing the two expressions for $Var(Y)$ we obtain $\overline{\sigma}^2 = \sigma^2 e^c$, which confirms (8.7).

Taking $c = -\mu$ in (8.7) gives

$$Y = \mu + ED(0,\sigma^2 e^{-\mu}), \tag{8.8}$$

which displays the distribution $ED(\mu,\sigma^2)$ in its location parameter form. As we saw for the normal distribution, when μ is a location parameter, we may obtain a more general convolution result than the standard convolution result (3.2) for exponential dispersion models. Thus, if $Y_1,\ldots,Y_n$ are independent and

$$Y_i \sim ED(\mu_i,\sigma^2/w_i),$$

then by (3.2) and (8.8) we have

$$\frac{\sum w_i e^{\mu_i} Y_i}{w.} \sim ED(\overline{\mu},\sigma^2 \frac{e^{\overline{\mu}}}{w.}) \tag{8.9}$$

where

$$\overline{\mu} = \frac{\sum w_i e^{\mu_i}\mu_i}{w.}$$

and

$$w. = \sum w_i e^{\mu_i},$$

and where all summations are from $i = 1$ to n.

As predicted, we have hence obtained a more general convolution formula in this special case than the usual convolution formula (3.2). However, formula (8.9) uses the weights $w_i e^{-\mu_i}$, giving a complicated, although curious, formula. In any case, formula (8.9) confirms the uniqueness of the result (3.3) for the normal distribution.

2.8.2 Exponential Dispersion Models Closed Under Translation.

We have now seen that the exponential dispersion model corresponding to the variance function (8.1) is closed under translation, a property shared by the normal distribution. The following argument shows that, under certain conditions, these are the only exponential dispersion models closed under translation.

Assume that we have a continuous exponential dispersion model, denoted $ED(\mu, \sigma^2)$, with support $\mathbb{R} = \Omega$, closed under translation. If $Y \sim ED(\mu, \sigma^2)$ and $c \in \mathbb{R}$ we assume that there exists a function $f(c, \sigma^2)$ such that

$$Y + c \sim ED(\mu + c, f(c, \sigma^2)). \tag{8.10}$$

Note that the expectation parameter of (8.10) must be $\mu + c = E(Y + c)$. Since $Y \sim ED(\mu, \sigma^2)$ we have

$$Var(Y) = \sigma^2 V(\mu),$$

where $V(\mu)$ is the variance function of $ED(\mu, \sigma^2)$, and by (8.10) we have

$$Var(Y) = f(c, \sigma^2) V(\mu + c).$$

Taking $\sigma^2 = 1$, and writing $g(c) = f(c, 1)$, the two expressions for the variance of Y gives

$$V(\mu) = g(c) V(\mu + c). \tag{8.11}$$

Since V is differentiable and positive, g is also differentiable, and $g(0) = 1$. Differentiating (8.11) with respect to c we hence find

$$V'(\mu) = -g'(0) V(\mu). \tag{8.12}$$

The solution to (8.12) is

$$V(\mu) = k \exp\{-g'(0)\mu\}, \quad \mu \in \mathbb{R},$$

where k is a constant. This gives us the exponential variance function, which we investigated in Section 2.8.1. In particular, $g'(0) = 0$ gives the normal distribution.

We have hence found all exponential dispersion models that satisfy the translation formula (8.10) that is, models which are closed with respect to translation. We have not, however solved the more general problem where f is allowed to depend also on μ.

§2.9 Notes

The idea of an exponential dispersion model goes back to Tweedie (1947), who noticed many of the important properties and special cases of exponential dispersion models. However, Tweedie's paper remained virtually unnoticed for a long time, and in particular, Nelder and Wedderburn (1972) seem to have been unaware of Tweedie's paper. A third independent line of development of exponential dispersion models has taken place in the form of the study of certain exponential families of stochastic processes, see for example Küchler (1982) and references therein. A systematic study of the statistical properties of exponential dispersion models was initiated by Jørgensen (1986, 1987a), who introduced the term exponential dispersion model. The classification of quadratic variance functions is due to Morris (1982). Power variance functions were first studied systematically in the context of exponential dispersion models by Tweedie (1984). However, a number of authors have studied this idea independently of Tweedie, see for example Morris (1981), Hougaard (1986) and Bar-Lev and Enis (1986). Exponential variance functions were mentioned by McCullagh (1983), and the connection with extreme stable distributions mentioned in the discussion of Jørgensen (1987a).

Exercises

Exercise 2.1: Find the natural exponential family generated by the uniform distribution on (0,1).

Exercise 2.2: Show that the Poisson distribution is a natural exponential family, find the cumulants of the distribution, and find the variance function.

Exercise 2.3: Show that the support of a natural exponential family is independent of the canonical parameter θ.

Exercise 2.4: Show that the logarithmic distribution, given by

$$p^*(z;\rho) = \rho^z/\{-z\log(1-\rho)\}, \quad z = 1, 2, \ldots,$$

is a natural exponential family, and find its mean and variance. Answer: let $b(\rho) = -\log(1-\rho)$. Then $E(z) = \rho/\{b(\rho)(1-\rho)\}$ and $Var(z) = \rho\{1 - \rho/b(\rho)\}/\{b(\rho)(1-\rho)^2\}$.

Exercise 2.5: Make a plot of the variance function for the logarithmic distribution in Exercise 2.4. Hint: Plot the points $(\kappa'(\theta), \kappa''(\theta))$ for a suitable set of θ-values.

Exercise 2.6: Let Z be a discrete random variable, and define the probability generating function of Z by

$$q(u) = E(u^Z).$$

Find the relation between the probability generating function and the moment generating function of Z. Find the probability generating function for a discrete exponential dispersion model.

Exercise 2.7: Show that the exponential distribution is a natural exponential family, find the cumulants of the distribution and find the variance function.

Exercise 2.8: Let the probability density function a be defined by $a(z) = \exp\{-|z|\}/2$. Find the natural exponential family generated by a. Find the cumulants of this family, and its variance function.

Exercise 2.9: Show that the ith cumulant κ_i of a natural exponential family satisfies the relation

$$\kappa_{i+1} = V(\mu)\frac{\partial \kappa_i}{\partial \mu}, \quad i = 1, 2, \ldots,$$

where μ denotes the mean value parameter.

Exercise 2.10: Assume that X follows a natural exponential family. Show that the distribution of $Y = \alpha + \beta X, \beta \neq 0$, follows a natural exponential family, and find the cumulant generator and the variance function.

Exercise 2.11: Let $x_1 < x_2 < \cdots < x_k$ be given numbers in $\mathbb{R}$, and define the discrete uniform distribution on $\{x_1, \ldots, x_k\}$ by $a(x_i) = 1/k, \quad i = 1, \ldots, k$. Find the natural exponential family generated by $a(z)$, in particular the cumulant generator. Show that the canonical parameter domain is $\Theta = \mathbb{R}$ and that the mean domain is $\Omega = (x_1, x_k)$.

Exercise 2.12: Let the probability function $a(z)$ be defined by

$$a(z) = \begin{cases} \frac{1}{4} & \text{for} \quad z = \pm 1 \\ \frac{1}{2} & \text{for} \quad z = 0. \end{cases}$$

Find the natural exponential family generated by $a(z)$, and show that its variance function is $v(\mu) = (1-\mu^2)/2$, $|\mu| < 1$.

Exercise 2.13: Let X follow a beta distribution with parameters θ_1 and θ_2. Show that for θ_2 known, $Y = \log X$ follows a natural exponential family.

Exercise 2.14: The Hermite distribution is a discrete distribution defined by the probability generating function

$$q(u) = \exp\{a_1(u-1) + a_2(u^2-1)\},$$

where $a_1, a_2 > 0$. Show that this distribution is a discrete exponential dispersion model, find its variance function, and show that it is infinitely divisible. Hint: Find the discrete exponential dispersion model generated by the distribution given by $a_1 = a_2 = 1$.

Exercise 2.15: Assume that $Z = Z_1 + \cdots + Z_n$, where $Z_1, \ldots, Z_n$ are independent and identically distributed. Prove that if $P(|Z| \le c) = 1$ for some constant c, then $Var(Z) \le c^2/n$. Use this result to show that a convolution family with bounded support can not be infinitely divisible.

Exercise 2.16: Find the skewness and kurtosis of respectively a discrete exponential dispersion model and a continuous exponential dispersion model.

Exercise 2.17: Let Y follow a continuous exponential dispersion model. Show that $U = \alpha + \beta Y, \beta \neq 0$, follows a continuous exponential dispersion model, and find the cumulant generator of this model.

Exercise 2.18: Let Z follow a convolution family $ED^*(\theta, \lambda)$. Show that $V = \alpha\lambda + \beta Z$ follows a convolution family, and find the cumulant generator of this model.

Exercise 2.19: Make a statistical analysis of the energy expenditure data, using the model $\overline{Y}_i \sim N(\overline{\mu}_i, \sigma^2/w_i)$, $\overline{\mu}_i = \beta_1 \overline{x}_{i1} + \beta_2 \overline{x}_{i2}$.

Exercise 2.20: Show that the normal distribution is a convolution family, and relate the convolution formula (3.1) for this case to the standard convolution result for the normal distribution.

Exercise 2.21: Define, for $x > 0$,

$$F(x) = \Phi\{(\lambda/x)^{1/2}(x(-2\psi)^{1/2} - 1)\} \\ + \exp\{2\lambda(-2\psi)^{1/2}\}\Phi\{-(\lambda/x)^{1/2}(x(-2\psi)^{1/2} + 1)\},$$

where Φ is the standard normal distribution function. Show that F is a distribution function for every (ψ, λ) in $(-\infty, 0] \times (0, \infty)$, and show that the corresponding distribution is $IG^*(\psi, \lambda)$. Hint: Show that $F'(x)$ corresponds to the inverse Gaussian density function.

Exercise 2.22: Show that if $Z \sim IG^*(0, \lambda)$, then $\lambda/Z \sim \chi^2(1)$.

Exercise 2.23: Let $Y \sim Ga(\mu, \sigma^2)$. Show, by direct calculation, that the density function of the variable $W = (Y - \mu)/\sigma$ converges to the normal density $N(0, \mu^2)$ for $\sigma^2 \to 0$. Hint: Use Stirling's formula $\Gamma(\lambda) \simeq (2\pi)^{1/2}\lambda^{\lambda-1/2}e^{-\lambda}$.

Exercise 2.24: Let $Y \sim IG(\mu, \sigma^2)$. Show, by direct calculation, that the density function of the variable $W = (Y - \mu)/\sigma$ converges to the density of the normal distribution $N(0, \mu^3)$ for $\sigma^2 \to 0$.

Exercise 2.25: Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables. If there exists constants a_n and b_n such that for every $n > 0$ the two variables

$$X_1 + \cdots + X_n \text{ and } a_n X_1 + b_n$$

have identical distributions, then the distribution of X_1 is called *stable*. Show that the normal distribution is stable. Show that the distribution $IG^*(0, \lambda)$ is stable. One may show that $a_n = n^{1/\alpha}$ for some $\alpha \in (0, 2]$. The value of α is called the (stability) index of the stable distribution. Show that the stability index of the distribution $IG^*(0, \lambda)$ is $\frac{1}{2}$.

Exercise 2.26: Give an interpretation of the convolution formula (3.1) for the binomial and the negative binomial (with λ an integer) distributions in terms of a sequence of Bernoulli experiments.

Exercise 2.27: Assume that $Y \sim Po(\mu)$, and let $Z = cY$. Show that Z follows a natural exponential family, and show that the variance function of Z is $V(\mu) = c\mu, \quad \mu > 0$.

Exercise 2.28: Consider the Hermite distribution in Exercise 2.14. Show that this distribution converges to the Poisson distribution for $a_2 \to 0$.

Exercise 2.29: Consider the random variable S_n defined by

$$S_n = X_1 + \cdots + X_n - n,$$

where $X_1, X_2, \ldots$ are independent and identically distributed according to the logarithmic distribution in Exercise 2.4. Give conditions under which S_n converges to the Poisson distribution for $n \to \infty$. Hint: Show that S_n follows a discrete exponential dispersion model with dispersion parameter n.

Exercise 2.30: Let V be the variance function of a discrete exponential dispersion model. Show that $\lambda V(m/\lambda) \to m$ for $\lambda \to \infty$. Hint: use the proof of Theorem 2.5.2.

Exercise 2.31: Find the exponential dispersion model corresponding to the variance function $V(\mu) = \mu^2, \quad \mu < 0$.

Exercise 2.32: Find the exponential dispersion model corresponding to the variance function $V(\mu) = \lambda + \beta\mu$.

Exercise 2.33: Carry out in detail the transformation $z = \log\{u/(1-u)\}/\pi$ leading from (6.9) to (6.10).

Exercise 2.34: Consider the distribution in Exercise 2.12 with variance function $V(m) = (1-m^2)/2$. Determine to which category (i) to (vi) for quadratic variance functions this function belongs, and explain the relation with the corresponding standard distribution.

Exercise 2.35: Find the skewness and kurtosis for the generalized hyperbolic secant distribution.

Exercise 2.36: Find the exponential dispersion model corresponding to the variance function $V(\mu) = (\mu - \alpha)^3, \quad \mu > \alpha$, where $\alpha \in I\!R$ is a given constant.

Exercise 2.37: Let $Y \sim ED(\mu, \sigma^2)$, for a given exponential dispersion model, and assume that $\Omega = I\!R_+$. Show that if there exists a function f such that for any $c > 0$,

$$cY \sim ED(c\mu, f(c, \sigma^2)),$$

then the model has a power variance function.

Exercise 2.38: Explain the meaning of the scale transformation property (7.9) in the case of the Poisson distribution.

Exercise 2.39: Let $\kappa(\theta) = \theta \log \theta, \quad \theta \geq 0$. Find the corresponding variance function, and find the corresponding exponential dispersion model.

Exercise 2.40: Let X be a random variable with moment generating function $M(s)$. Show that $M(s) \geq 1 + sE(X)$ for $s \in I\!R$.

Chapter 3

ESTIMATION AND ANALYSIS OF DEVIANCE

Analysis of deviance is a method for making inferences in generalized linear models and other regression models, analogous to analysis of variance for linear normal models. The main body of the theory concerns dispersion models and continuous exponential dispersion models (Section 3.1-3.7). In Section 3.8 we consider inference for discrete exponential dispersion models.

§3.1 Inference For Regression Models

Consider a regression model with random component defined by

(i) $Y_1, \ldots, Y_n$ are independent random variables.

(ii) $Y_i \sim DM(\mu_i, \sigma^2/w_i), \quad i = 1, \ldots, n$, where $w_1, \ldots, w_n$ are known weights and $DM(\mu, \sigma^2)$ is a given dispersion model, defined by

$$p(y; \mu, \sigma^2) = a(\sigma^{-2}, y) \exp\{\sigma^{-2} t(y, \mu)\}, \quad y \in \mathbb{R}. \tag{1.1}$$

In particular, for a continuous exponential dispersion model $Y_i \sim ED(\mu_i, \sigma^2/w_i)$, we have

$$t(y, \mu) = y\tau^{-1}(\mu) - \kappa(\tau^{-1}(\mu)).$$

The domain of variation for the parameters of (1.1) is given by $(\mu, \sigma^{-2}) \in \Omega \times \Lambda$, where Ω is an interval. Unless otherwise stated, we assume $\Lambda = \mathbb{R}_+$.

We are mainly interested in hypotheses concerning μ, whereas σ^2 is either known, or is unknown and varies freely in Λ^{-1}. Consider three nested hypotheses, with parameters $\mu = (\mu_1, \ldots, \mu_n)^T$, β, and γ, respectively, defined by

$$
\begin{aligned}
&H_0: \mu \in \Omega^n \qquad \text{(the saturated hypothesis)}\\
&H_1: \mu = \mu(\beta), \quad \dim(\beta) = k_1\\
&H_2: \mu = \mu(\beta) \text{ and } \beta = \beta(\gamma), \quad \dim(\gamma) = k_2,
\end{aligned}
$$

where $k_2 < k_1 < n$.

Example 3.1.1: *Energy expenditure data 3*. In Section 2.3 we proposed the following model for the energy expenditure data

$$\overline{Y}_i \sim ED(\overline{\mu}_i, \sigma^2/w_i) \quad i = 1, \ldots, n$$

where the $\overline{Y}_i$s are independent. Define H_1 by

$$H_1\colon\ \overline{\mu}_i = \beta_1 \overline{x}_{i1} + \beta_2 \overline{x}_{i2}, \quad i = 1, \ldots, n$$

which is the main hypothesis for these data. A sub-hypothesis of H_1 of considerable interest is

$$H_2\colon \beta_2 = 0,$$

which states that the energy expenditure in fat tissue is zero. In the analysis of these data, we are interested in testing the goodness of fit of H_1 and a test for H_2 under H_1. ∎

In the above example there were just two hypotheses of real interest. In general one may be interested in a sequence of nested hypotheses $\Omega^n = H_0 \supseteq H_1 \supseteq H_2 \supseteq \ldots$, the H_is being subsets of Ω^n, each hypothesis representing a further restriction on the parameters of the model. In this case we proceed iteratively in the inference process. First, we check the goodness of fit of H_1, which is often done by an analysis of residuals. Then we test H_2 under H_1. If H_2 is accepted we test H_3 under H_2 and so on. When a significant result is obtained we either keep the last accepted model as the simplest model that explains the data, or we try an alternative sequence of models, possibly leading to a different final model. At this stage it may often be wise to take a look at the fit of the final model to see if some important aspect of the goodness of fit was overlooked in the first check.

In the following we consider the two hypotheses H_1 and H_2, as defined above, parametrized by respectively β and γ, this case being sufficiently general to explain the theory. The saturated hypothesis H_0, which has the same number of parameters as the number of observations, serves as a convenient reference, being the largest possible hypothesis for the given data.

§3.2 The Deviance

3.2.1 Definition of the Deviance.

To test hypotheses concerning μ, we shall mainly rely on the likelihood ratio test and various modifications of it. The models considered here allow the likelihood ratio test to be specified very conveniently via the deviance, which generalizes the sum of squares of residuals from normal theory.

Consider the regression model defined in Section 3.1, and let $\mathbf{y} = (y_1, \dots, y_n)^T$ represent the vector of observations. The log likelihood function for the parameters μ and σ^2 is

$$(2.1) \qquad L(\mu, \sigma^2) = \sum_{i=1}^{n} \log a(\sigma^{-2} w_i, y_i) + \sigma^{-2} \sum_{i=1}^{n} t(y_i, \mu_i) w_i.$$

The *deviance* for the parameter μ given the data vector $\mathbf{y}$ is defined by

$$(2.2) \qquad D(\mathbf{y}, \mu) = 2 \sum_{i=1}^{n} \{\sup_{\mu \in \Omega} t(y_i, \mu) - t(y_i, \mu_i)\} w_i.$$

Consider estimation under the hypothesis H_1. By (2.1) the maximum likelihood estimate for β, for σ^2 known, may be obtained by maximizing $\sum t(y_i, \mu_i(\beta)) w_i$ with respect to β, and does not depend on the value of σ^2. In particular, the estimate of β remains the same even if σ^2 is unknown, and we denote the estimate by $\hat{\beta}$. From (2.2) we see that $\hat{\beta}$ may be obtained by minimizing the deviance $D(\mathbf{y}, \mu(\beta))$ with respect to β.

The minimum value of the deviance under H_1 is

$$D_1 = D(\mathbf{y}, \mu(\hat{\beta})),$$

and similarly

$$D_2 = D(\mathbf{y}; \mu(\beta(\hat{\gamma})))$$

is the minimum value of the deviance under H_2, where $\hat{\gamma}$ is the maximum likelihood estimate of γ under H_2. We call D_1 and D_2 the *deviances of the* *hypotheses* H_1 and H_2, respectively. Note that the deviance is always non-negative, and D_1 and D_2 satisfy

$$0 \leq D_1 \leq D_2,$$

due to the nesting $H_0 \supseteq H_1 \supseteq H_2$. In particular, the deviance of H_0 is zero.

The deviance is closely related to the various log-likelihood ratio test statistics related to the hypotheses H_0, H_1 and H_2. In fact $D_1/(2\sigma^2)$ is the log-likelihood ratio statistic for H_1 under H_0, and similarly $D_2/(2\sigma^2)$ for H_2 under H_0. The statistic $(D_2 - D_1)/(2\sigma^2)$ is the log-likelihood ratio statistic for H_2 under H_1, assuming in all three cases that σ^2 is known.

Example 3.2.1: Most of the dispersion models considered in Section 1.3 has $t(y, \mu)$ of the form $t(y - \mu)$, as in Example 1.3.1. If furthermore $t(y)$ has maximum for $y = 0$, we obtain

$$\sup_{\mu \in \Omega} t(y - \mu) = t(y - y) = t(0),$$

which gives the deviance

$$D(\mathbf{y}, \mu) = 2\{w_. t(0) - \sum_{i=1}^{n} t(y_i - \mu_i) w_i\},$$

where $w_. = \sum w_i$. The von Mises-Fisher distribution has $t(y, \mu) = \cos(y - \mu)$ of this form, which gives the deviance

$$D(\mathbf{y}, \mu) = 2\{w. - \sum_{i=1}^{n} w_i \cos(y_i - \mu_i)\},$$

generalizing Example 1.5.1. ▮

Example 3.2.2: For the gamma distribution $Ga(\mu, \sigma^2)$ we have

$$\theta y - \kappa(\theta) = \theta y + \log(-\theta), \tag{2.3}$$

and hence

$$\frac{\partial}{\partial \theta}\{\theta y - \kappa(\theta)\} = y + 1/\theta.$$

If follows that the supremum of (2.3) is obtained for $\theta = -1/y$, and is

$$\sup_{\theta < 0}\{\theta y - \kappa(\theta)\} = -1 + \log(1/y).$$

Hence, the deviance is

$$\begin{aligned} D(\mathbf{y}, \mu) &= 2\sum_{i=1}^{n} w_i \left[-1 + \log(1/y_i) - \{-y_i/\mu_i + \log(1/\mu_i)\}\right] \\ &= 2\sum_{i=1}^{n} w_i\{-y_i/\mu_i + \log(\mu_i/y_i)\}. \end{aligned}$$ ▮

3.2.2 Steepness.

For the general discussion of the deviance we need the concept of steepness, defined as follows.

Definition 3.2.3: The exponential dispersion model $Y \sim ED(\mu, \sigma^2)$ is called *steep* if the interior of the convex support of Y is equal to the mean domain Ω for all values of λ. For a discrete exponential dispersion model $Z \sim ED^*(\theta, \lambda)$, the model is called steep if the interior of the convex support of Z/λ is equal to Ω for all parameter values. ▮

Example 3.2.4: The gamma distribution $Ga(\mu, \sigma^2)$ has support $\mathbb{R}_+$, which is a convex set. Hence the interior of the convex support is $\mathbb{R}_+$, which is equal to the mean domain Ω, and hence the gamma model is steep. ▮

It is equally simple to see that the normal and inverse Gaussian distributions are steep, see Exercise 3.7.

Steepness plays a role in the determination of the supremum entering in the deviance. Let us analyze the function

$$\ell(\theta) = \theta y - \kappa(\theta)$$

for a given exponential dispersion model with cumulant generator κ. We have

$$\ell'(\theta) = y - \tau(\theta), \quad \ell''(\theta) = -V(\mu).$$

Hence ℓ is strictly concave, because $\ell''(\theta) < 0$, and has a unique maximum. If the point $\tilde{\theta}$ where the maximum is obtained belongs to int Θ, then $\tilde{\theta}$ satisfies the equation

$$y = \tau(\tilde{\theta}), \tag{2.4}$$

and $\tilde{\theta}$ is unique. The solution to (2.4) exists if and only if $y \in \Omega = \tau(\text{ int } \Theta)$.

For a continuous steep model the boundary of the support has probability zero, and hence the solution to (2.4) exists with probability one in this case. In the discrete case, the boundary has positive probability, so in this case (2.4) may not have a solution in int Θ even if the model is steep. The discrete case will be discussed in Section 3.8.

In the continuous steep case, the solution to (2.4) is $\tilde{\theta} = \tau^{-1}(y)$, and hence the deviance is

$$D(\mathbf{y}, \mu) = 2\sum_{i=1}^{n} w_i[y_i\tilde{\theta}_i - \kappa(\tilde{\theta}_i) - \{y_i\theta_i - \kappa(\theta_i)\}],$$

where $\tilde{\theta}_i = \tau^{-1}(y_i)$ and $\theta_i = \tau^{-1}(\mu_i)$. As we have mentioned already, the three most important continuous exponential dispersion models are steep. Later we shall see that in the discrete case the binomial, negative binomial and Poisson distributions are also steep. However, it is not difficult to find examples of non-steep models, as the following example shows.

Example 3.2.5: Consider the exponential dispersion model with variance function $V(\mu) = \mu^p$ for $p < 0$ (Section 2.7.2). From Table 2.4 we have that the support is $\mathbb{R}$, whereas $\Omega = \mathbb{R}_+$, and hence this model is not steep. Since $\kappa_\alpha(\theta) = ((\alpha-1)/\alpha)(\theta/(\alpha-1))^\alpha$ we obtain

$$\frac{\partial}{\partial\theta}\{\theta y - \kappa_\alpha(\theta)\} = y - (\theta/(\alpha-1))^{\alpha-1}, \tag{2.5}$$

and the equation $y = (\theta/(\alpha-1))^{\alpha-1}$ has a solution only if $y > 0$. If $y \leq 0$, (2.5) is negative for all $\theta > 0$, and hence the supremum of $\theta y - \kappa_\alpha(\theta)$ is obtained for the value

$$\tilde{\theta} = \begin{cases} 0 & \text{for } y \leq 0 \\ (\alpha-1)y^{1/(\alpha-1)} & \text{for } y > 0. \end{cases}$$

In this case, the deviance hence takes the form

$$D(\mathbf{y},\mu) = 2\sum_{i=1}^n w_i \left[y_i\tilde{\theta}_i - \frac{\alpha-1}{\alpha}\{\tilde{\theta}_i/(\alpha-1)\}^\alpha \right.$$
$$\left. - \left\{ y_i(\alpha-1)\mu_i^{1/(\alpha-1)} - \frac{\alpha-1}{\alpha}\mu_i^{\alpha/(\alpha-1)} \right\} \right].$$

It follows that the high-order derivatives with respect to y of the deviance has a discontinuity at $y = 0$ (the boundary of Ω), a general feature for non-steep families. Maximum likelihood estimates also show non-standard behaviour in this case. Thus, if $y_1, \ldots, y_n$ are independent and identically distributed observations from the distribution considered here, then the maximum likelihood estimate for μ is

$$\hat{\mu} = \max\left\{0, \frac{1}{n}\sum_{i=1}^n y_i\right\}.$$

If μ is near 0, the distribution of $\hat{\mu}$ may be for from normal, depending on the value of σ^2, because of the discontinuity of the distribution of $\hat{\mu}$ at $\hat{\mu} = 0$. From a numerical point of view, the fact that the maximum likelihood estimate may fall on the boundary of the parameter space may cause problems for iterative maximum likelihood procedures. ▮

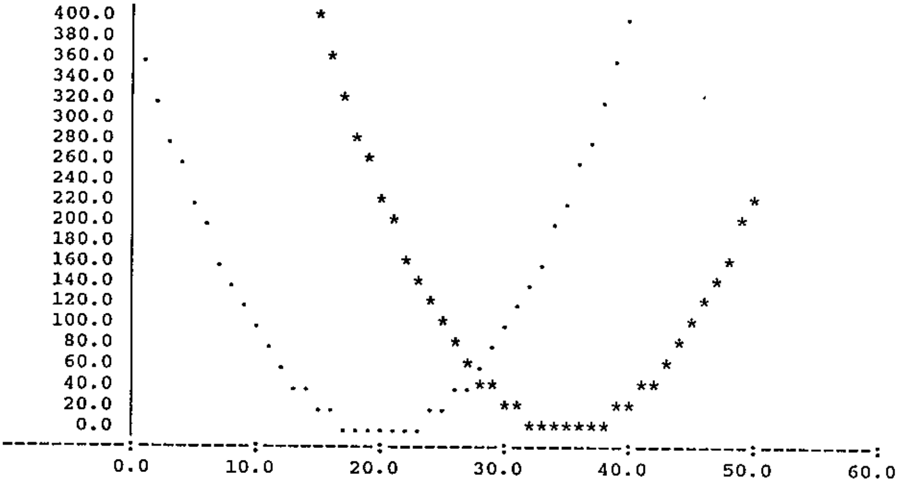

Figure 3.1 Plots of the deviance for the normal distribution

Table 3.1: *The deviance for some continuous exponential dispersion models* ($n = 1$ *and* $w = 1$)

Model	Deviance
$N(\mu, \sigma^2)$	$(y-\mu)^2$
$Ga(\mu, \sigma^2)$	$2\{y/\mu - 1 + \log(\mu/y)\}$
$IG(\mu, \sigma^2)$	$(y-\mu)^2/(\mu^2 y)$
$GHS(\mu, \sigma^2)$	$2\left[y\{\tan^{-1}(y) - \tan^{-1}(\mu)\} + \log\frac{\cos\{\tan^{-1}(y)\}}{\cos\{\tan^{-1}(\mu)\}}\right]$
Stable , $\alpha = 1$	$2\{e^{\mu}(\mu - 1 - y) + e^{y}\}$
$V(\mu) = \mu^p$ $(1 \leq p, p \neq 2)$	$2(\alpha-1)\{y^{\alpha/(\alpha-1)}(\alpha-1)/\alpha + \mu^{\frac{1}{(\alpha-1)}}(\mu/\alpha - y)\}$

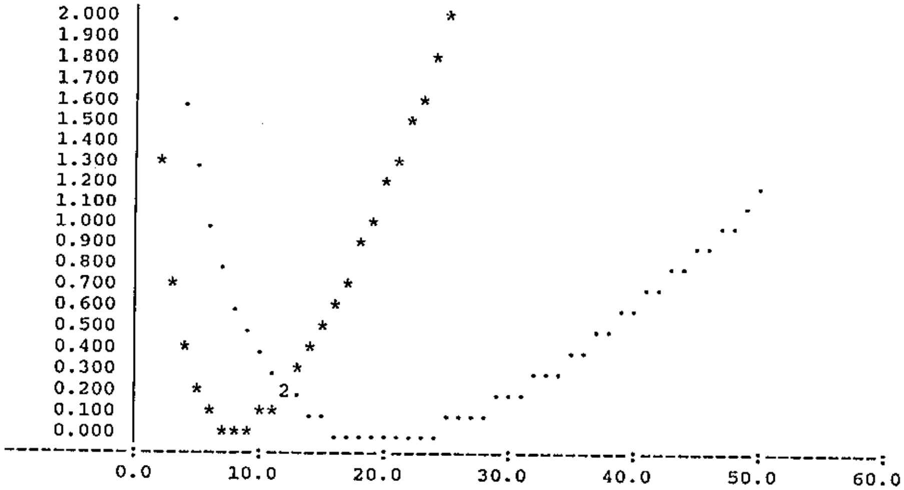

Figure 3.2 Plots of the deviance for the gamma distribution

To understand the role of the deviance as a measure of fit, it may be useful to plot the deviance of a single observation as a function y for a given value of μ. For the normal distribution, these plots are parabolas, whereas for other distributions the plots may be far from parabolic. Figure 3.1 and 3.2 shows such plots for the normal and the gamma distribution, respectively. Table 3.1 gives the form of the deviance for a number of continuous models.

§3.3 The Saddlepoint Approximation

3.3.1 The Saddlepoint Approximation for Exponential Dispersion Models.

The saddlepoint approximation is a numerical approximation to the probability density function of an exponential dispersion model. It is quite accurate in many cases, and has a simple statistical interpretation, because its main ingredients are the deviance and the variance function. We shall use it repeatedly in the following.

In the present section we consider a single observation y from the model $ED(\mu, \sigma^2)$, with deviance

$$D(y, \mu) = 2[y\tilde{\theta} - \kappa(\tilde{\theta}) - \{y\tau^{-1}(\mu) - \kappa(\tau^{-1}(\mu))\}],$$

where $\tilde{\theta}$ is the value of θ that maximizes $y\theta - \kappa(\theta)$. We assume that $ED(\mu, \sigma^2)$ is continuous, with probability density function

$$p(y; \theta, \lambda) = a(\lambda, y) \exp[\lambda\{\theta y - \kappa(\theta)\}], \quad y \in \mathbb{R},$$

where $\mu = \tau(\theta)$ and $\sigma^2 = 1/\lambda$.

Theorem 3.3.1. *For σ^2 tending to zero we have*

$$p(y; \theta, \lambda) \simeq \{2\pi\sigma^2 V(y)\}^{-1/2} \exp\{-D(y, \mu)/(2\sigma^2)\}, \quad y \in \Omega \tag{3.1}$$

which is called the saddlepoint approximation. The approximation is defined as zero outside Ω. The convergence in (3.1) is uniform in y on any compact subset of Ω.

A detailed proof of the theorem is outside the scope of the present text. For completeness, the main arguments of the proof are sketched in Section 3.3.3.

Example 3.3.2: For the gamma distribution, $Ga(\mu, \sigma^2)$ the probability density function is

$$p(y; \theta, \lambda) = \frac{\lambda^\lambda}{\Gamma(\lambda)} y^{\lambda-1} \exp[\lambda\{\theta y + \log(-\theta)\}], \quad y > 0. \tag{3.2}$$

Using Example 3.2.2 we find the saddlepoint approximation

$$\begin{aligned} p(y; \theta, \lambda) &\simeq (2\pi\sigma^2 y^2)^{-1/2} \exp\{-D(y, \mu)/(2\sigma^2)\} \\ &= \lambda^{1/2}(2\pi)^{-1/2} y^{-1} \exp[-\lambda\{-1 + \log(1/y)\} \\ &+ \lambda\{y\theta + \log(-\theta)\}] \\ &= \lambda^{1/2} e^{\lambda} (2\pi)^{-1/2} y^{\lambda-1} \exp[\lambda\{\theta y + \log(-\theta)\}]. \end{aligned} \tag{3.3}$$

Comparing with (3.2) we find that the saddlepoint approximation amounts to Stirling's approximation

$$\Gamma(\lambda) \simeq (2\pi)^{1/2} \lambda^{\lambda-1/2} e^{-\lambda} \tag{3.4}$$

applied to the gamma function in (3.2). This approximation is quite accurate, even for moderately small values of λ. For example, the relative error of (3.4) is about 1% for $\lambda = 8$. In the present example the relative error of the saddlepoint approximation is independent of y, and depends on λ and μ trough λ only. ▮

As shown in Exercise 3.9 and 3.10, the saddlepoint approximation is exact (i.e. is equal to the probability density function) for the normal and the inverse Gaussian distributions. In particular, we may write the inverse Gaussian distribution in the form

$$p(y;\theta,\lambda) = (2\pi\sigma^2 y^3)^{-1/2} \exp\{-(y-\mu)^2/(2\sigma^2\mu^2 y)\}, \quad y > 0,$$

where $(y-\mu)^2/(y\mu^2)$ is the deviance of the distribution. The saddlepoint approximation is exact after renormalization for the gamma distribution, in the sense that if we divide (3.3) by its integral with respect to y, we obtain the exact density (3.2). It may be shown (Daniels, 1980) that the normal, the inverse Gaussian and the gamma distributions are the only cases where the renormalized saddlepoint approximation is exact. Renormalization generally improves the accuracy of the saddlepoint approximation.

The saddlepoint approximation may be viewed as a refinement of the normal approximation given in Section 2.3. In particular, the deviance is approximately a quadratic form in y for σ^2 small. Thus, a quadratic expansion of $D(y,\mu)$ as a function of y around μ yields, for $y \in \Omega$,

$$D(y,\mu) \simeq (y-\mu)^2/V(\mu) \quad for \quad \sigma^2 \to 0. \tag{3.5}$$

This follows, because for $y \in \Omega$ we have $\tilde{\theta} = \tau^{-1}(y)$, and hence

$$\begin{aligned}\frac{\partial D}{\partial y} &= 2\{\tau^{-1}(y) + y/V(y) - y/V(y) - \tau^{-1}(\mu)\} \\ &= 2\{\tau^{-1}(y) - \tau^{-1}(\mu)\}\end{aligned}$$

and

$$\frac{\partial^2 D}{\partial y^2} = 2/V(y).$$

Since y converges to μ in probability as σ^2 tends to zero, we replace $V(y)$ by $V(\mu)$ in (3.1), and using (3.5), we obtain

$$p(y;\theta,\lambda) \simeq (2\pi\sigma^2 V(\mu))^{-1/2} \exp[-(y-\mu)^2/\{2\sigma^2 V(\mu)\}],$$

which is the probability density function of the normal distribution $N(\mu,\ \sigma^2\ V(\mu))$.

Note that for a steep model, the domain of y is Ω, and that in this case, the saddlepoint approximation (3.1) is defined throughout the domain of y. In the non-steep case, however, the situation is somewhat less satisfactory, because the saddlepoint approximation (3.1) is not defined on the set $\mathbb{R} - \Omega$, which has positive probability in this case. However, for σ^2 small, the distribution of y becomes concentrated near μ, and hence the probability that $y \notin \Omega$ becomes negligible for σ^2 small.

3.3.2 The Saddlepoint Approximation for Dispersion Models.

The saddlepoint approximation may be generalized to arbitrary dispersion models under certain conditions. The next theorem presents the simplest and most useful case for which the saddlepoint approximation my be derived.

Theorem 3.3.3. *Consider a dispersion model for the random variable Y, of the form*

$$p(y;\mu,\sigma^2) = a(\sigma^{-2})\exp\{t(y-\mu)/\sigma^2\}, \quad a < y < b, \tag{3.6}$$

where t satisfies the conditions

(i) $t(y)$ has global maximum at $y = 0$, and there exists a bounded neighbourhood ω of 0 such that t has no other stationary points in ω, and

$$\sup_{y \notin \omega} t(y) < t(0).$$

(ii) $t(y)$ is twice differentiable at $y = 0$.

Let $V = -t''(0)^{-1}$. Then

$$(Y-\mu)/\sigma \xrightarrow{d} N(0,V) \quad \text{for} \quad \sigma^2 \to 0 \tag{3.7}$$

and, for $\sigma^2 \to 0$ and $a < y < b$,

$$p(y;\mu,\sigma^2) \simeq (2\pi\sigma^2 V)^{-1/2} \exp\{-D(y,\mu)/(2\sigma^2)\}, \tag{3.8}$$

where $D(y,\mu) = 2\{t(0) - t(y-\mu)\}$ is the deviance. We call (3.8) the saddlepoint approximation to (3.6).

We sketch the proof of this theorem in Section 3.3.3.

Example 3.3.4: Consider the von Mises-Fisher distribution with probability density function

$$p(y;\mu,\sigma^2) = a(\sigma^{-2})\ \exp\{\cos(y-\mu)/\sigma^2\}, \quad 0 < y < 2\pi.$$

The condition of Theorem 3.3.3 are clearly satisfied in this case, and hence we have the saddlepoint approximation

$$p(y;\mu,\sigma^2) \simeq (2\pi\sigma^2)^{-1/2} \exp[\{\cos(y-\mu) - 1\}/\sigma^2]. \tag{3.9}$$

Hence, this amounts to the approximation

$$a(\sigma^{-2}) \simeq (2\pi\sigma^2)^{-1/2} \exp(-1/\sigma^2). \tag{3.10}$$

Let us write

$$a(\lambda) = \{2\pi I_0(\lambda)\}^{-1},$$

where I_0 is the modified Bessel function of the first kind and order 0. In terms of I_0, the approximation (3.9) amounts to the result

$$I_0(\lambda) \simeq (2\pi\lambda)^{-1/2} e^{\lambda},$$

which is known from the theory of Bessel functions, see e.g.. Abramowitz and Stegun (1972, p. 377). ∎

As for exponential dispersion models, the saddlepoint approximation may be viewed as a refinement of the normal approximation (3.7). Thus, expanding $t(\cdot)$ around 0 we obtain

$$D(y,\mu) \simeq (y-\mu)^2/V. \tag{3.11}$$

Inserting this in (3.8) we obtain the normal approximation (3.7). Note here that V plays the role of the variance function in the corresponding formula (3.5) for exponential dispersion models. We shall explore this analogy further in the following.

3.3.3 Derivation of The Saddlepoint Approximation.

We shall now sketch a proof of the saddlepoint approximation under various conditions. We begin with Theorem 3.3.1.

Proof of Theorem 3.3.1 (sketch)

(i) *The continuous case*

Assume that $\phi(t)$ is the characteristic function of a random variable Y. If $\phi(t)$ is absolutely integrable then, by the Fourier inversion Theorem, the probability density function of Y is

$$p(y) = (2\pi)^{-1} \int_{-\infty}^{\infty} \phi(t) e^{-ity} dt,$$

where i is the complex imaginary unit. For $Y \sim ED(\mu, \sigma^2)$, the characteristic function is

$$M(it; \theta, \lambda) = \exp[\lambda\{\kappa(\theta + it/\lambda) - \kappa(\theta)\}],$$

and hence the probability density function of Y is, for $\theta \in \operatorname{int} \Theta$,

$$\begin{aligned} p(y; \theta, \lambda) &= (2\pi)^{-1} \int_{-\infty}^{\infty} \exp[\lambda\{\kappa(\theta + it/\lambda) - \kappa(\theta)\} - ity] dt \\ &= \frac{\lambda}{2\pi} \int_{-\infty}^{\infty} \exp[\lambda\{\kappa(\theta + is) - \kappa(\theta) - isy\}] ds, \end{aligned} \tag{3.12}$$

where we have made the substitution $s = t/\lambda$. Since the integrand is analytic, we may move the path of integration from $(-\infty, \infty)$ to $i(\theta - \tilde{\theta}) + (-\infty, \infty)$, provided $\tilde{\theta} \in \operatorname{int} \Theta$. This gives

$$\begin{aligned} p(y; \theta, \lambda) &= \frac{\lambda}{2\pi} \int_{-\infty}^{\infty} \exp[\lambda\{\kappa(\tilde{\theta} + is) - \kappa(\theta) - (\tilde{\theta} + is)y + \theta y\}] ds \\ &= \frac{\lambda}{2\pi} \exp[\lambda\{\theta y - \kappa(\theta)\}] \int_{-\infty}^{\infty} \exp[\lambda\{\kappa(\tilde{\theta} + is) - (\tilde{\theta} + is)y\}] ds. \end{aligned} \tag{3.13}$$

We now choose $y \in \Omega$ and let $\tilde{\theta} = \tau^{-1}(y) \in \operatorname{int}\Theta$. By a Taylor expansion around $\tilde{\theta}$ of the exponent in the integrand of (3.13), we obtain

$$\begin{aligned} \kappa(\tilde{\theta} + is) - (\tilde{\theta} + is)y &\simeq \kappa(\tilde{\theta}) - \tilde{\theta} y + \frac{1}{2}(is)^2 \kappa''(\tilde{\theta}) \\ &= \kappa(\tilde{\theta}) - \tilde{\theta} y - \frac{1}{2} s^2 V(y). \end{aligned}$$

Introducing this approximation in (3.13), and using the definition of the deviance, we obtain

$$\begin{aligned} p(y; \theta, \lambda) &\simeq \frac{\lambda}{2\pi} \exp\{-\lambda D(y, \mu)/2\} \int_{-\infty}^{\infty} \exp\{-\lambda V(y) s^2/2\} ds \\ &= [\lambda/\{2\pi V(y)\}]^{1/2} \exp\{-\lambda D(y, \mu)/2\} \end{aligned}$$

for $y \in \Omega$, which is the saddlepoint approximation (3.1).

(ii) *The discrete case.*

In the discrete case the inversion formula takes the form

$$p(z) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \phi(t)e^{-itz}dt,$$

where $p(z)$ is the probability function of the distribution with characteristic function $\phi(t)$. For a discrete exponential dispersion model we hence have

$$p(z;\theta,\lambda) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \exp[\lambda\{\kappa(\theta+it)-\kappa(\theta)\} - itz]dt.$$

By the analyticity and periodicity of the integrand, proceeding by analogy with the continuous case, we may change the path of integration to $(-\pi,\pi)+i(\theta-\tilde{\theta})$. Hence, for $z/\lambda = \tau(\tilde{\theta}) \in \Omega$, we obtain

$$\begin{aligned} p(z;\theta,\lambda) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \exp[\lambda\{\kappa(\tilde{\theta}+it)-\kappa(\theta)\} - (\tilde{\theta}+it)z + \theta z]dt \\ &\simeq \frac{1}{2\pi}\exp\{-\lambda D(z/\lambda,\mu)/2\}\int_{-\pi}^{\pi} \exp\{-\lambda V(z/\lambda)t^2/2\}dt \\ &\simeq \frac{1}{2\pi}\exp\{-\lambda D(z/\lambda,\mu)/2\}\int_{-\infty}^{\infty} \exp\{-\lambda V(z/\lambda)t^2/2\}dt \\ &= \{2\pi\lambda V(z/\lambda)\}^{-1/2}\exp\{-\lambda D(z/\lambda,\mu)/2\} \end{aligned} \tag{3.14}$$

for σ^2 tending to zero. This is the saddlepoint approximation in the discrete case. The approximation is uniform in $y = z/\lambda$ on any compact subset of Ω.

Proof of Theorem 3.3.3 (sketch)

Introducing the deviance in the density (3.6), and using (3.11), we obtain

$$\begin{aligned} p(y;\mu,\sigma^2) &= a(\sigma^{-2})\exp\{t(0)/\sigma^2 - D(y,\mu)/(2\sigma^2)\} \\ &\simeq a(\sigma^{-2})\exp\{t(0)\sigma^2 - (y-\mu)^2/(2V\sigma^2)\}. \end{aligned} \tag{3.15}$$

Integrating with respect to y, we obtain

$$\begin{aligned} [a(\sigma^{-2})\exp\{t(0)/\sigma^2\}]^{-1} &\simeq \int_a^b \exp\{-(y-\mu)^2/(2V\sigma^2)\}dy \\ &\simeq \int_{-\infty}^{\infty} \exp\{-(y-\mu)^2/(2V\sigma^2)\}dy \\ &= (2\pi\sigma^2 V)^{1/2}. \end{aligned}$$

Inserting this in (3.15) gives the saddlepoint approximation (3.8). As we have already shown, (3.8) implies (3.7). This concludes the proof. ▮

Note that in the proof of Theorem 3.3.1, the terminology "saddlepoint approximation" comes from the fact that the exponent of the integrand of (3.12) has a saddlepoint at $\tilde{\theta}$, whereas the technique used in the proof of (3.8) is called Laplace's method, see Bleistein and Handelsman (1975). However, to simplify the terminology, we use the term "saddlepoint approximation" for either of the results (3.1), (3.8) or (3.14).

§3.4 The Fisher Information Matrix

3.4.1 The Information Matrix for Dispersion Models.

As indicated earlier, the deviance plays a key role in the inference for dispersion models, and in the previous sections we have worked out the basic properties of the deviance. A second key ingredient in the inference is the Fisher information matrix, whose structure reflects the structures of the inference process, in particular the fact that the parameters μ and σ^2 are orthogonal.

Let us return to the model defined in Section 3.1, with log-likelihood

$$L(\mu, \sigma^2) = \sum_{i=1}^{n} \log a(w_i/\sigma^2, y_i) + \sigma^{-2} \sum_{i=i}^{n} w_i t(y_i, \mu_i).$$

In the following we assume that the model is regular, that is, satisfies the usual regularity conditions of large-sample theory. In particular, the log likelihood is assumed to be differentiable, and the operations of differentiation and integration, for derivatives of the log-likelihood, may be interchanged, see Exercise 3.23 and 3.24.

By the definition of the deviance, the log likelihood may be written on the form

$$L(\mu, \sigma^2) = c(\mathbf{y},\sigma^2) - D(\mathbf{y},\mu)/(2\sigma^2),$$

where

$$c(\mathbf{y}, \sigma^2) = \sum_{i=1}^{n} \{\log a(w_i/\sigma^2, y_i) + \sup_{\mu \in \Omega} t(y_i, \mu) w_i/\sigma^2\}.$$

Hence, the components of the score vector for the parameter (μ, σ^2) are

$$\frac{\partial L(\mu, \sigma^2)}{\partial \mu_i} = u(y_i, \mu_i) w_i / \sigma^2, \tag{4.1}$$

where $u(y_i, \mu_i) = t'(y_i, \mu_i)$, and

$$\begin{aligned} \frac{\partial L(\mu, \sigma^2)}{\partial \sigma^2} &= u_{n+1}(\mathbf{y}, \mu, \sigma^2) \\ &= c'(\mathbf{y}, \sigma^2) + D(\mathbf{y}, \mu)/(2\sigma^4), \end{aligned} \tag{4.2}$$

where primes denote derivatives of t and c with respect to μ_i and σ^2, respectively.

Since the model is assumed to be regular, the expectations of (4.1) and (4.2) are zero, and hence

$$E_{\mu,\sigma^2}\{t'(Y_i, \mu_i)\} = 0 \tag{4.3}$$

and

$$E_{\mu,\sigma^2}\{c'(\mathbf{Y}, \sigma^2)\} = -E_{\mu,\sigma^2}\{D(\mathbf{Y}, \mu)\}/(2\sigma^4). \tag{4.4}$$

To find the Fisher (expected) information matrix, we calculate the expectation of minus the second derivative of the log likelihood.

Now, for any $i = 1, \ldots, n$

$$\frac{\partial^2 L(\mu, \sigma^2)}{\partial \sigma^2 \partial \mu_i} = -t'(y_i, \mu_i) w_i / \sigma^4. \tag{4.5}$$

From (4.3) it follows that (4.5) has zero expectation, a property expressed by saying that μ_i and σ^2 are *orthogonal parameters*. Hence, the Fisher information matrix for (μ, σ^2) is given by

$$\left\{ \begin{array}{cc} \mathbf{i}(\mu \mid \sigma^2)\mathbf{W}/\sigma^2 & 0 \\ 0 & i(\sigma^2 \mid \mu) \end{array} \right\}, \tag{4.6}$$

where $\mathbf{i}(\mu \mid \sigma^2)$ and $\mathbf{W}$ are diagonal matrices with ith diagonal elements defined by

$$\mathbf{i}_{ii}(\mu \mid \sigma^2) = -E_{\mu,\sigma^2}\{t''(Y_i, \mu_i)\}, \tag{4.7}$$

and $\mathbf{W}_{ii} = w_i$, respectively, and where

$$(4.8) \qquad i(\sigma^2 \mid \mu) = E_{\mu,\sigma^2}\{-c''(\mathrm{Y},\sigma^2) + D(\mathrm{Y},\mu)/\sigma^6\}.$$

Hence, the information matrix (4.6) is diagonal, the off-diagonal zeroes occurring either because of the independence of the observations, or because of the orthogonality of the parameters μ_i and σ^2 for $i = 1, \ldots, n$.

Consider now the hypothesis

$$H_1 : \mu = \mu(\beta), \quad \beta = (\beta_1, \ldots, \beta_{k_1})^T,$$

as defined in Section 3.1. The regularity conditions for the model require the function $\mu(\,\cdot\,)$ to be twice differentiable. Moreover, defining

$$\mathbf{X}(\beta) = \frac{\partial \mu}{\partial \beta^T},$$

we assume that $\mathbf{X}(\beta)$ $(n \times k_1)$ has full rank for every value of β.

Using (4.6), we find that (β, σ^2) has information matrix

$$(4.9) \qquad \left\{ \begin{matrix} \mathbf{i}(\beta \mid \sigma^2)/\sigma^2 & 0 \\ 0 & i(\sigma^2 \mid \mu(\beta)) \end{matrix} \right\},$$

where

$$\mathbf{i}(\beta \mid \sigma^2) = \mathbf{X}(\beta)^T \mathbf{W} \mathbf{i}(\mu(\beta) \mid \sigma^2) \mathbf{X}(\beta).$$

Hence, (4.9) shows that β_j and σ^2 are orthogonal for every $j = 1, \ldots, k_1$. We express this by saying that β and σ^2 are orthogonal. By the orthogonality of β and σ^2, it follows that $\mathbf{i}(\beta \mid \sigma^2)$ is identical to the expected information matrix for β when σ^2 is known, and $i(\sigma^2 \mid \mu)$ is the expected information matrix for σ^2 when μ is known.

The second derivative of the log likelihood function with respect to β is

$$\begin{aligned} \frac{\partial^2 L}{\partial \beta \partial \beta^T} &= \mathbf{X}(\beta)^T \frac{\partial^2 L}{\partial \mu \partial \mu^T} \mathbf{X}(\beta) + \sum_{i=1}^{n} \frac{\partial^2 \mu_i}{\partial \beta \partial \beta^T} \frac{\partial L}{\partial \mu_i} \\ &= \left\{ -\mathbf{X}(\beta)^T \mathbf{j}(\mu) \mathbf{W} \mathbf{X}(\beta) + \sum_{i=1}^{n} \frac{\partial^2 \mu_i}{\partial \beta \partial \beta^T} u(y_i, \mu_i) w_i \right\} / \sigma^2 \\ &= -\mathbf{j}(\beta)/\sigma^2, \end{aligned}$$

say, where $\mathbf{j}(\mu)$ is the diagonal matrix with diagonal elements

$$\mathbf{j}_{ii}(\mu) = -t''(y_i, \mu_i), \quad i = 1, \ldots, n.$$

The observed information matrix for β is

$$\mathbf{j}(\hat{\beta})/\sigma^2, \tag{4.10}$$

the second derivative of the log likelihood at the maximum $\hat{\beta}$, which is positive-definite. Since $\mathbf{i}(\beta \mid \sigma^2) = E_{\mu,\sigma^2}(\mathbf{j}(\beta))$, we find that

$$\mathbf{j}(\hat{\beta}) \xrightarrow{P} \mathbf{i}(\beta \mid \sigma^2) \quad \text{for} \quad n \to \infty. \tag{4.11}$$

We shall now examine (4.9) in some special cases, in particular for exponential dispersion models and the small-dispersion case.

3.4.2 The Information Matrix for Exponential Dispersion Models.

For an exponential dispersion model $Y_i \sim ED(\mu_i, \sigma^2/w_i), \quad i = 1, \ldots, n$, we have

$$t(y, \mu) = y\tau^{-1}(\mu) - \kappa(\tau^{-1}(\mu)), \quad \mu \in \Omega.$$

Hence

$$t'(y, \mu) = (y - \mu)/V(\mu) \tag{4.12}$$

and

$$t''(y, \mu) = \frac{-V(\mu) - (y - \mu)V'(\mu)}{V(\mu)^2},$$

which implies

$$\mathbf{i}_{ii}(\mu \mid \sigma^2) = 1/V(\mu). \tag{4.13}$$

To emphasize that $\mathbf{i}(\mu \mid \sigma^2)$ does not depend on σ^2 in this case, we shall write $\mathbf{i}(\mu)$ and $\mathbf{i}(\beta)$ instead of $\mathbf{i}(\mu \mid \sigma^2)$ and $\mathbf{i}(\beta \mid \sigma^2)$, respectively, whenever we are dealing with an exponential dispersion model.

To verify that the model is regular, we must show (cf. Exercise 3.23), that

$$E_{\mu,\sigma^2}\{(Y_i - \mu_i)/V(\mu_i)\} = 0, \tag{4.14}$$

which is (4.3), and

$$\mathbf{i}_{ii}(\mu)w_i/\sigma^2 = Var_{\mu,\sigma^2}[(Y_i - \mu_i)w_i/\{\sigma^2 V(\mu_i)\}]. \tag{4.15}$$

However, (4.14) and (4.15) easily follow from the standard results $E_{\mu,\sigma^2}(Y_i) = \mu_i$ and $Var_{\mu,\sigma^2}(Y_i) = \sigma^2 V(\mu_i)/w_i$.

3.4.3 Small-Dispersion Asymptotics and the Information Matrix.

We shall now examine the information matrix in the case where the dispersion parameter is small. For the distribution $Y_i \sim DM(\mu_i, \sigma^2/w_i)$ this means that σ^2/w_i is small. The most convenient assumption is to take σ^2 as fixed and consider the limit

$$\min\{w_1, \ldots, w_n\} \to \infty. \tag{4.16}$$

We use the notation "$\mathbf{w} \to \infty$" for (4.16), where $\mathbf{w} = (w_1, \ldots, w_n)^T$. We refer to (4.16) as the "small-dispersion" case, and the corresponding results are called "small-dispersion" results. Similarly, we shall refer to the case "$n \to \infty$" as the "large-sample case" etc.

Using the saddlepoint approximation (Theorem 3.3.1 or 3.3.3) we find for $\mathbf{w} \to \infty$

$$c(\mathbf{y}, \sigma^2) \simeq -\frac{1}{2}\sum_{i=1}^{n}[\log\{2\pi V(y_i)/w_i\} + \log\sigma^2]. \tag{4.17}$$

Here, and in the following, we let $V(y_i)$ or $V(\mu_i)$ denote $V = -1/t''(0)$ in the case of a dispersion model (Theorem 3.3.3). For $\mathbf{w} \to \infty$, (4.2) becomes

$$u_{n+1}(\mu, \sigma^2) \simeq -n/(2\sigma^2) + D(\mathbf{y}, \mu)/(2\sigma^4),$$

which shows that the maximum likelihood estimate of σ^2 under H_1 is approximately

$$\sigma^2 \simeq D(\mathbf{y}, \mu(\hat{\beta}))/n \quad \text{for} \quad \mathbf{w} \to \infty. \tag{4.18}$$

The equation (4.4) becomes

$$E_{\mu,\sigma^2}\{D(\mathbf{Y}, \mu)\} \simeq n\sigma^2 \quad \text{for} \quad \mathbf{w} \to \infty. \tag{4.19}$$

Since Y_i tends to μ_i in probability in the small-dispersion case, we find the approximation, from (4.7),

$$\mathbf{i}_{ii}(\mu \mid \sigma^2) \simeq 1/V \quad \text{for} \quad \mathbf{w} \to \infty, \tag{4.20}$$

providing an analogue of (4.13) for the general case.

From (4.8), (4.17) and (4.19) we find

$$i(\sigma^2 \mid \mu) \simeq n/(2\sigma^4) \quad \text{for} \quad \mathbf{w} \to \infty. \tag{4.21}$$

We note that (4.21) does not depend on the model under consideration. In particular we have $i(\sigma^2 \mid \mu) = n/(2\sigma^4)$ for the normal and inverse Gaussian distributions, see Exercise 3.25. For a linear normal model the estimator for σ^2 is proportional to a χ^2-variate. Hence, (4.21) suggests that a χ^2-approximation may be better than a normal approximation for the distribution of $\hat{\sigma}^2$ in the small-dispersion case. This will be confirmed later, cf. Section 3.6.

§3.5 Parameter Estimation

3.5.1 Maximum Likelihood Estimation.

In Section 3.2 we saw that the maximum likelihood estimate of β may be found by minimizing the deviance $D(\mathbf{y}, \mu(\beta))$ with respect to β. Using the results of Section 3.4, we find that the likelihood equation for β is

$$X(\beta)^T \mathbf{W}\mathbf{u}(\mathbf{y}, \mu(\beta)) = 0, \tag{5.1}$$

where $u_i(\mathbf{y}, \mu) = u(y_i, \mu_i) = t'(y_i, \mu_i)$. Equation (5.1) is in general nonlinear in β, and hence it must be solved by iterative methods.

Given the estimate $\hat{\beta}$, the maximum likelihood estimate of σ^2 may be found as the solution to the equation

$$c'(\mathbf{y}, \sigma^2) + D_1/(2\sigma^4) = 0, \tag{5.2}$$

where $D_1 = D(\mathbf{y}, \mu(\hat{\beta}))$ is the deviance for the hypothesis H_1 under consideration. This equation is also nonlinear, and generally has to be solved by iterative methods.

By standard asymptotic theory we know that the estimate $(\hat{\beta}, \hat{\sigma}^2)$ is asymptotically normally distributed, with variance matrix given by the inverse of the information matrix (4.9). By the orthogonality of β and σ^2, and the consequent block diagonal structure of (4.9), the two blocks of the matrix may be inverted separately. In particular, $\hat{\beta}$ and $\hat{\sigma}^2$ are asymptotically independent, and the asymptotic normal distributions for $\hat{\beta}$ and $\hat{\sigma}^2$ are

$$\hat{\beta} \sim N(\beta, \sigma^2 \mathbf{i}^{-1}(\beta \mid \sigma^2)) \tag{5.3}$$

and

$$\hat{\sigma}^2 \sim N(\sigma^2, i^{-1}(\sigma^2 \mid \mu(\beta))) \tag{5.4}$$

The main condition for these results, apart from regularity conditions, which we have already discussed, is that the information matrix $\mathbf{i}(\beta \mid \sigma^2)$ tends to infinity. This is the case, for example, if the limit

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{i}(\beta \mid \sigma^2) \tag{5.5}$$

exists, and is positive-definite. A more practical way of stating this condition is to say that for each $j = 1, \ldots, k_1$ many observations contribute to the estimation of each parameter β_j.

By (5.3) the asymptotic standard error of the estimate $\hat{\beta}_j$ is

$$\mathrm{se}(\hat{\beta}_j, \sigma^2) = \sigma\{\mathbf{i}^{jj}(\beta \mid \sigma^2)\}^{1/2}, \tag{5.6}$$

where by A^{jk} we denote the jk-th element of the inverse A^{-1} of a quadratic matrix. If σ^2 is known, an estimate of the asymptotic standard error of $\hat{\beta}$ may be obtained by inserting $\hat{\beta}$ for β in (5.6). If σ^2 is unknown, we must also insert an estimate for σ^2 in (5.6). Here we may use either the maximum likelihood estimate $\hat{\sigma}^2$ or any other consistent estimate. In a similar way we may estimate the asymptotic covariance between $\hat{\beta}_j$ and $\hat{\beta}_k$ via $\sigma^2\mathbf{i}^{jk}(\beta \mid \sigma^2)$.

The asymptotic standard error for $\hat{\sigma}^2$ may be calculated as $i(\sigma^2 \mid \mu(\beta))^{-1/2}$. However, in general, a much better approach is to approximate the distribution of $\hat{\sigma}^2$ by a χ^2-distribution, as will be discussed in Section 3.6.

3.5.2 Estimation of The Dispersion Parameter.

The maximum likelihood estimate of σ^2 has certain undesirable properties, and for this reason we shall suggest some alternative estimators with better properties. This follows a long tradition for linear normal models, for which the unbiased estimate

$$\frac{1}{n-k_1}\sum_{i=1}^{n} w_i(y_i - \hat{\mu}_i)^2 \tag{5.7}$$

is always used instead of the maximum likelihood estimate

$$\frac{1}{n}\sum_{i=1}^{n} w_i(y_i - \hat{\mu}_i)^2.$$

Let us analyze the situation in the small-dispersion case. Consider the two hypotheses H_1 and H_2 from Section 3.1, and let D_1 and D_2 denote the corresponding deviances. Then, by (4.18), the estimates of σ^2 under H_1 and H_2, respectively, are approximately

$$H_1: \hat{\sigma}_1^2 \simeq \frac{1}{n} D_1 \quad for \quad \mathbf{w} \to \infty$$

$$H_2: \hat{\sigma}_2^2 \simeq \frac{1}{n} D_2 \quad for \quad \mathbf{w} \to \infty.$$

Since $D_2 > D_1$, due to the resting of the hypotheses, we have $\hat{\sigma}_2^2 > \hat{\sigma}_1^2$, (approximately). This seems an undesirable property for two estimators of the same parameter σ^2, considering the case where H_1 and H_2 are both true. However, it is easy to understand why the estimators have this property. By (4.19) we have

$$E_{\mu,\sigma^2}\{D(\mathbf{Y},\mu)/n\} \simeq \sigma^2 \quad \text{for} \quad \mathbf{w} \to \infty.$$

Hence, if μ were known, the estimator $D(\mathbf{y},\mu)/n$ would be perfectly adequate (in the small-dispersion case). However, since $\hat{\beta}$ is obtained by minimizing $D(\mathbf{y},\mu(\beta))$ with respect to β, it is clear that we have approximately, in some sense,

$$E_{\mu,\sigma^2}\{D(\mathbf{Y},\mu(\hat{\beta}))\} < E_{\mu,\sigma^2}\{D(\mathbf{Y},\mu)\}.$$

In fact we show in Section 3.6 that in the small-dispersion case we have

$$E_{\mu,\sigma^2}\{D(\mathbf{Y},\mu(\hat{\beta}))\} \simeq (n-k_1)\sigma^2 \quad \text{for} \quad \mathbf{w} \to \infty. \tag{5.8}$$

Hence, any reasonable estimator for σ^2 should probably be approximately equal to $D(\mathbf{y},\mu(\hat{\beta}))/(n-k_1)$ in the small-dispersion case.

A simple way to obtain an estimator with this property is to define the estimate as the solution to the equation

$$E_{\mu,\sigma^2}\{D(\mathbf{Y},\mu(\hat{\beta}))\} = D(\mathbf{y},\mu(\hat{\beta})).$$

However, this approach has two disadvantages. First, the expectation of the deviance $D(\mathbf{Y},\mu(\hat{\beta}))$ may be difficult to calculate, and it depends on the hypothesis under consideration. Second, the expectation of the deviance in (5.8) in general depends on the value of μ, making it necessary to insert an estimate for μ in the expression for the expectation. For these reasons, the practical value of the estimate based on (5.8) for general use is fairly limited.

In certain special cases of particular interest, such as certain special models in the normal and the inverse Gaussian distribution, arguments based on sufficiency or ancillarity principles may suggest the correct way to estimate σ^2. However, such arguments rarely apply within the entire class of models under consideration here. From a practical point of view there are basically three estimators that we need to consider.

3.5.3 The Modified Profile Likelihood Estimate.

The modified profile log likelihood for σ^2 is defined by

$$L^0(\sigma^2) = \frac{k_1}{2}\log\sigma^2 + L(\mu(\hat{\beta}), \sigma^2) \tag{5.9}$$

under H_1. The value of σ^2 that maximizes (5.9), denoted $\hat{\sigma}_0^2$, is called the modified profile likelihood estimate. The reason for this name is that the term $L(\mu(\hat{\beta}), \sigma^2)$ in (5.9) is known as the profile log likelihood for σ^2. Using (4.17), we find that in the small-dispersion case

$$\frac{\partial L^0}{\partial\sigma^2} \simeq \frac{k_1 - n}{2\sigma^2} + \frac{D(\mathbf{y}, \mu(\hat{\beta}))}{2\sigma^4} \quad \text{for} \quad \mathbf{w} \to \infty,$$

giving the approximation

$$\hat{\sigma}_0^2 \simeq \frac{D(\mathbf{y}, \mu(\hat{\beta}))}{n - k_1} \quad \text{for} \quad \mathbf{w} \to \infty. \tag{5.10}$$

Hence, $\hat{\sigma}_0^2$ has the right limiting value for $\mathbf{w} \to \infty$. Since $\hat{\sigma}_0^2$ is equivalent to $\hat{\sigma}^2$ for n large, this estimate is consistent for σ^2 in the large-sample case.

3.5.4 The Deviance-Based Estimate.

A very simple estimate is obtained by using the asymptotic value (5.10), obtaining the estimate

$$\tilde{\sigma}^2 = \frac{D(\mathbf{y}, \mu(\hat{\beta}))}{n - k_1}.$$

This estimate may be considered as the analogue of the estimate (5.7) for the normal distribution, giving it an immediate intuitive appeal. However, we shall see, in Example 3.5.3, that $\tilde{\sigma}^2$ is not in general a consistent estimator for σ^2 in the large-sample case. For this reason, $\tilde{\sigma}^2$ is not recommended for general use, but it may be used if it is known that $\sigma^{-2}\min\{w_1, \ldots, w_n\}$ is reasonably large.

3.5.5 The Pearson Estimate.

A reasonably simple estimator with the correct asymptotic behaviour may be obtained as the solution $\overline{\sigma}^2$ to the equation

$$\sigma^2 = \frac{1}{n-k_1} X^2(\mu(\hat{\beta}), \sigma^2), \tag{5.11}$$

where the statistic X^2 is defined by

$$X^2(\mu, \sigma^2) = \sum_{i=1}^{n} t'(y_i, \mu_i)^2 w_i / \mathrm{i}_{ii}(\mu \mid \sigma^2). \tag{5.12}$$

The statistic (5.12) is known as the generalized Pearson statistic. The form of (5.11) suggests a simple iterative procedure for calculating $\overline{\sigma}^2$.Using (5.11) to update a preliminary estimate gives the sequence $\sigma_1^2, \sigma_2^2, \ldots$ defined by

$$\sigma_{m+1}^2 = \frac{1}{n-k_1} X^2(\mu(\hat{\beta}), \sigma_m^2).$$

Any estimator, such as $\tilde{\sigma}^2$, may be used to initiate the iterative process.

To justify (5.11), consider first the case of an exponential dispersion model $Y_i \sim ED(\mu_i, \sigma^2/w_i)$. By (4.12) and (4.13), (5.11) becomes

$$\overline{\sigma}^2 = \frac{1}{n-k_1} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 w_i / V(\hat{\mu}_i), \tag{5.13}$$

where $\hat{\mu}_i = \mu_i(\hat{\beta}), \quad i = 1, \ldots, n$. In this case no iteration is needed, so $\overline{\sigma}^2$ is a convenient alternative to the modified profile likelihood estimate.

Consider now a dispersion model with $t(y, \mu) = t(y - \mu)$. Under the conditions of Theorem 3.3.3 we find, expanding $t'(\,\cdot\, - \mu)$ around μ, using $V = -1/t''(0)$,

$$t'(y_i - \mu_i) \simeq -(y_i - \mu_i)/V.$$

Hence by (4.20)

$$\overline{\sigma}^2 \simeq \frac{1}{n-k_1} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 w_i / V \quad \text{for} \quad \mathbf{w} \to \infty, \tag{5.14}$$

which is an approximate analogue of (5.13).

Using (3.5) and (3.11), we hence find the asymptotic relation

$$\overline{\sigma}^2 \simeq \frac{1}{n-k_1} D(\mathbf{y}, \mu(\hat{\beta})) \quad \text{for} \quad \mathbf{w} \to \infty, \tag{5.15}$$

valid under the assumptions of either Theorem 3.3.1 or 3.3.3. Hence, $\overline{\sigma}^2$ has the right limiting behaviour in the small-dispersion case.

Now, let us show that $\overline{\sigma}^2$ is consistent for σ^2 in the large-sample case. By the regularity of the model we have the following relation

$$Var_{\mu,\sigma^2}\{t'(Y_i,\mu_i)w_i/\sigma^2\} = \mathbf{i}_{ii}(\mu \mid \sigma^2)w_i/\sigma^2. \tag{5.16}$$

Using (4.3), this implies

$$\begin{aligned} E_{\mu,\sigma^2}\{t'(Y_i,\mu_i)^2\} &= Var_{\mu,\sigma^2}\{t'(Y_i,\mu_i)\} \\ &= \mathbf{i}_{ii}(\mu \mid \sigma^2)\sigma^2/w_i. \end{aligned}$$

Hence, by (5.12) we find

$$E_{\mu,\sigma^2}\{X^2(\mu,\sigma^2)\} = n\sigma^2,$$

which implies that $\overline{\sigma}^2$ is consistent in the large-sample case, because a consistent estimator, $\hat{\beta}$, was used in (5.11).

3.5.6 Comparison of The Three Estimators for The Dispersion Parameter.

We now consider some examples, that illustrate the behaviour of the three estimators.

Example 3.5.1: For the normal distribution, the three estimators $\hat{\sigma}_0^2, \tilde{\sigma}^2$ and $\overline{\sigma}^2$ are identical, and correspond to the usual estimate (5.7). This is shown in Exercise 3.26. ▮

Example 3.5.2: Consider the inverse Gaussian distribution $Y_i \sim IG(\mu_i, \sigma^2/w_i)$. In this case the density is identical to the saddlepoint approximation, and hence the approximation (5.10) becomes exact. Hence

$$\hat{\sigma}_0^2 = \tilde{\sigma}^2 = \frac{1}{n-k_1} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2 w_i}{y_i \hat{\mu}_i^2}, \tag{5.17}$$

so in this case the deviance-based estimate $\tilde{\sigma}^2$ is consistent in the large-sample case. By (5.13), the Pearson estimate is

$$\bar{\sigma}^2 = \frac{1}{n-k_1}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 w_i/\hat{\mu}_i^3. \tag{5.18}$$

The slight difference between (5.17) and (5.18) is probably unimportant in practice, but due to the occurrence of y_i in the denominator of (5.17), this estimate is more sensitive to very small or very large observations. ▮

Example 3.5.3: In the case of the gamma distribution, $Y_i \sim Ga(\mu_i, \sigma^2 /w_i)$, $c'(\mathbf{y}, \sigma^2)$ does not depend on y, because

$$c(\mathbf{y}, \sigma^2) = \sum_{i=1}^{n}\{\lambda_i \log \lambda_i - \lambda_i - \log \Gamma(\lambda_i) - \log y_i\},$$

where $\lambda_i = w_i/\sigma^2$, $i = 1, \ldots, n$. Hence, writing $c'(\sigma^2) = c'(\mathbf{y}, \sigma^2)$, we find that $\hat{\sigma}_0^2$ is the solution to the equation

$$k_1/(2\sigma^2) + c'(\sigma^2) + D_1/(2\sigma^4) = 0,$$

where

$$c'(\sigma^2) = \sum_{i=1}^{n}\{\psi(w_i/\sigma^2) - \log(w_i/\sigma^2)\}w_i/\sigma^4,$$

with $\psi(\lambda) = \Gamma'(\lambda)/\Gamma(\lambda)$ denoting the digamma function. Using (4.4) and an asymptotic expansion for ψ (cf. Abramowitz and Stegun, 1972, p 259), we obtain

$$\begin{aligned} E_{\mu,\sigma^2}\{D(\mathbf{Y}, \mu)\} &= 2\sum_{i=1}^{n} w_i\{\log(w_i/\sigma^2) - \psi(w_i/\sigma^2)\} \\ &= n\sigma^2 + \sum_{i=1}^{n}\{\frac{\sigma^4}{6w_i} - \frac{\sigma^6}{60w_i^2} + \ldots). \end{aligned}$$

Consequently, the deviance-based estimate $\tilde{\sigma}^2$ has bias approximately σ^4 $(1/w_1 + \cdots + 1/w_n)/\{6(n-k_1)\}$, which persists for n large. We conclude that $\tilde{\sigma}^2$ is inconsistent in the large- sample case. Hence, for the gamma distribution, one should use either σ_0^2 or the Pearson estimate, which is

$$\bar{\sigma}^2 = \frac{1}{n-k_1}\sum_{i=1}^{n} w_i\{(y_i - \hat{\mu}_i)/\hat{\mu}_i\}^2. \quad ▮$$

Example 3.5.4: Using the Bessel function I_0, cf. Example 3.3.4, the probability density function of the von Mises-Fisher distribution may be written on the form

$$(5.19) \qquad p(y;\mu_i,\sigma^2/w_i) = \frac{1}{2\pi I_0(\frac{w_i}{\sigma^2})}\ \exp\{\frac{w_i}{\sigma^2}\cos(y-\mu_i)\},$$

for $0 < y < 2\pi$. In this case $c(\mathbf{y},\sigma^2)$ does not depend on $\mathbf{y}$, and may be written as

$$c(\sigma^2) = \sum_{i=1}^{n}[w_i/\sigma^2 - \log\{2\pi I_0(w_i/\sigma^2)\}].$$

The modified profile likelihood estimate may hence be obtained as the solution of the equation

$$k_1/(2\sigma^2) + c'(\sigma^2) + D_1/(2\sigma^4) = 0,$$

where

$$c'(\sigma^2) = \sum_{i=1}^{n}\{I_1(w_i/\sigma^2)/I_0(w_i/\sigma^2) - 1\}w_i/\sigma^4.$$

Here I_0 is the modified Bessel function of the first kind of order 1, defined by

$$I_1(\lambda) = I_0'(\lambda) = \frac{1}{2\pi}\int_0^{2\pi}\cos y \exp\{\lambda\cos y\}dy.$$

In particular, we obtain the relation

$$E\{\cos(Y_i-\mu_i)\} = I_1(w_i/\sigma^2)/I_0(w_i/\sigma^2),$$

where the random variable Y_i has probability density function (5.19).

Using (4.4) we obtain

$$E_{\mu,\sigma^2}\{D(\mathbf{Y},\mu)\} = -c'(\sigma^2)2\sigma^4$$

which, like for the gamma distribution, implies that the estimator $\tilde{\sigma}^2$ is inconsistent in the large-sample case.

To derive the estimate $\overline{\sigma}^2$, note that $t'(y-\mu) = -\sin(y-\mu)$ and $t''(y-\mu) = -\cos(y-\mu)$. Hence

$$\begin{aligned}\mathbf{i}_{ii}(\mu \mid \sigma^2) &= -E_{\mu,\sigma^2}\{t''(Y_i-\mu_i)\}\\ &= I_1(w_i/\sigma^2)/I_0(w_i/\sigma^2).\end{aligned}$$

Hence $\overline{\sigma}^2$ is the solution to the equation

$$\sigma^2 = \frac{1}{n-k_1}\sum_{i=1}^{n}\sin^2(y_i - \hat{\mu}_i)w_i I_0(w_i/\sigma^2)/I_1(w_i/\sigma^2).$$

By (5.14) and the fact that $V = 1$, we obtain the approximation

$$\overline{\sigma}^2 \simeq \frac{1}{n-k_1}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 w_i. \tag{5.20}$$

However, this approximation is problematic for the von Mises-Fisher distribution, because μ_i is defined modulo 2π, and hence (5.20) may depend on the choice of origin for the data. A better approximation to $\overline{\sigma}^2$ is obtained by inserting (4.20) in (5.12), to obtain

$$\overline{\sigma}^2 \simeq \frac{1}{n-k_1}\sum_{i=1}^{n}\sin^2(y_i - \hat{\mu}_i)w_i. \tag{5.21}$$

However, the estimates (5.20) and (5.21) are both inconsistent in the large-sample case. ▮

The von Mises-Fisher distribution gives occasion to discuss get another estimate of σ^2. This estimate is obtained by inserting the approximation

$$\begin{aligned}\mathbf{i}_{ii}(\mu \mid \sigma^2) &= -E_{\mu,\sigma^2}\{t''(Y_i, \mu_i)\}\\ &\simeq -t''(y_i, \mu_i)\end{aligned}$$

in (5.12), which yields the estimate

$$\breve{\sigma}^2 = \frac{1}{n-k_1}\sum_{i=1}^{n} t'(y_i,\mu_i)^2 w_i/\{-t''(y_i,\mu_i)\}.$$

This estimate has the advantage, compared with $\overline{\sigma}^2$, that it may be calculated without iteration. However, $\breve{\sigma}^2$ is not in general consistent for σ^2 in the large-sample case, and hence it is not recommended for general use. In the case of the von Mises-Fisher distribution, this estimate becomes

$$\breve{\sigma}^2 = \frac{1}{n-k_1}\sum_{i=1}^{n}\sin^2(y_i - \hat{\mu}_i)w_i/\cos(y_i - \hat{\mu}_i).$$

In this and other cases where $t(\,\cdot\,)$ is not concave, $-t''(y_i, \mu_i)$ may be zero or negative, and hence $\check{\sigma}^2$ may be negative, or even infinite. If a quick estimate is needed, then clearly the deviance-based estimate $\tilde{\sigma}^2$ or the approximation (5.14) are preferable.

§3.6 Asymptotic Theory

3.6.1 The Results.

We shall now present the basic asymptotic results for the models under consideration. The proof of the results with appropriate regularity conditions will be outlined in Section 3.6.3.

We shall use the notation "$n \to \infty$" and "$\mathbf{w} \to \infty$" to indicate respectively large-sample results and small-dispersion results, as before. We write "$n, \mathbf{w} \to \infty$" if a result is valid in both limits. The notation $\overset{H_1}{\sim}$ indicates that the result is valid under the hypothesis H_1 etc. The hypotheses H_1 and H_2 and their deviances D_1 and D_2 are as defined in Section 3.1 and 3.2.

The three fundamental results that form the basis of the asymptotic theory are

$$\hat{\beta} \overset{H_1}{\sim} N(\beta, \sigma^2 \mathbf{i}^{-1}(\beta \mid \sigma^2)) \qquad \text{for} \quad n, \mathbf{w}, \to \infty \tag{6.1}$$

$$D_1/\sigma^2 \overset{H_1}{\sim} \chi^2(n - k_1) \quad \text{for} \quad \mathbf{w} \to \infty \tag{6.2}$$

$$(D_2 - D_1)/\sigma^2 \overset{H_2}{\sim} \chi^2(k_1 - k_2) \quad \text{for} \quad n, \mathbf{w} \to \infty. \tag{6.3}$$

Furthermore, $\hat{\beta}$ and D_1 are asymptotically independent under H_1 for $\mathbf{w} \to \infty$, and D_1 and $D_2 - D_1$ are asymptotically independent under H_2 for $\mathbf{w} \to \infty$.

Result (6.1) was justified in the large-sample case in Section 3.5.1 (Equation (5.1)). If σ^2 is known, the statistics in (6.2) and (6.3) are the log likelihood ratio tests for respectively H_1 under H_0 and H_2 under H_1. Hence (6.3) follows immediately from standard large-sample theory, whereas (6.2), as indicated, is generally not valid in the large-sample case, because the dimension of H_0 depends on n.

The fact that (6.1), (6.2) and (6.3) are valid in the small-dispersion case is, for exponential dispersion models, a consequence of the standard convolution formula for exponential dispersion models (Section 2.3). The details of the argument is given in Section 3.6.2, together with a proof of the small-dispersion results for the dispersion model case.

From (6.1)-(6.3) follow three results which are basic for analysis of deviance, namely

$$\text{(6.4)} \qquad \frac{\hat\beta_j - \beta_j}{\hat\sigma_0 \mathbf{j}^{jj}(\hat\beta)^{1/2}} \overset{H_1}{\sim} t(n-k_1) \quad \text{for} \quad n, \mathbf{w} \to \infty$$

$$\text{(6.5)} \qquad (n-k_1)\hat\sigma_0^2/\sigma^2 \overset{H_1}{\sim} \chi^2(n-k_1) \quad \text{for} \quad \mathbf{w} \to \infty$$

$$\text{(6.6)} \qquad \frac{D_2 - D_1}{\hat\sigma_0^2(k_1-k_2)} \overset{H_2}{\sim} F(k_1-k_2, n-k_1) \quad \text{for} \quad n, \mathbf{w} \to \infty.$$

Here $t(f_1)$ and $F(f_1, f_2)$ denote the t- and F-distributions with degrees of freedom as indicated, and $\hat\sigma_0^2$, $\overline{\sigma}^2$ and $\tilde\sigma^2$ denote estimates under H_1. The denominator $\hat\sigma_0 \mathbf{j}^{jj}(\hat\beta)^{1/2}$ in (6.4) is the estimated standard error of $\hat\beta_j$ based on the observed information (4.10). For an exponential dispersion model we may replace $\mathbf{j}^{jj}(\hat\beta)$ by the corresponding quantity $\mathbf{i}^{jj}(\hat\beta)$ based on the expected information. This is not possible for a general dispersion model, because $\mathbf{i}(\beta \mid \sigma^2)$ then depends on σ^2, and if σ^2 is replaced by $\hat\sigma_0^2$, the asymptotic distribution of (6.4) for $\mathbf{w} \to \infty$ is altered.

Whereas, by (6.1), $\hat\beta$ is consistent for β in the small-dispersion case, (6.5) shows that $\hat\sigma^2$ is *not* consistent for σ^2 in the small-dispersion case, because the asymptotic variance of $\hat\sigma_0^2$ does not tend to zero for $\mathbf{w} \to \infty$ and apparently it is not possible to estimate σ^2 consistently in this limit. This is in contract to the large-sample case, where σ^2, as well as β, may be estimated consistently by for example the maximum likelihood estimator.

To show (6.4)-(6.6), consider first the small-dispersion case, where we have

$$\text{(6.7)} \qquad \hat\sigma_0^2 \simeq \tilde\sigma^2 = D_1/(n-k_1).$$

Hence, by the asymptotic independence of $\hat\beta$ and D_1, the result (6.4) follows from (6.1), (6.2) and the definition of the t-distribution, because, by (4.11) and Slutsky's theorem, $\mathbf{j}$ may be replaced by $\mathbf{i}(\beta \mid \sigma^2)$ in the limit. Similarly, (6.5) follows from (6.7) and (6.2), and (6.6) follows from (6.2), (6.3), (6.7) and the asymptotic independence of D_1 and $D_2 - D_1$.

In the large-sample case, the consistency of $\mathbf{j}$, (4.11), implies that the ratio in (6.4) converges to a standard normal distribution. However, the $t(n-k_1)$-distribution is approximately a standard normal distribution for n large and k_1 fixed, and hence (6.4) is a valid approximation for n large. Similarly, (6.6) follows from (6.3) and the fact that $\hat{\sigma}_0^2$ is consistent, because the distribution $F(f_1, f_2)$ is approximately a $\chi^2(f_1)/f_1$-distribution for f_2 large. Note that $\hat{\sigma}_0^2$ and $D_2 - D_1$ are asymptotically independent for $n, \mathbf{w} \to \infty$, so that (6.6) is likely to hold in the area where $D_2 - D_1$ and $\hat{\sigma}_0^2$ have approximate χ^2-distributions.

The idea behind the results (6.4) and (6.6) is that, since they are valid in both the small-dispersion case and the large-sample case, they are likely to be valid in a quite wide range of cases, and may be applied without making an explicit choice between the two asymptotic frameworks. The proof of the results (6.1)-(6.3) shows that the error in (6.4) is proportional to $\sigma/\{nw\}^{1/2}$, where $w = \min\{w_1, \ldots, w_n\}$. The actual error depends on the model under consideration, and it is difficult to make more precise general statements about the accuracy of the approximations.

In the case where σ^2 is known, the results (6.1)-(6.3) may be applied directly in the inference on the parameter β. However, the discussion of this case is more relevant in the discrete case (Section 3.8), and the conclusions from the discrete case apply, with obvious modifications, in the continuous case.

3.6.2 The Relation Between The F-Test and The Likelihood Ratio Test.

The F-statistic in (6.6) may be used for testing H_2 under H_1, and in fact this test is the main tool of inference in the analysis of deviance. This raises the question of whether the F-test has any optimality properties. A partial answer to this question may be found by analyzing the relation between the F-test and the likelihood ratio test, the latter being known as asymptotically optimal and optimal in certain simple special cases, and as giving acceptable results in wide generality. We have already noted that the statistic $(D_2 - D_1)/(2\sigma^2)$ is the log likelihood ratio test for H_2 under H_1 when σ^2 is known. However, when σ^2 is unknown, the log likelihood ratio test is a more complicated function of the observations, and we resort to an investigation of the asymptotic relation between the two tests.

In the small-dispersion case, the maximum likelihood estimate of σ^2 under H_j is approximately $\hat{\sigma}_j^2 = D_j/n$, $j = 1, 2$. Using the saddlepoint approximation, the maximized log likelihood under H_1 is approximately

$$L(\hat{\mu}^{(j)}, \hat{\sigma}_j^2) \simeq -\frac{1}{2}\sum_{i=1}^{n}[\log\{2\pi V(y_i)/w_i\} + \log(D_j/n)] - \frac{n}{2},$$

where $\hat{\mu}^{(j)}$ is the maximum likelihood estimate under H_j, $j = 1, 2$. Hence, the log likelihood ratio test for H_2 under H_1 is approximately

$$\begin{aligned} LR_{1,2} &= 2\{L(\hat{\mu}^{(1)}, \hat{\sigma}_1^2) - L(\hat{\mu}^{(2)}, \hat{\sigma}_2^2)\} \\ &\simeq n \log(D_2/D_1) \\ &= n \log\{1 + (D_2 - D_1)/D_1\} \end{aligned}$$

which is a monotone function of the ratio

$$\frac{(D_2 - D_1)/(k_1 - k_2)}{D_1/(n - k_1)}.$$

Hence, by (5.10), $LR_{1,2}$ is asymptotically equivalent to the F-test in (6.6).

In the large-sample case we have that $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are asymptotic ally equivalent to $\hat{\sigma}_{01}^2$ (the estimate $\hat{\sigma}_0^2$ calculated under H_1). Hence

$$\begin{aligned} LR_{1,2} &= 2\{c(\mathbf{y}, \hat{\sigma}_1^2) - c(\mathbf{y}, \hat{\sigma}_2^2)\} - D_1/\hat{\sigma}_1^2 + D_2/\hat{\sigma}_2^2 \\ &\simeq (D_2 - D_1)/\hat{\sigma}_{01}^2. \end{aligned}$$

Hence, the likelihood ratio test is again equivalent to the F-test in (6.6).

Since the F-test is asymptotically equivalent to the likelihood ratio test in both limits, we conclude that the F-test is approximately optimal within the range where the F-approximation (6.6) is valid. The t-test (6.4) may be viewed as approximately equivalent to a signed version of the F-test, and hence the same conclusion holds for the t-test.

The same arguments as above show that the conclusion about asymptotic equivalence of the F- and likelihood ratio tests holds in the small-dispersion case if $\hat{\sigma}_0^2$ is replaced by either $\tilde{\sigma}^2$ or $\overline{\sigma}^2$. However, since $\overline{\sigma}^2$ and $\tilde{\sigma}^2$ are not asymptotically equivalent to $\hat{\sigma}_0^2$ for $n \to \infty$, this conclusion does not hold in the large-sample case.

3.6.3 Regularity Conditions and Proofs.

We shall now discuss the necessary regularity conditions needed in order to prove the asymptotic results presented in the previous section. The proof of the results is outlined, in the large-sample case and in the small-dispersion case.

Regularity conditions for Theorem 3.6.1. The functions $t(y, \mu)$ and $f(\lambda, y) = \log a(\lambda, y) + \lambda t(y, \mu)$ are three times differentiable with respect to the parameters, and there exist continuous and integrable functions A_j and B_j, $j = 1, 2, 3$, and a neighbourhood ω of the true value μ, such that for every μ in ω.

$$\left|t^{(j)}(y, \mu)\right| \leq A_j(y), \quad j = 1, 2, 3 \tag{6.8}$$

$$\left|f^{(j)}(\lambda, y)\right| \leq B_j(y), \quad j = 1, 2, 3. \tag{6.9}$$

The function $\mu(\,\cdot\,)$ is three times continuously differentiable, and $\mathbf{X}(\beta)$ has rank k_1 for every β. The matrix $\mathbf{i}(\mu \mid \sigma^2)$ is positive-definite for every μ and σ^2 and $i(\sigma^2 \mid \mu) > 0$ for every μ and σ^2. There exists a positive-definite matrix $\bar{\mathbf{i}}(\beta \mid \sigma^2)$ and a positive number $\bar{i}(\sigma^2 \mid \mu)$, such that

$$\frac{1}{n}\mathbf{i}(\beta \mid \sigma^2) \to \bar{\mathbf{i}}(\beta \mid \sigma^2) \quad \text{for} \quad n \to \infty \tag{6.10}$$

$$\frac{1}{n}i(\sigma^2 \mid \mu) \to \bar{i}(\sigma^2 \mid \mu) \quad \text{for} \quad n \to \infty. \quad \blacksquare \tag{6.11}$$

These regularity conditions were chosen in order to give simple and easily verifiable conditions. It is possible to weaken the conditions considerably, see for example Fahrmeir and Kaufmann (1985), who considered asymptotic theory for generalized linear models.

Theorem 3.6.1. *Under the regularity conditions outlined above, results (6.1) and (6.2) hold. If furthermore there exists a parametrization of H_1 such that H_2 corresponds to the condition $\beta_{k_2+1} = \cdots = \beta_{k_1} = 0$, then (6.3) holds.* ∎

The condition (6.8) is easily seen to hold for an arbitrary exponential dispersion model. Thus, if $Y_i \sim ED(\mu_i, \sigma^2/w_i)$, we have

$$t'(y, \mu) = (y - \mu)/V(\mu).$$

Since $V(\mu)$ is continuous, we have

$$|t'(y, \mu)| \leq \frac{|y - a| + |y - b|}{\inf_{a \leq \mu \leq b} V(\mu)} \quad \text{for } a \leq \mu \leq b$$

The same type of argument applies to $t^{(2)}$ and $t^{(3)}$, and hence we may take ω to be an interval contained in Ω.

Condition (6.9) may be verified for exponential dispersion models using the Fourier inversion formula for characteristic functions. Thus, in the continuous case we obtain from (3.12)

$$a(\lambda, y)\exp(\lambda\theta y) = \frac{\lambda}{2\pi}\int_{-\infty}^{\infty} \exp[\lambda\{\kappa(\theta + is) - is\}]ds. \tag{6.12}$$

Based on (6.12), straightforward analytic arguments show that condition (6.9) is satisfied.

The conditions (6.10) and (6.10) have to be checked for each given sequence of models under consideration. A considerable simplification is obtained in the case of a generalized linear model, where $\mu(\,\cdot\,)$ is differentiable if the link function is differentiable. To illustrate the conditions (6.10) and (6.11) in this case, we consider an example.

Example 3.6.2: Suppose $Y_1, \ldots, Y_n$ are independent and $Y_i \sim Ga(\beta_1 + \beta_2 x_i, \sigma^2)$, $i = 1, 2, \ldots, n$. Then $V(\mu) = \mu^2$ and

$$X(\beta) = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix}^T.$$

Hence

$$\mathbf{i}(\beta) = \sum_{i=1}^{n} \left\{ \begin{matrix} 1/(\beta_1 + \beta_2 x_i)^2 & x_i/(\beta_1 + \beta_2 x_i)^2 \\ x_i/(\beta_1 + \beta_2 x_i)^2 & x_i^2/(\beta_1 + \beta_2 x_i)^2 \end{matrix} \right\}.$$

If (6.10) holds with a positive definite limiting matrix, then

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i^2/(\beta_1 + \beta_2 x_i) > 0 \tag{6.13}$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} 1/(\beta_1 + \beta_2 x_i) > 0 \tag{6.14}$$

$$\lim_{n\to\infty} \left[\frac{1}{n}\sum_{i=1}^{n} \frac{x_i^2}{\beta_1 + \beta_2 x_i} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\beta_1 + \beta_2 x_i} - \left\{ \frac{1}{n}\sum_{i=1}^{n} \frac{x_i}{\beta_1 + \beta_2 x_i} \right\}^2 \right] > 0. \tag{6.15}$$

The conditions (6.13) and (6.14) imply that x_i cannot tend too quickly to zero or too quickly to infinity. For example, (6.14) is violated by the sequence $x_i = i$, and (6.13) is violated by the sequence $x_i = 1/i$. On the other hand, the condition that the sequence of x_is is bounded, together with the restriction $\beta_1 + \beta_2 x_i > 0$, implies (6.13) and (6.14). The condition (6.14) implies that the x_is have a reasonable spread, asymptotically. Thus, if all x_is are equal, except for a finite number of values, then (6.15) is violated. In summary, the conditions (6.13), (6.14) and (6.15) are satisfied if, for example, the x_is are well spread out in a finite interval, a sensible requirement from a practical point of view anyway.

By Example 3.5.3 we have that the information function for σ^2 is of the form

$$\begin{aligned} i(\sigma^2 \mid \mu) = -c''(\sigma^2) \\ + E_{\mu,\sigma^2}\{D(\mathbf{Y},\mu)\}/\sigma^6, \end{aligned}$$

where each of the two terms are proportional to n. Hence $i(\sigma^2 \mid \mu)/n$ is constant, and (6.11) is satisfied. ▮

As the example illustrates, the condition (6.11) is generally very mild, due to the fact that all of the data contribute to the estimation of the dispersion parameter. However, if the weights are not all equal, (6.11) implies certain conditions on the weights, which are generally met if the weights belong to a bounded interval.

Proof of Theorem 3.6.1 (outline). We begin with the large-sample case, outlining the standard arguments leading to (6.1) and (6.3). Let $L(\psi)$ denote the log likelihood for the parameter $\psi = (\beta, \sigma^2)$, and define

$$\mathbf{u}(\psi) = \frac{\partial L}{\partial \psi} \quad \mathbf{i}(\psi) = Var_{\psi}\{\mathbf{u}(\psi)\}.$$

By assumption $\mathbf{i}(\psi)$ is positive-definite. The first step in the argument is to show consistency of the maximum likelihood estimator $\hat{\psi}$, or more precisely to show that with probability tending to 1 for n tending to infinity, the likelihood equation

(6.16) $$\mathbf{u}(\hat{\psi}) = 0$$

has a consistent root. The main argument for this is that, due to $\mathbf{i}(\hat{\psi})$ being positive-definite, $L(\psi)$ will have a local maximum near the true value of ψ for n large.

By the independence of the observations, the score vector $\mathbf{u}(\psi)$ is the sum of n terms. By condition (6.11), the variance of each component becomes negligible compared with the total variance for n large. Hence we may apply the Central Limit Theorem to $\mathbf{u}(\psi)$, obtaining

$$\mathbf{u}(\psi) \sim N(0, \mathbf{i}(\psi)) \tag{6.17}$$

approximately for n large. By expanding $\mathbf{u}(\psi)$ around $\hat{\psi}$ we obtain, by (6.16) and (4.11),

$$\mathbf{u}(\psi) \simeq \mathbf{u}(\hat{\psi}) + \left.\frac{\partial^2 L}{\partial \psi}\right|_{\psi=\hat{\psi}} (\psi - \hat{\psi}) \simeq \mathbf{i}(\psi)(\hat{\psi} - \psi)$$

or, equivalently,

$$\hat{\psi} - \psi \simeq \mathbf{i}^{-1}(\psi)\mathbf{u}(\psi). \tag{6.18}$$

Hence, by (6.17) we obtain

$$\hat{\psi} \sim N(\psi, \mathbf{i}^{-1}(\psi)),$$

approximately for n large. Since $\mathbf{i}(\psi)$ is block diagonal with blocks $\mathbf{i}(\beta \mid \sigma^2)$ and $i(\sigma^2 \mid \mu)$, we have shown (6.1) and (5.4).

To show (6.3), we continue using the above notation, except that we now let $\psi = \beta$, taking σ^2 as known. Expanding $L(\psi)$ around $\hat{\psi}$ and using (4.11) and (6.18), we obtain

$$\begin{aligned} 2\{L(\psi) - L(\hat{\psi})\} &\simeq (\psi - \hat{\psi})^T \left.\frac{\partial^2 L}{\partial \psi \partial \psi^T}\right|_{\psi=\hat{\psi}} (\psi - \hat{\psi}) \\ &\simeq -(\hat{\psi} - \psi)^T \mathbf{i}(\psi)(\hat{\psi} - \psi) \\ &\simeq -\mathbf{u}(\psi)^T \mathbf{i}^{-1}(\psi)\mathbf{u}(\psi). \end{aligned} \tag{6.19}$$

We assume that H_1 has been parametrized such that H_2 is equivalent to $\psi_{k_2+1} = \cdots = \psi_{k_1} = 0$, and we let ψ_0 denote the true value of ψ. There exists a lower triangular matrix $\mathbf{i}^{1/2}(\psi_0)$ such that

$$\mathbf{i}(\psi_0) = \mathbf{i}^{1/2}(\psi_0)\mathbf{i}^{1/2}(\psi_0)^T.$$

Now let us parametrize H_2 by the parameter $\phi = \mathbf{i}^{1/2}(\psi_0)\psi$, which has score function

$$\tilde{\mathbf{u}}(\phi) = \frac{\partial \psi^T}{\partial \phi}\mathbf{u}(\psi(\phi)) = \mathbf{i}^{-1/2}(\psi_0)\mathbf{u}(\psi(\phi))$$

and information matrix

$$\mathbf{i}^{-1/2}(\psi_0)\mathbf{i}(\psi(\phi))\mathbf{i}^{-1/2}(\psi_0)^T.$$

In particular, the information matrix is the identity matrix for $\phi = \phi_0$. Hence, (6.19) takes the form, for $\psi = \psi_0$,

$$\begin{aligned} 2\{L(\hat{\psi}) - L(\psi_0)\} &\simeq \tilde{\mathbf{u}}(\phi_0)^T \tilde{\mathbf{u}}(\phi_0) \\ &= \sum_{j=1}^{k_1} \tilde{u}_j(\phi_0)^2. \end{aligned} \tag{6.20}$$

Now, since $\mathbf{i}^{-1/2}(\psi_0)^T$ is upper triangular, H_2 is equivalent to $\phi_{k_2+1} = \cdots = \phi_{k_1} = 0$. Hence, arguments similar to the above show that, letting $\hat{\psi}^{(2)}$ denote the estimate of ψ under H_2,

$$2\{L(\hat{\psi}^{(2)}) - L(\psi_0)\} \simeq \sum_{j=1}^{k_2} \tilde{u}_j(\phi_0)^2. \tag{6.21}$$

It follows that the difference in deviance between H_2 and H_1, which is the difference between (6.20) and (6.21), is

$$(D_2 - D_1)/\sigma^2 \simeq \sum_{j=k_2+1}^{k_1} \tilde{u}_j(\phi_0)^2. \tag{6.22}$$

By (6.17), and the fact that the information matrix for ϕ is diagonal at ϕ_0, we find that the terms of (6.22) are approximately independent and distributed as $\chi^2(1)$ for n large. This implies (6.3), and hence we have concluded the proof in the large-sample case.

In the small-dispersion case, consider first the case of an exponential dispersion model $Y_i \sim ED(\mu_i, \sigma^2/w_i)$. We show the results under the condition $w_i = r\overline{w}_i$, $i = 1, \ldots, n$, where $\overline{w}_i$ is fixed and r is an integer which tends to infinity. By the standard convolution formula for exponential dispersion models, we may write

$$Y_i = \frac{1}{r}\sum_{j=1}^{r} Y_{ij},$$

where $Y_{i1}, \ldots, Y_{ir}$ are independent, and independent for different i, and $Y_{ij} \sim ED(\mu_i, \sigma^2/\overline{w}_i)$. The log-likelihood for μ based on Y_{ij} is, disregarding terms that to not depend on μ, and letting $\theta_i = \tau^{-1}(\mu_i)$,

$$
\begin{aligned}
L(\mu) &= \sum_{i=1}^{n}\sum_{j=1}^{r}(\overline{w}_i/\sigma^2)\{y_{ij}\theta_i - \kappa(\theta_i)\} \\
&= \sum_{i=1}^{n}(\overline{w}_i/\sigma^2)\{ry_i\theta_i - r\kappa(\theta_i)\} \\
&= \sum_{i=1}^{n}(w_i/\sigma^2)\{y_i\theta_i - \kappa(\theta_i)\}.
\end{aligned}
$$

Hence the likelihood depends on the observations only through $y_i = (y_{i1} + \cdots + y_{ir})/r$, and depends on y_i and μ_i in exactly the same way as for the original observations. Consequently, maximum likelihood estimates, likelihood ratio tests and their distributions are exactly the same as for the original observations. Hence, we may prove the asymptotic results by applying the large-sample results to the rn independent variables Y_{ij}.

The information matrix for $\hat{\beta}$ is $\mathbf{i}(\beta)/\sigma^2$, where

$$\mathbf{i}(\beta) = r\mathbf{X}(\beta)^T\overline{\mathbf{W}}\mathbf{i}(\mu)\mathbf{X}(\beta),$$

where $\overline{\mathbf{W}} = \text{diag}\,\{\overline{w}_1, \ldots, \overline{w}_n\}$. Hence $\mathbf{i}(\beta)/(r\sigma^2)$ is constant as a function of r, which implies condition (6.10) for $rn \to \infty$ with n fixed. By the results of the proof in the large-sample case, we conclude that (6.1) and (6.3) hold for r tending to infinity. Since n is fixed we may now apply (6.3) to the test of H_1 against H_0, and since the difference in deviance between these two hypotheses is D_1, we have shown (6.2) in the case where r tends to infinity. One may show that (6.1)-(6.3) hold under the weaker condition $\min\{w_1, \ldots, w_n\} \to \infty$ too.

Finally, we turn to the case of a dispersion model $Y_i \sim DM(\mu_i, \sigma^2)$ with $t(y, \mu) = t(y - \mu)$, the model for which we showed the saddlepoint approximation. In particular, we assume that the conditions of Theorem 3.3.3 are fulfilled. By (3.7) we obtain

$$Y_i \sim N(\mu_i, \sigma^2 V/w_i), \tag{6.23}$$

approximately for w_i large, where $V = -1/t''(0)$. Expanding $t'(\,\cdot\,)$ around 0 we obtain

$$t'(Y_i - \mu_i) \simeq t'(0) + t''(0)(Y_i - \mu_i) = -(Y_i - \mu_i)/V.$$

Hence, by (6.23) we obtain

$$(w_i/\sigma^2)t'(Y_i - \mu_i) \sim N(0, w_i/(\sigma^2 V)), \tag{6.24}$$

approximately, for w_i large. Furthermore, by (4.20), the information matrix for μ is given by

$$\mathbf{i}_{ii}(\mu \mid \sigma^2) \simeq w_i/(\sigma^2 V). \tag{6.25}$$

In terms of inference on the parameter $\psi = \beta$, (6.24) and (6.25) imply that the key results (6.17) and (6.18) hold approximately for $\mathbf{w} \to \infty$. Continuing the proof as in the large-sample case, we may hence show (6.1)-(6.3). We note that the proof requires (6.8) and (6.9) to be satisfied for continuous functions A_i and B_i, whereas in the large-sample case A_i and B_i are required to be integrable.

In the more general case where $t(y, \mu)$ is not of the form $t(y - \mu)$, the proof of the asymptotic results is given, under further regularity conditions, in Jørgensen (1987b).

§3.7 Analysis of Deviance for Continuous Models

3.7.1 General Points.

We shall now consider the application of the theory developed in the present chapter to some data examples, and make some general considerations about analysis of deviance. At this point, the reader may perhaps want to go back to Chapter 1 and review the approach to data analysis outlined there.

A typical data analysis proceeds via the following steps:

(i) Initial choice of model and estimation of parameters.
(ii) Verification and modification of the model.
(iii) Hypothesis testing.
(iv) Conclusions.

The choice of model often involves several cycles of model verification and subsequent modification of the model. This part also involves analysis of residuals which, in spite of its importance, has not been included in this preliminary version of the text.

Once a well-fitting and theoretically satisfactory model has been found, one may proceed to hypothesis testing (iii). This part may involve the assessment of well-defined scientific hypotheses, or may involve an exploratory search for a parsimonious model, with a descriptive purpose. In the simplest case, the process involves successive reductions of the model, until the smallest model, in terms of the number of parameters, consistent with the data has been found.

The conclusion (iv) involves the interpretation and communication of the results of the analysis within the specific context of the data.

We shall often present the results of the hypothesis testing process in the form of an *analysis of deviance table*. This is a parallel to the analysis of variance table for linear normal models, although the analysis of deviance table is slightly different in form, reflecting the sequential nature of the testing process. The table has the following form

Model	Deviance	d.f	ΔD	$\Delta d.f$	$\hat{\sigma}^2$	F
H_1	D_1	f_1			$\hat{\sigma}^2_{(1)}$	
H_2	D_2	f_2	$D_2 - D_1$	$f_2 - f_1$	$\hat{\sigma}^2_{(2)}$	F_2
H_3	D_3	f_3	$D_3 - D_1$	$f_3 - f_2$	$\hat{\sigma}^2_{(3)}$	F_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Here $H_1 \supseteq H_2 \supseteq \ldots$ is a sequence of nested hypotheses (models), H_i having deviance D_i and degrees of freedom f_i. The table also gives the successive differences between deviances, and between degrees of freedom. Finally it gives estimates for σ^2 (based on a specific estimator) and the corresponding F-test

$$F_i = \frac{(D_i - D_{i-1})/(f_i - f_{i-1})}{\hat{\sigma}^2_{(i-1)}}$$

for testing H_i under H_{i-1}. We use the symbols $\hat{\sigma}_0^2$, $\overline{\sigma}^2$ and $\tilde{\sigma}^2$ to denote respectively the modified profile, Pearson and deviance-based estimate for σ^2.

The F-test may be used to test H_i under H_{i-1}, successively, starting with H_2 under H_1. If the test does not show significance, H_i is accepted. The process continues until a significance is found, and the last model accepted becomes the final model.

A hypothesis of the form $\beta_j = \beta_j^{(0)}$ for a parameter β_j may be tested using the t-test

$$t = (\hat{\beta}_j - \beta_j^{(0)})/\text{se}(\hat{\beta}_j),$$

where $\text{se}(\hat{\beta}_j)$ is the estimated standard error for $\hat{\beta}_j$. We write $\hat{F}_0$, $\overline{F}$, $\hat{t}_0$, $\bar{t}$ etc. to indicate the estimate of σ^2 used. We shall quote the standard error of an estimate in brackets after the estimate, for example $\hat{\beta}_1 = 0.025(0.30)$ means $\hat{\beta}_1 = 0.25$ and $\text{se}(\hat{\beta}_1) = 0.30$. If an estimate for σ^2 is quoted in the same table, this estimate is also used to calculate the standard error.

3.7.2 Energy Expenditure Data.

Consider the energy expenditure data from Section 1.1. In Section 2.3 we argued in favour of a gamma model for these data, of the form

$$\begin{aligned} \overline{Y}_i &\sim Ga(\overline{\mu}_i, \sigma^2/w_i), \quad i = 1, \ldots, n, \\ \overline{\mu}_i &= \beta_1 \overline{x}_{i1} + \beta_2 \overline{x}_{i2}, \quad i = 1, \ldots, n, \end{aligned} \tag{7.1}$$

where $\overline{Y}_i$ is the average energy expenditure per unit of body mass, w_i is the body mass and $\overline{x}_{i1}$ and $\overline{x}_{i2}$ are the proportions of fat and fat-free tissue for individual i. The analysis of deviance table is

Model	Deviance	d.f	$\hat{\sigma}_0^2$	$\tilde{\sigma}^2$	$\overline{\sigma}^2$	$\hat{F}_0$
$\beta_1 \overline{x}_{i1} + \beta_2 \overline{x}_{i1}$	130	102	1.092	1.277	1.296	
$\beta_2 \overline{x}_{i2}$	152	103	1.244	1.477	1.001	20.0

The two F-tests not quoted in table are $\tilde{F} = 17.23$ and $\overline{F} = 16.97$. Hence any of the tree F-tests show a rather strong significance. We conclude that $\beta_1 \neq 0$, which means that the generation of energy in fat tissue is not zero.

For $w = 50kg$ we have $\hat{\sigma}_0^2/w \doteq 0.0218$ under (7.1) which indicates that we may use small-dispersion asymptotics. Nevertheless, the three estimates for σ^2 are somewhat different, the closest agreement being between $\tilde{\sigma}^2$ and $\overline{\sigma}^2$. According to Example 3.5.3, $\tilde{\sigma}^2$ has asymptotic bias approximately 0.0023, calculated from $\hat{\sigma}_0^2$, which is not large enough to explain the differences between the estimates. However, according to the $\sigma^2\chi^2(f)/f$ distribution, the standard deviation of the estimates are approximately 0.15, which makes the differences between the three estimates seem more plausible.

It is also instructive to compare the three estimators for σ^2 under the hypothesis $\beta_1 = 0$. According to the F-test, this hypothesis is not true, and hence we should see some inflation in the estimates compared with H_1. This is the case for $\tilde{\sigma}^2$ and $\hat{\sigma}_0^2$, but not for $\overline{\sigma}^2$, which hence lends less credibility to the latter estimate.

The final estimates for the parameters of the model (7.1) are

$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma}_0^2$
0.304(0.068)	1.359(0.047)	1.277

```
   150.00 |
   145.00 |
   140.00 |                                   *
   135.00 |         *
   130.00 |                      *                 *
   125.00 |         3*                  **  *   *  * *   *
   120.00 |        ***           *    *    *
   115.00 |       *   *         *   * ******  3*
   110.00 |   *     *  *      *  2    2 *2 *  2     * *
   105.00 |                     * 22* 2*    2  ** *
   100.00 |                * **    *3 ** **       *
    95.00 |                     2      **  *
    90.00 |             **    *3     2* *  *      *
    85.00 |         *          2* *
    80.00 |*                ***
    75.00 |
    70.00 |
    65.00 |
    60.00 |         *
    55.00 |
    50.00 |
----------:---------:---------:---------:---------:---------:---------:
        40.0      60.0      80.0     100.0     120.0     140.0     160.0
```

Figure 3.3 The variance for Y as a function of w according to the gamma model, for the energy expenditure data

A 95% confidence interval for β_1 is hence $[0.169, 0.439]$. The estimated coefficient of variation for $\overline{Y}_i$ is, for $w = 50kg, \hat{\sigma}_0/w^{1/2} = 0.160$.

Since the normal distribution is often used in practice in an example like the present, it may be interesting to compare the above analysis with an analysis based on the normal distribution. This model, as introduced in Example 1.1.1, is

$$Y_i \sim N(\mu_i, \rho^2) \quad i = 1, \ldots, n$$
$$\mu_i = \beta_1 x_{i1} + \beta_2 x_{i2} \quad i = 1, \ldots, n,$$

where

$$Y_i = \overline{Y}_i w_i \tag{7.2}$$

is the total energy expenditure and

$$x_{i1} = \overline{x}_{i1} w_i \tag{7.3}$$

is the mass of fat tissue. The estimates for the parameters of this model are

β_1	β_2	ρ^2
0.306(0.073)	1.35(0.050)	107.54

The estimates for β_1 and β_2 and their standard errors are in close agreement with those based on the gamma model, and hence the normal model does not lead to substantially different conclusions. The parameters ρ^2 and σ^2 are not directly comparable, due to the different variance functions, and because of the transformation (7.2). By (7.2) we have

$$Var(Y_i) = w_i^2 Var(\overline{Y}_i)$$

and since, by the gamma model, $Var(\overline{Y}_i) = \sigma^2 \overline{\mu}_i^2 / w_i$, we have

$$Var(Y_i) = \sigma^2 \mu_i^2 / w_i, \tag{7.4}$$

compared with $Var(Y_i) = \rho^2$ for the normal model. However, since μ_i is an increasing function of w_i for the present data, the two variances are not necessarily very different in practice. Figure 3.3 shows the variance (7.4) as a function of w for the data, and although the plot shows a slight linear trend, it is not entirely in disagreement with the hypothesis $Var(Y)$ constant.

Whereas the present analysis does not show any preference for either of the two models, an analysis of residuals might possibly reveal which of the two variance functions, if any, is more appropriate.

3.7.3 Trees Data.

An interesting data set which has been analysed from several points of view in the literature, is the "trees data" (Ryan, Joiner end Ryan, 1985), which is shown in Table 3.2. The variables are diameter (d), height (h) and volume (v) for black cherry trees in Allegheny National Forest, Pennsylvania.

The relation between diameter, height and volume for trees depends on the shape of the tree, and we consider two possibilities. For a cylindrical shape we have the relation

$$v = \frac{1}{4}\pi d^2 h. \tag{7.5}$$

For a conic shape we have

$$v = \frac{1}{12}\pi d^2 h. \tag{7.6}$$

In both cases we have a linear relationship between $\log v$, $\log d$ and $\log h$,

$$\log v = \beta_0 + \beta_1 \log d + \beta_2 \log h. \tag{7.7}$$

Table 3.2: *Trees data, with variables d (diameter in inches at 4.5 feet above ground level), h (height of tree in feet) and v (volume of tree in cubic feet)*

d	h	v
8.3	70	10.3
8.6	65	10.3
8.8	63	10.2
10.5	72	16.4
10.7	81	18.8
10.8	83	19.7
11.0	66	15.6
11.0	75	18.2
11.1	80	22.6
11.2	75	19.9
11.3	79	24.2
11.4	76	21.0
11.4	76	21.4
11.7	69	21.3
12.0	75	19.1
12.9	74	22.2
12.9	85	33.8
13.3	86	27.4
13.7	71	25.7
13.8	64	24.9
14.0	78	34.5
14.2	80	31.7
14.5	74	36.3
16.0	72	38.3
16.3	77	42.6
17.3	81	55.4
17.5	82	55.7
17.9	80	58.3
18.0	80	51.5
18.0	80	51.0
20.6	87	77.0

We shall assume a gamma distribution for the distribution of volume given diameter and height, $V_i \sim Ga(v_i, \sigma^2)$ for the ith tree, the trees being independent. We hence have a generalized linear model with log link. A tentative justification for the gamma distribution is, analogously to the argument for the Energy Data in Section 2.3, that the process of growth of a tree may be described by a stochastic process with independent gamma increments. Although this model may be too simple, by for example not taking into account the environmental factors affecting the growth of the tree, the gamma distribution has the appropriate characteristics of being positive and right-skewed.

The estimates of the parameters of (7.7) are

β_0	−1.635(0.686)	
β_1	1.980(0.075)	
β_2	1.13(0.20)	
$\tilde{\sigma}^2$	0.00655	(d.f. = 28)

Hence β_1 and β_2 are not significantly different from their hypothesized values 2 and 1, respectively. A 95% confidence interval for β_0 is $[-3.04, -0.23]$. The values of β_0 corresponding to (7.5) and (7.6) are respectively −0.24 and −1.34. Both values belong to the confidence interval, although −0.24 (cylindrical shape) seems slightly less plausible than the other value. However, only a set of more extensive data can decide finally between the two models.

In the conic model, the diameter d is measured at the foot of the cone, whereas the diameter was actually measured 1.37m (4.5 feet) above ground level. The resulting model is analyzed in Exercise 3.34.

3.7.4 Permeability of a Building Material.

Plates of building material of a given thickness were subjected to a test, where the penetration time for water was measured. The plates were produced on three machines on each of nine days, with three replications for each combination of day and machine. The data are shown in Table 3.3.

A simple model for permeability is the following. Divide the plate into layers of thickness $w_1, \ldots, w_r$, such that $w = w_1 + \cdots + w_r$ is the total thickness, as indicated in Figure 3.4. We consider a particle travelling across the plate from left to right, and let Z_i be the time spent in the ith layer. If the process is stationary and the material homogeneous, it seems reasonable to assume that the distribution $P(w_i)$ of Z_i satisfies the relation

$$Z_i \sim P(w_i), Z_j \sim P(w_j) \Rightarrow Z_i + Z_j \sim P(w_i + w_j), \quad i \neq j.$$

Table 3.3: *Permeability of building material, given by time to penetration (seconds)*

		Machine		
		1	2	3
Day	1	25.35	20.23	85.51
		22.18	42.46	47.21
		41.50	25.70	25.06
	2	27.99	17.42	26.67
		37.07	15.31	58.61
		66.07	32.81	72.28
	3	82.04	32.06	24.10
		29.99	37.58	48.98
		78.34	44.57	22.96
	4	77.09	47.10	52.60
		30.55	23.55	33.73
		24.66	13.00	23.50
	5	59.16	16.87	20.89
		53.46	24.95	30.83
		35.08	33.96	21.68
	6	46.24	25.35	42.95
		34.59	28.31	40.93
		47.86	42.36	22.86
	7	82.79	16.94	21.28
		85.31	32.21	63.39
		134.59	27.29	24.27
	8	69.98	38.28	48.87
		61.66	42.36	177.01
		110.15	19.14	62.37
	9	34.67	43.25	50.47
		26.79	11.67	23.44
		50.58	24.21	69.02

We assume that $Z_1, \ldots, Z_r$ are independent. From Section 2.3 we know that a convolution family satisfies these requirements. Thus, if $Z_i \sim ED^*(\theta, w_i/\rho^2)$, $i = 1, \ldots, r$, then $Z = Z_1 + \cdots + Z_r$, the total time taken to traverse the plate, has distribution $ED^*(\theta, w/\rho^2)$. If w_i may be taken as arbitrarily small, which may be reasonable to assume if the material is completely homogeneous, then the model is infinitely divisible.

In this model ρ^2 may be in interpreted as the unit in which the thickness of the plate is measured, and since $E(Z_i) = \tau(\theta)w_i/\rho^2$, $\rho^2/\tau(\theta)$ is the average speed of the particle. Hence the parameters of the model have a direct physical interpretation.

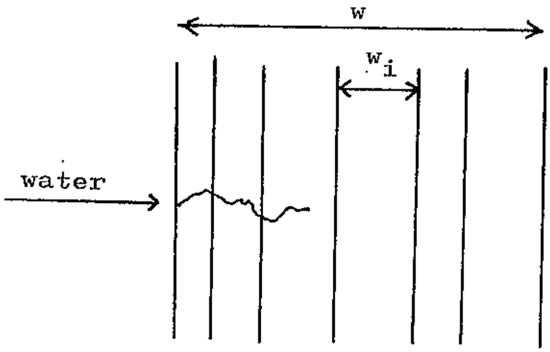

Figure 3.4 Model for water particle traversing a plate.

One further requirement is that the model should not depend on the unit of measurement for time. Hence the family of distributions of Z_i must be closed with respect to scale transformations. By Exercise 2.37 the model hence has power variance function, $V(\mu) = \mu^p$, say. Let us divide Z by w/ρ^2, to obtain an exponential dispersion model

$$\rho^2 Z/w \sim ED(\tau(\theta), \rho^2/w),$$

and finally, by the scale transformation property, we obtain

$$Y = Z/w \sim ED(\mu, \sigma^2/w),$$

where $\sigma^2 = \rho^{2(p-1)}$ and $\mu = \tau(\theta)/\rho^2$.

We conclude that the sample average time per unit of thickness is, according to our model, an exponential dispersion model with expectation μ, the average time for traversing a unit, and dispersion parameter σ^2/w, where w, which plays the role of weight, is the thickness of the plate, and σ^2 is related to the unit of measurement of the thickness. Since all the plates had the same thickness, we take $w_i = 1$ for all i, and hence the final model for the data in Table 3.3 is

$$Y_{ijk} \sim ED(\mu_{ij}, \sigma^2),$$

where i denotes machine, j day and k repetition, and where the variance function is μ^p.

If the particle travels across the plate according to a Brownian motion, the time taken for the particle to travel a given distance is known to be inverse Gaussian distributed, corresponding to $p = 3$. The inverse Gaussian distribution is then called the *first-hitting time distribution* for the Brownian motion process. Similarly, the gamma distribution ($p = 2$) may be interpreted as the first-hitting time distribution for the Poisson process, provided $\lambda = 1/\sigma^2$ is an integer. In fact, for any p in the interval $(1, 2]$ the corresponding exponential dispersion model is a first-hitting time distribution for some stochastic process.

To estimate p we consider the sample variance in group (i, j),

$$s_{ij}^2 = \frac{1}{2}\sum_{k=1}^{3}(Y_{ijk} - \overline{Y}_{ij.})^2,$$

where

$$\overline{Y}_{ij.} = \frac{1}{3}\sum_{k=1}^{3} Y_{ijk}.$$

If σ^2 is small, Y is approximately normally distributed, and hence, approximately

$$s_{ij}^2 \sim \sigma^2 \mu_{ij}^p \chi^2(2)/2,$$

where $\sigma^2 \mu_{ij}^p$ is the variance of Y_{ijk}. If μ_{ij} were known, this model would be a generalized linear model with log link and gamma distributed errors, $s_{ij}^2 \sim Ga(\sigma^2 \mu_{ij}^p, 1)$, with the dispersion parameter known and equal to 1. The systematic part of the model is

$$\log E\{s_{ij}^2\} = \log \sigma^2 + p \log \mu_{ij}. \tag{7.8}$$

Hence we shall estimate p by taking $\mu_{ij} = \overline{Y}_{ij.}$ in (7.8). The estimate of p for the present data is $\hat{p} = 2.29(0.49)$, giving a 95% confidence interval for p of $[1.31, 3.27]$. Hence, both the inverse Gaussian ($p = 3$) and the gamma ($p = 2$) distributions are likely candidates for the distribution of Y_{ijk}. Since $\hat{p}$ is closer to 2, we shall analyse the data using the gamma distribution, $Y_{ijk} \sim Ga(\mu_{ij}, \sigma^2)$.

We shall analyse the data using a two-factor model with interaction and log link,

$$\log \mu_{ij} = \alpha_i + \beta_j + \delta_{ij}.$$

The analysis of deviance table for this model is

Model	Deviance	d.f	$\tilde{\sigma}^2$	$\tilde{F}$	d.f
$\alpha_i + \beta_j + \delta_{ij}$	10.570	54	0.182		
$\alpha_i + \beta_j$	15.362	70	0.219	1.60	16/54
α_i	19.051	78	0.235	2.19	8/70
α	24.640	80	0.294	11.89	2/78

By the F-test for $\delta_{ij} = 0$, there is no interaction between day and machine, and since $F(8, 70)_{0.95} = 2.07$, there is weak evidence of an effect of day. Removing the effect of day, there is a significant effect of machine ($p < 0.0005$).

We shall hence accept the model with only machine effect. The estimates for this model is

Machine	Coefficient	Original scale
1	4.001(0.095)	54.7(5.2)
2	3.360(0.095)	28.8(2.7)
3	3.830(0.095)	46.1(4.4)
$\tilde{\sigma}^2 = 0.244$	d.f = 78	

The estimates show the effect of machine on the log-scale, while the third column of the table shows the effects on the original scale.

As mentioned earlier, the gamma distribution may be interpreted as the first-hitting time distribution for a Poisson process, if $\lambda = 1/\sigma^2$ is an integer. This case corresponds to a physical model in which the plate consists of grains or layers of material, and where a particle travels across the plate from grain to grain or layer to layer. The value $\tilde{\lambda} = 1/\tilde{\sigma}^2 = 4.1$ indicates that, if the Poisson-gamma model is correct, the number of particles or layers is small compared with the thickness of the material.

If our model for permeability is correct the negative of the parameter p represents "coarseness" of the material. The model thus allows us to distinguish between the case of a Poisson process, as above, which corresponds to a gamma distribution for the penetration time, and the case of a Brownian motion, which corresponds to an inverse Gaussian distribution for the penetration time, the latter corresponding to a continuous and uniform type of material. However, a precise estimation of the parameter p requires more information, in the form of a larger sample size, a larger spread of the μ-values, or plates of varying thickness.

3.7.5 Failures of Airconditioning Equipment in Airplanes.

In several cases we have been able to relate the convolution property of an exponential dispersion model with a physical model for the phenomenon under study. This was the case for the example in the previous section, and for the examples in Section 2.3. There are many other possible applications of this kind. For example, the model developed in the previous section for permeability may be a reasonable model for the occurrence of faults in a machine etc. Thus, the time it takes for a crack in a piece of material to develop into a fault may be described by a first-hitting time distribution, provided the development of the crack follows the corresponding stochastic process. However, the following analysis shows that the fault-generating process might be more complicated than this.

Table 3.4 shows the intervals between successive failures of airconditioning equipment in 13 Boeing 720 aircraft (Prochan, 1963). Jørgensen (1982) analyzed this data set using the generalized inverse Gaussian distribution, which is a three-parameter distribution that includes as special cases the gamma distribution, the inverse Gaussian distribution, and the reciprocals of these distributions. Jørgensen's analysis showed that the reciprocal of an inverse Gaussian distribution fits these data, while significant deviations from the gamma distribution and the inverse Gaussian distribution were found. We shall hence analyse the data using the reciprocal inverse Gaussian model.

We assume that the times between failures are $1/Y_{ij}$, where $i = 1, \ldots, 12$ denotes aircraft and $j = 1, \ldots, n_i$ denotes failures. Here we have excluded aircraft 11, which has only two failures. We assume that the Y_{ij} are independent and

$$Y_{ij} \sim IG(1/\mu_i, \sigma_i^2).$$

For σ_i^2 small, μ_i is approximately the expectation of the failure time $1/Y_{ij}$.

We first test equality of $\sigma_1^2, \ldots, \sigma_{12}^2$. As estimates of σ_i^2 we take $\tilde{\sigma}_i^2 = D_i/(n_i - 1)$, where D_i is the deviance for the ith group of observations. The

Table 3.4: *Numbers of operating hours between successive failures of airconditioning equipment in 13 aircraft.*

					Aircraft							
1	2	3	4	5	6	7	8	9	10	11	12	13
194	413	90	74	55	23	97	50	359	50	130	487	102
15	14	10	57	320	261	51	44	9	254	493	18	209
41	58	60	48	56	87	11	102	12	5		100	14
29	37	186	29	104	7	4	72	270	283		7	57
33	100	61	502	220	120	141	22	603	35		98	54
181	65	49	12	239	14	18	39	3	12		5	32
	9	14	70	47	62	142	3	104			85	67
	169	24	21	246	47	68	15	2			91	59
	447	56	29	176	225	77	197	438			43	134
	184	20	386	182	71	80	188				230	152
	36	79	59	33	246	1	79				3	27
	201	84	27	15	21	16	88				130	14
	118	44	153	104	42	106	46					230
	34	59	26	35	20	206	5					66
	31	29	326		5	82	5					61
	18	118			12	54	36					34
	18	25			120	31	22					
	67	156			11	216	139					
	57	310			3	46	210					
	62	76			14	111	97					
	7	26			71	39	30					
	22	44			11	63	23					
	34	23			14	18	13					
		62			11	191	14					
		130			16	18						
		208			90	163						
		70			1	24						
		101			16							
		208			52							
					95							

deviance for the inverse Gaussian distribution is known to be exactly (not just asymptotically) χ^2-distributed,

$$\tilde{\sigma}_i^2 \sim \sigma_i^2 \frac{\chi^2(n_i-1)}{n_i-1} = Ga(\sigma_i^2, 2/(n_i-1)).$$

Hence, we use the gamma distribution, with known dispersion parameter and weights $n_i - 1$ for testing the hypothesis $\sigma_1^2 = ... = \sigma_{12}^2$. Under the hypothesis, the scaled deviance (deviance divided by the value of the dispersion parameter) for the gamma distribution is 17.0, which is not significant compared with the $\chi^2(11)$-distribution.

We proceed to test equality of $\mu_1, \ldots, \mu_{12}$. The result is shown in the following analysis of deviance table.

Model	Deviance	d.f	$\tilde{\sigma}^2$	F	d.f
μ_i	13391	199	67.29		
μ	15028	210	71.56	2.21	11/199

It may be shown that the F-test for the hypothesis $\mu_1 = \cdots = \mu_{12}$ is exactly F-distributed in this particular case, which is comfortable, because the large value of $\tilde{\sigma}^2$ does not allow us to rely on small-dispersion asymptotics. The F-test shows significance at the 2.5% level, and hence we have moderately strong evidence that the aircraft are different in terms of the time between failures of the airconditioning equipment. The estimates of the μ_is are

Aircraft	Estimate
1	36.0 (20.1)
2	30.7 (9.5)
3	42.4 (9.9)
4	40.2 (13.4)
5	61.1 (17.1)
6	11.2 (5.0)
7	14.5 (6.0)
8	18.3 (7.2)
9	8.6 (8.0)
10	17.7 (14.1)
11	14.8 (9.1)
12	42.9 (13.4)
$\tilde{\sigma}^2 = 67.29$	$d.f = 199$

It is worth noting that the estimate $1/\hat{\mu}_i$, being the average of $Y_{i1}, \ldots, Y_{in_i}$, has distribution $IG(1/\mu_i, \sigma^2/n_i)$.

It remains to interpret the reciprocal inverse Gaussian distribution for these data. A possible interpretation comes from the fact that for $1/U \sim IG(1/\mu, \sigma^2)$, U may be decomposed as

$$U = U_1 + U_2, \tag{7.8}$$

where U_1 and U_2 are independent, $U_1 \sim Ga(\sigma^2, 2)$ and $U_2 \sim IG(\mu, \mu^2/\sigma^2)$. This result may be shown by using moment generating functions. By (7.8), the expectation of U is $\mu + \sigma^2$ which, by the value of $\tilde{\sigma}^2$, is considerably larger than μ for the aircraft data.

We may think of U_1 and U_2 as respectively the time it takes before a new crack in a piece of material starts to develop, and the time it takes for the crack to develop into a failure. The latter process could be described by a Brownian motion, leading to an inverse Gaussian first-hitting time distribution for U_2. Such a model might be realistic for complicated machinery, where the internal forces produced during use provoke minute cracks from existing microscopic weaknesses in the material, after which the crack develops into a break, producing a failure.

3.7.6 Data on Wind Direction.

In Sections 1.3 and 1.5 we considered data on wind directions. The proposed model was the von Mises-Fisher distribution,

$$Y_{ij} \sim vM(\mu_i, \sigma^2),$$

where i denotes season and j denotes measurement within season.

One possible justification for this model is as follows. Suppose that, for a given season, there is a predominant wind direction, determined by the direction between a low-pressure centre and a high-pressure centre. Suppose the actual locations of the centres are independent, and follow two-dimensional normal distributions,

$$U_i \sim N_2(v_i, \rho^2 I), \quad i = 1, 2.$$

The wind direction is determined by the difference between the two centres,

$$U = U_2 - U_1 \sim N_2(v_2 - v_1, 2\rho^2 I).$$

The conditional distribution of the direction $U/\|U\|$ given $\|U\| = u$ is then a von Mises-Fisher distribution $vM(\mu, \sigma^2)$, where μ is the angle between v_1 and v_2 and

$$\sigma^2 = 2\rho^2/ru, \tag{7.9}$$

where $r = \|v_2 - v_1\|$, as shown in Exercise 3.34.

It might seem more natural to consider the marginal distribution of the direction, instead of the conditional distribution given $\|U\|$. However, the dispersion parameter (7.9) has the right qualitative form. It decreases with increasing distance between the actual wind centres v_1 and u_2, and it is proportional to the ratio between the variance parameter for $U_2 - U_1$ and the distance between the theoretical wind centres.

The analysis of deviance table for comparison of the four seasons is

Model	Deviance	d.f	$\tilde{\sigma}^2$	$\hat{\sigma}_0^2$	$\tilde{F}$
μ_i	69.4510	45	1.54	∞	
μ	86.2458	48	1.80	∞	3.63

As noted in Section 1.5, the value of $\tilde{\sigma}^2$ indicates that σ^2 is not small enough to justify the use of small-dispersion asymptotics for these data, which is confirmed by the infinite value for $\hat{\sigma}_0^2$. In fact, $\sigma^2 = \infty$ corresponds to the uniform distribution, that is, no preferred wind direction. Hence, although the value $\tilde{F} = 3.63$ is significant at the 2.5% level compared with the $F(3, 44)$ distribution, we can not definitively reject the hypothesis of a common wind direction for the four seasons.

The estimates of the wind directions for the four seasons are (in degrees)

Winter	272 (24)
Spring	330 (26)
Summer	57 (26)
Autumn	232 (28)
$\tilde{\sigma}^2 = 1.54$	$d.f. = 45$

The estimates indicate that the wind directions could be equally spaced around the circle, corresponding to the hypothesis

$$\mu_i = (\mu + i\pi/2) \bmod 2\pi, \quad i = 1, 2, 3, 4.$$

More sophisticated models for the seasonal variation of wind direction are possible, but the present data do not contain much more information in this respect.

3.7.7 Use of Outside Labour Power for Amazonian Peasants.

In a study of the conditions of migrants in the Amazonian area of Brazil, Botelho (1989) followed 210 families from two settlements. One economic and social indicator for a family is the use of outside labour power. We analyse the variable $Y = Z/w$,where Z is the amount (Cr$) spent by the family per year on hiring outside labour power, and w denotes the number of working members of the family. The reason for standardizing by w is that the economic condition of a given family is known to be in fairly direct proportion to the number of working members of the family.

The variable Y may be zero with positive probability, because some families do not use any outside labour power. One distribution with this characteristic is an exponential dispersion model with power variance function, provided the power p belongs to the interval $1 < p < 2$. In the present analysis we chose $p = 1.75$, although in principle, p should be estimated from the data.

It is difficult to give more detailed arguments in favour of this particular exponential dispersion model, but by the standard convolution formula for exponential dispersion models, the model has the attractive feature that if we wish to combine the values of Y for two families, this must be done by weighted averaging using w as weight, corresponding to the sum of the values of Z.

Letting $Y_i \sim ED(\mu_i, \sigma^2/w_i)$ denote the distribution of Y for the ith family, we shall use the log link, and assuming independence of the Y_is, we thus have a generalized linear model. The data contained a large number of independent variables, and we eliminated many of the variables using the t-test based on the estimate $\overline{\sigma}^2$. The parameter estimates for the final model were

Variable	Coefficient	se
Gross income	0.0030	0.0004
Wages received	−0.022	0.004
Family labour power	−0.0060	0.0007
Family size (settlement 1)	0.17	0.08
Family size (settlement 2)	−0.21	0.18
No debt (settlement 1)	−5.3	0.3
No debt (settlement 2)	−5.7	0.3
Debt (settlement 1)	−27	12
Debt (settlement 2)	−6.9	0.8
$\overline{\sigma}^2$	1199	d.f. = 201
$\tilde{\sigma}^2$	990	

The estimates show that family labour power has a significant negative effect on the use of outside labour power. Thus, families with many working members tend to hire less outside labour power per working family member. Similarly, gross income has a positive effect, and wages received has a negative effect on the use of outside labour power. The family size (number of working and non-working members) seems to have opposite effects in the two settlements, although the effects are hardly significant.

For the qualitative variable debt/no debt, there seems to be a significant difference between the two settlements, in that families in settlement 1 with debt hardly use any outside labour power at all, compared with other families in either settlement. For the quantitative variables, the mean of the variables was subtracted in each group, and hence the conclusion holds for an "average" family in the settlement. This difference between settlements indicates a more dramatic economic differentiation (in terms of the use of outside labour power) in settlement 1, which consists of spontaneous settlers, compared with settlement 2, which was sponsored by the federal government. An overall comparison of the two settlements based on an F-test ($F = 3.73$, d.f. $= 3, 201$, $p \simeq 0.01$) shows that there probably is a real difference between the settlements. For comparison, note that $\tilde{F} = 4.52$, which is rather different, although in the present case $\tilde{F}$ would lead to the same conclusion.

§3.8 Analysis of Deviance for Discrete Exponential Dispersion Models

3.8.1 General Points.

Most of what has already been said about analysis of deviance in the continuous case (Section 3.7.1), continues to apply in the discrete case. The main difference is that in the discrete case, the dispersion parameter is nearly always known. This is the case, for example, for the binomial distribution and for the Poisson distribution, the latter having only one parameter. In Section 2.5.3 we considered two parameters for the Poisson distribution, but since these parameters are not identifiable, there is in practice just one parameter.

A discrete exponential dispersion model is necessarily on the convolution family form,

$$p^*(z;\theta,\lambda) = a^*(\lambda, z)\exp\{\theta z - \lambda\kappa(\theta)\}, \quad z \in \mathbf{N}_0.$$

In this case, we define the deviance as $D(\mathbf{z}/\lambda, \mu)$ for a vector $\mathbf{z} = (z_1, \ldots, z_n)^T$ of independent observations.

We shall use the likelihood ratio test, or equivalently, the difference between deviances, for testing hypotheses. In the analysis of deviance table we report the deviance, degrees of freedom, and their first differences. For a table like

Model	Deviance	d.f.	ΔD_i	$\Delta d.f$
H_1	D_1	f_1		
H_2	D_2	f_2	$D_2 - D_1$	$f_2 - f_1$
H_3	D_3	f_3	$D_3 - D_2$	$f_3 - f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

we test H_i under H_{i-1} by the χ^2-test

$$\Delta D_i/\sigma^2 = (D_i - D_{i-1})/\sigma^2$$

and compare with a $\chi^2(f_i - f_{i-1})$-distribution. The quantity D_i/σ^2 is called the ***scaled deviance.*** We normally assume that the model is parametrized such that $\sigma^2 = 1$, which makes the scaled and unscaled deviances equal. If $\sigma^2 \neq 1$ its value will be given in the table. For testing individual parameters, we use a normal reference distribution. Thus, if $\hat{\beta}_j$ is an estimate with standard error $\text{se}(\hat{\beta}_j)$, the test

$$u = \frac{\hat{\beta}_j - \beta_j^0}{\text{se}(\hat{\beta}_j)}$$

has an asymptotic normal $N(0, 1)$ distribution for $\beta_j = \beta_j^0$.

The deviance D_i for a hypothesis H_i is an obvious measure of fit of the hypothesis, and in the small-dispersion case D_i has an asymptotic χ^2-distribution, allowing us to make a formal test for goodness of fit of H_i. However, since the χ^2-approximation does not hold in the large-sample case, one should be very careful with this test. Moreover, for the purpose of checking the model, a single measure of fit is not enough. Only a thorough inspection of the fit, including an analysis of residuals, can do this. In short, we shall emphasize the use of *differences* between deviances, rather than the deviances themselves.

3.8.2 Beetle Data.

In Section 1.2, we considered data on beetle mortality as a function of the dose x of a poison. Our model for this data is $Y_1, \ldots, Y_n$ independent, where

Y_i is the number of dead insects out of n_i in the ith experiment. If the insects in a given experiment are independent with respect to the action of the poison, and all have the same probability of dying, we have a binomial distribution

$$Y_i \sim Bi(n_i, \mu_i).$$

Assuming a generalized linear model with logit link we obtain the model

$$\log \frac{\mu_i}{1-\mu_i} = \beta_0 + \beta_1 x_i. \tag{8.1}$$

The analysis of deviance table for testing $\beta_1 = 0$ is

Model	Deviance	d.f.	ΔD_i	d.f.
$\beta_0 + \beta_1 x_i$	11.23	6		
β_0	284.20	7	273	1

Hence the χ^2-test for $\beta_1 = 0$ is 273 with 1 degree of freedom, which shows an overwhelming significance against this hypothesis. This result confirms the toxic effect of the poison.

The parameter estimates under the model (8.1) are

$\hat{\beta}_0 = -60.72(5.18)$	$\hat{\beta}_1 = 34.27(2.91)$

This is an example where the estimated correlation between estimates is very high, with a value of -0.9997. This is caused by the fact that the x-values are concentrated in a narrow interval far from the origin. To correct for this fact, we subtract the average value $\overline{x}$ from each x-value, leading to the parametrization

$$\log \frac{\mu_i}{1-\mu_i} = \alpha + \beta_1(x_i - \overline{x}),$$

where $\alpha = \beta_0 + \beta_1\overline{x}$. The estimate for α is $\hat{\alpha} = 0.74(3.57)$, and the estimated correlation between $\hat{\alpha}$ and β_1 is reduced to the value 0.020. The estimated probability of dying for $x = \overline{x}$ is $e^{\hat{\alpha}}/(1+e^{\hat{\alpha}}) = 0.6769$.

To obtain reliable estimates of the probability of an insect dying for a given dose x, particularly for high or low value of x if is important to analyse the data using a correct link function. For this purpose it may be necessary to try alternative link functions, or to use a parametrized family of link functions, the latter case meaning that we are effectively estimating the link function from the data,

3.8.3 Paired Comparisons.

The quality of an industrial product is difficult to measure in quantitative terms, because it may involve many different aspects, such as taste, colour, appearance etc. However, if presented with two items for comparison, a person may express his preference for one or the other, thus producing a binary outcome. This technique is called the *method of paired comparisons.*

Assume that k objects are compared in pairs, and let n_{ij} be the number of times i is compared with j. If we use different judges for each of the n_{ij} comparisons, and if the judges all have the same probability μ_{ij} for preferring item i over item, j, we have a binomial model,

$$Y_{ij} \sim Bi(n_{ij}, \mu_{ij}), \quad 1 \leq i < j \leq k,$$

where Y_{ij} is the number of times i was preferred to j. We also assume that the Y_{ij}s are independent. Table 3.5 shows a set of data on the effect of an additive, monosodium glutamate, on the flavour of apple sauce (Sinclair, 1982).

Table 3.5: *Data for effect of monosodium glutamate on apple sauce*

Items compared i j	No. of times i preferred to j x_{ij}	No. of times i compared to j n_{ij}
1 2	3	4
1 3	3	4
1 4	3	4
2 3	3	4
2 4	4	4
3 4	0	4

A simple model for the probability of a judge preferring i over j is

$$\mu_{ij} = \frac{\delta_i}{\delta_i + \delta_j},$$

where $\delta_1, \ldots, \delta_k$ are positive parameters. On the logistic scale, the model is

$$\log \frac{\mu_{ij}}{1 - \mu_{ij}} = \beta_i - \beta_j,$$

where $\beta_i = \log \delta_i$. The model thus ranks the observations on an interval scale with $\beta_i - \beta_j$ indicating the magnitude of the preference for i over j on the logistic scale. The model is a generalized linear model with logit link. Since the model is over-parametrized, we take $\beta_k = 0$.

The deviance for the model, for the data in Table 3.5 is 6.937 on 3 degrees of freedom, and the parameter estimates are

$\hat{\beta}_1 = 1.21(0.84)$	$\hat{\beta}_2 = 0.89(0.81)$
$\hat{\beta}_3 = -1.00(0.87)$	$\beta_4 = 0\ (-)$

The treatments corresponding to $i = 1, 2, 3$ corresponded to increasing amounts of additive, and treatment 4 is a control with no additive. The two smaller doses of additive ($i = 1, 2$) were thus well accepted by the judges compared with the control, whereas the highest dose of additive ($i = 3$) was less well accepted. However, the standard errors of the estimates are fairly large, and hence it is not possible to draw any firm conclusions from the present data.

3.8.4 Highway Accidents.

In Section 2.3 we considered data Y_{ij}, the number of road accidents on straight highways in Denmark in 1963, where $j = 1, 2, 3, 4$ is season and $i = 1, 2$ is road type, respectively 3-lane and 4-lane. We argued in Section 2.3 in favour of a discrete exponential dispersion model for Y_{ij}, the Y_{ij}s being independent, and here we shall analyse the simplest case, a Poisson distribution. According to the properties of the Poisson process, this model is appropriate if cars act independently of each other, accidents in non-overlapping time intervals occur independently of each other, and if the number of cars is large, and each car has a small probability of being involved in an accident in a given small time interval, this probability being the same for all cars.

If $Y_{ij} \sim Po(\mu_{ij})$, we assume that

$$\log \mu_{ij} = \alpha_i + \beta_j, \tag{8.2}$$

which means that we have a generalized linear model with log link. We take $\alpha_1 = 0$, such that α_2 is the difference between the two road types. The analysis of deviance table for this model is

Model	Deviance	d.f.	ΔD_i	d.f.
$\alpha_i + \beta_j$	7.2818	3		
β_j	481.55	4	474.3	1
α_i	24.843	6	17.56	3

The table shows the differences in deviance relative to the first model, for both the second and the third model.

The deviance for the model (8.2), compared with $\chi^2(3)_{0.95} = 7.81$, suggests a certain lack of fit for this model, although the χ^2-approximation to the distribution of the deviance may be questioned, as mentioned in Section 3.8.1. Figure 3.5 shows a plot of $\log Y_{ij}$ versus j, which shows a reasonable constancy of $\log Y_{1j} - \log Y_{2j}$, confirming the model. The parameter estimates are

$\alpha_1 = 0(\text{-})$	$\hat\beta_1 = 3.744(0.092)$
	$\hat\beta_2 = 3.769(0.092)$
$\hat\alpha_2 = 1.512(0.079)$	$\hat\beta_3 = 3.924(0.088)$
	$\hat\beta_4 = 4.057(0.086)$

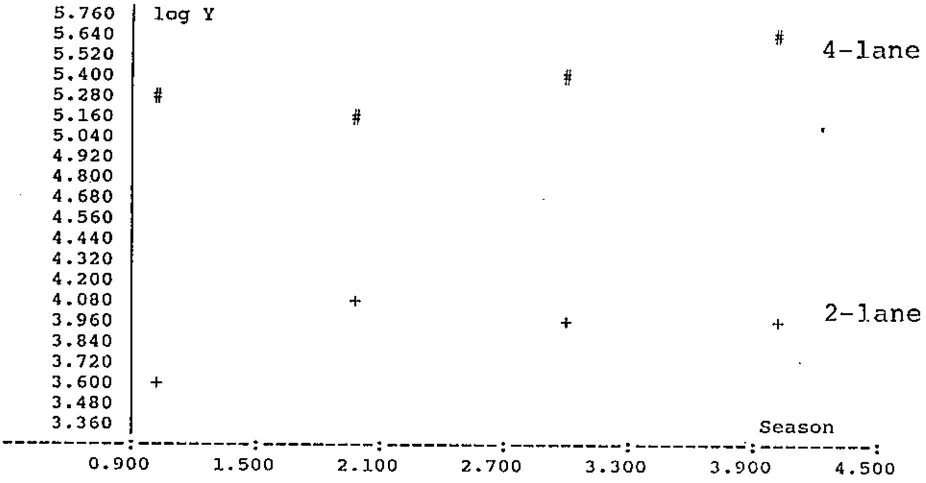

Figure 3.5 Plots of $\log Y_{ij}$ versus j for the data in Table 2.6

The estimates for β_j show that there is an increasing number of accidents through the year for both road types.

§3.9 Notes

One of the important achievements of Nelder and Wedderburn's (1972) paper, in which generalized linear models were introduced, was to suggest analysis of deviance as a unified method for inference in a number of different statistical models. Analysis of deviance thus generalizes analysis of variance for normal models, analysis of log-linear models for contingency tables, and probit analysis for binomial data.

The generalization of analysis of deviance to dispersion models was proposed by Sweeting (1981) and Jørgensen (1983), and Sweeting (1984) showed that analysis of deviance may, to a certain extent, be generalized to location and scale models,

Small-dispersion asymptotic results have been in use for long in various areas, for example in analysis of contingency tables and binomial data. As a general method for dispersion models and exponential dispersion models, it was proposed by Jørgensen (1987a,b).

The saddlepoint approximation was introduced by Daniels (1954), and has recently enjoyed renewed interest, see e.g. Barndorff-Nielsen and Cox (1979) and Lugannani and Rice (1980). A recent survey of this topic was given by Daniels (1987).

Exercises

Exercise 3.1: Write and test a set of GLIM macros for fitting the exponential dispersion model with variance function $V(\mu) = \exp(-\mu)$. The test should include fitting one or more models for a set of data. Use the identity link, or a link of your own choice.

Exercise 3.2: Show that the deviance for the normal distribution has the form

$$D(\mathbf{y}, \mu) = \sum_{i=1}^{n} w_i(y_i - \mu_i)^2.$$

Exercise 3.3: Show that the deviance for the inverse Gaussian distribution has the form

$$D(\mathbf{y}, \mu) = \sum_{i=1}^{n} w_i \frac{(y_i - \mu_i)^2}{y_i \mu_i^2}$$

Show that if $D_1 = D(\mathbf{y}, \hat{\mu})$ is the deviance for a given hypothesis, then the maximum likelihood estimate of σ^2 is D_1/n.

Exercise 3.4: Find the deviance of the t-distribution (Section 1.3), with probability density function

$$p(y; \mu, \lambda) = a(\lambda)(1 + (y - \mu)^2)^{-\lambda}, \quad y \in \mathbb{R}.$$

Exercise 3.5: The hyperbola distribution is defined by the probability density function

$$p(y; \mu, \lambda) = a(\lambda) y^{-1} \exp\{-(\lambda/2)(y/\mu + \mu/y)\}, \quad y > 0,$$

where $\lambda > 0$ and $\mu > 0$ are parameters. Show that this is a dispersion model, and find the deviance. Show that the distribution of $u = \log y$ is of the form (3.6).

Exercise 3.6: The symmetric hyperbolic distribution is defined by the probability density function

$$p(y; \mu, \lambda) = a(\lambda) \exp[-\lambda\{1 + (y - \mu)^2\}^{1/2}], \quad y \in \mathbb{R}.$$

Show that this is a dispersion model, and find the deviance.

Exercise 3.7: Show that the normal and inverse Gaussian distributions are steep.

Exercise 3.8: Plot the deviance of the inverse Gaussian distribution ($n = 1$) as a function of y for some values of μ.

Exercise 3.9: Show that the saddlepoint approximation is exact for the normal distribution.

Exercise 3.10: Show that the saddlepoint approximation is exact for the inverse Gaussian distribution.

Exercise 3.11: Calculate the deviance for the generalized hyperbolic secant distribution, verifying the result in Table 3.1. Find the saddlepoint approximation to the distribution.

Exercise 3.12: Calculate the deviance for the exponential dispersion model with variance function $V(\mu) = e^{-\mu}$, verifying the result in Table 3.1. Find the saddlepoint approximation for this distribution.

Exercise 3.13: Show that the exponential dispersion model with power variance function $V(\mu) = \mu^p$ is steep in the cases $1 < p < 2$ and $p > 2$. Calculate the deviance of the model in these cases, verifying the result in Table 3.1, and find the saddlepoint approximation to the distribution.

Exercise 3.14: Find the saddlepoint approximation to the t-distribution in Exercise 3.4.

Exercise 3.15: Find the saddlepoint approximation to the hyperbola distribution in Exercise 3.5. Hint: consider the distribution of $\log y$.

Exercise 3.16: Make a numerical or graphical evaluation of the approximation $D(y, \mu) \simeq (y - \mu)^2/V(\mu)$ for the gamma distribution, and for the inverse Gaussian distribution.

Exercise 3.17: Make a numerical or graphical evaluation of the approximation $D(y, \mu) \simeq (y - \mu)^2/V$ for the symmetric hyperbolic distribution in Exercise 3.6, and for the von Mises-Fisher distribution.

Exercise 3.18: Let $D(y, \mu)$ denote the deviance of a single observation from an exponential dispersion model $ED(\mu, \sigma^2)$. Show that $D(y, \mu)$ is convex as a function of y for any given μ, and strictly convex when the model is steep.

Exercise 3.19: Make a plot of the deviance $D(y, \mu)$, for a single observation from the exponential dispersion model with power variance function $V(\mu) = \mu^p$ $(p < 0)$, as a function of y. Make sure that the plots illustrates the fact that the deviance is convex, but not strictly convex, as a function of y, due to the fact that the model is not steep, cf. Exercise 3.18.

Exercise 3.20: Find $E(\log Y)$ in the case $Y \sim Ga(\mu, \sigma^2)$. Hint: use the expression for the expectation of the deviance.

Exercise 3.21: Find $E(1/Y)$ in the case $Y \sim IG(\mu, \sigma^2)$. Hint: see Exercise 3.20.

Exercise 3.22: Consider n independent observations $Y_1, \ldots, Y_n$ with $Y_i \sim IG(\mu_i, \sigma^2/w_i)$, $i = 1, \ldots, n$. Show that the deviance has a χ^2-distribution

$$D(\mathbf{Y}, \mu)/\sigma^2 \sim \chi^2(n).$$

Hint: Find the moment generating function of the deviance in the case $n = 1$.

Exercise 3.23: Consider a model with log likelihood

$$L(\mu) = \sum_{i=1}^{n} \log p(y_i; \mu),$$

where $p(\,\cdot\,; \mu)$ is a probability density function and $\mu = (\mu_1, \ldots, \mu_k)^T$ is a parameter vector. Define the score vector by $\mathbf{u}(\mu) = \partial L/\partial \mu$, and let $\mathbf{i}(\mu) =$

$Var_\mu(\mathbf{u}(\mu))$ denote the Fisher (expected) information matrix for μ. Show that if the operation of integration and differentiation may be interchanged, then
(i) $E_\mu(\mathbf{u}(\mu)) = 0$
(ii) $\mathbf{i}(\mu) = E_\mu\{\mathbf{u}(\mu)\mathbf{u}(\mu)^T\} = -E\{\frac{\partial^2 L}{\partial\mu\partial\mu^T}\}$.
Hint: differentiate the relation $\int p(y;\mu)dy = 1$ twice with respect to μ.

Exercise 3.24: Using the same notation and assumptions as in Exercise 3.23, consider the hypothesis

$$H_1 : \mu = f(\beta),$$

where $\beta = (\beta_1, \ldots, \beta_{k_1})^T$. Assume that f is differentiable, and define

$$\mathbf{X}(\beta) = \frac{\partial f}{\partial \beta^T}.$$

Show that the score vector for β is

$$\mathbf{u}^{(1)}(\beta) = \mathbf{X}(\beta)^T\mathbf{u}(f(\beta)).$$

Show that the expected information matrix for β is

$$\mathbf{i}(\beta) = \mathbf{X}(\beta)^T\mathbf{i}(f(\beta))\mathbf{X}(\beta).$$

Show that the observed information matrix for β is $\mathbf{j}(\hat{\beta})$, where

$$\mathbf{j}(\beta) = \mathbf{X}(\beta)^T\{-\frac{\partial^2 L}{\partial\mu\partial\mu^T}\}\mathbf{X}(\beta) - \sum_{i=1}^{n}\frac{\partial^2\mu_i}{\partial\beta\partial\beta^T}u_i(\mu).$$

Exercise 3.25: Find $i(\sigma^2 \mid \mu)$ for each of the three models $Y_i \sim N(\mu_i, \sigma^2/w_i)$, $Y_i \sim Ga(\mu_i, \sigma^2/w_i)$ and $Y_i \sim IG(\mu_i, \sigma^2/w_i)$ where $Y_1, \ldots, Y_n$ are independent random variables, and $\mu = (\mu_1, \ldots, \mu_n)^T$ is given by the model $\mu = \mu(\beta)$.

Exercise 3.26: Show that $\hat{\sigma}_0^2 = \overline{\sigma}^2 = \tilde{\sigma}^2$ for the normal model.

Exercise 3.27: Consider the model

$$Y_i \sim Ga(\beta_0 + \beta_1 x_i, \sigma^2/w), \quad i = 1, 2, \ldots, 10$$

where $x_i = i$, $i = 1, 2, \ldots, 10$. Conduct a small simulation experiments, comparing the estimators $\tilde{\sigma}^2$ and $\overline{\sigma}^2$. Use for example the values $\beta_0 = 1$, $\beta_1 = 2$, $\sigma^2 = 1$ and $w = 1, 2, 5$. Hint: $\sigma^2/w = 1$ corresponds to the exponential distribution, which is easily simulated. To obtain simulated gamma variables for $w = 2$ or 5, use the convolution formula.

Exercise 3.28: Write a small computer program to calculate the estimator $\tilde{\sigma}_0^2$ given the value of the deviance, assuming a gamma distribution, and that the weights w_i are all equal to 1. Use the approximation

$$\psi(\lambda) \simeq \log\lambda - 1/(2\lambda) - 1/(12\lambda^2)$$

for the digamma function.

Exercise 3.29: Consider the hyperbola distribution on the form

$$p(y;\eta,\lambda) = \{2K_0(\lambda)y\}^{-1}\exp\{-(\lambda/2)(y/\eta+\eta/y)\}, \quad y > 0$$

where

$$K_0(\lambda) = (1/2)\int_0^\infty y^{-1}\exp\{-(\lambda/2)(y+y^{-1})\}dy$$

is the modified Bessel function of the second kind and order 0. Define

$$K_1(\lambda) = (1/2)\int_0^\infty \exp\{-(\lambda/2)(y+y^{-1})\}dy,$$

which is the modified Bessel function of the second kind and order 1. Derive the relevant formulas and equations for the estimates $\tilde{\sigma}_0^2$, $\tilde{\sigma}^2$ and $\overline{\sigma}^2$ for this distribution, expressed in terms of K_0 and K_1.

Exercise 3.30: Consider the model $Y_i \sim ED(\mu,\sigma^2)$, for $Y_1,\ldots,Y_n$ independent. Define the Pearson statistic by

$$X^2 = \sum_{i=1}^n (Y_i-\hat{\mu})^2/V(\hat{\mu}),$$

where $\hat{\mu} = n^{-1}(Y_1+\cdots+Y_n)$. Show that X^2 is asymptotically normally distributed for n tending to infinity, and find the parameters in the asymptotic distribution. Hint: Express X^2 in terms of $\hat{\mu}$ and $S = \sum Y_i^2$, show that $(\hat{\mu}, S)$ is asymptotically normally distributed, and use the δ-method.

Exercise 3.31: Show that the von Mises-Fisher distribution satisfies the conditions of Theorem 3.6.1.

Exercise 3.32: Let $Y_1,\ldots,Y_n$ be independent with distribution

$$Y_i \sim Ga((\beta_0+\beta_1 x_i)^2,\sigma^2), \quad i = 1,\ldots,n.$$

Examine the condition that $\lim n^{-1}\mathbf{i}(\beta)$ be positive-definite for this model.

Exercise 3.33: Analyse the Energy Expenditure Data (Example 1.1.1) using the model

$$\overline{Y}_i \sim N(\beta_2+(\beta_1-\beta_2)\overline{x}_{i1},\sigma^2/w_i), \quad i = 1,\ldots,n,$$

in the notation of the example. Compare the analysis with the analysis in Section 3.7.2.

Exercise 3.34: If the diameter of a tree with conic shape is measured at distance k above ground level, show that the volume of the tree is

$$V = \frac{\pi}{12} d^2 h^3 / (h - k)^2.$$

Analyse the data in Section 3.7.3 using this model with $k = 1.37(4.5ft)$.

Exercise 3.34: Let Y follow a two-dimensional normal distribution $N_2(\mu, \sigma^2 I)$.

a) Show that the density of Y may be written in the form

$$f(y_1, y_2) = (2\pi)^{-1} \exp\left[-\frac{1}{2\sigma^2}\{y_1^2 + y_2^2 + \mu_1^2 + \mu_2^2 - 2(y_1\mu_1 + y_2\mu_2)\}\right].$$

b) Show that the conditional distribution of Y/R given $R = r$, where $R^2 = Y_1^2 + Y_2^2$, is a von Mises-Fisher distribution $vM(\theta, \sigma^2/rm)$, where $m^2 = \mu_1^2 + \mu_2^2$ and θ is the angle given by $\tan\theta = \mu_2/\mu_1$.

REFERENCES

Apart from references cited in the text, the list contains some additional references.

Abramowitz, M. and Stegun, I.A. (1972) *Handbook of Mathematical Functions*. New York: Dover.

Aitkin, M. and Clayton, D. (1980). The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Appl. Statist.* **29**, 156-163.

Baker, R.J. and Nelder, J.A. (1978). The GLIM System, Release 3. Oxford: Numerical Algorithms Group.

Bar-Lev, S.K. and Enis, P. (1986). Reproducability and natural exponential families with power variance functions. *Ann. Statist* **14**, 1507-1522.

Barndorff-Nielsen, O. (1978a). *Information and Exponential Families in Statistical Theory*. Chichester: Wiley.

Barndorff-Nielsen, O. (1978b). Hyperbolic distributions and distributions on hyperbolae. *Scand. J. Statist.* **5**, 151-157.

Barndorff-Nielsen, O. (1985). Properties of modified profile likelihood. In *Contributions to Probability and Statistics in Honour of Gunner Blum* (Ed.J. Lanke et al.) Lund University.

Barndorff-Nielsen, O. and Cox, D.R. (1979). Edgeworth and saddlepoint approximations with statistical applications. *J.R. Statist. Soc.* **B 41**, 279-312.

Blæsild, P. and Jensen, J.L. (1985). Saddlepoint formulas for reproductive exoponential models. *Scand. J. Statist.* **12**, 193-202.

Bleistein, N. and Handelsman, R.A. (1975). *Asymptotic Expansions of Integrals*. New York: Holt, Rinehart and Winston.

Bliss, C.I. (1935) The calculation of the dosage-mortality curve. *Ann. Appl. Biol.* **22**, 134-167.

Botelho, V.L. (1989). *The Peasantry in The Expansion of The Amazonian Frontier*. Ph.D. thesis, University of London.

Chhikara, R.S. and Folks, J.L. (1978). The inverse Gaussian distribution and its statistical application - a review (with discussion). *J.R. Statist. Soc.* **B 40**, 263-289.

Cordeiro, G.M. (1986) *Modelos Lineares Generalizados*. Universidade Estadual de Campinas, São Paulo, Brasil.

Cox, D.R. and Snell, E.J. (1981). *Applied Statistics*. London: Chapman and Hall.

Daniels, H.E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Statist.* **25**, 631-650.

Daniels, H.E. (1980). Exact saddlepoint approximations. *Biometrika* **67**, 59-63.

Daniels, H.E. (1987). Tail probability approximations. *Int. Statist. Rev.* **55**, 37-48.

Dobson, A.J. (1983). *An Introduction to Statistical Modelling*. London: Chapman and Hall.

Eaton, M.L., Morris, C. and Rubin, H. (1971). On extreme stable laws and some applications. *J. Appl. Prob.* **8**, 794-801.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc.* **81**, 709-721.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.* **13**, 342-368.

Garby, L., Garrow, J.S., Jørgensen, B., Lammert, O., Madsen, K., Sørensen, P. and Webster, J. (1988) Relation between energy expenditure and body composition in man: Specific energy expenditure in *vivo* of fat and fat-free tissue. *Eur. J. Clinic. Nutr.* **42**, 301-305.

Gehan, E.A. (1965) A generalized Wilcoxon test for comparing arbitrary single-censured samples. *Biometrika* **52**, 203-224.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika* **73**, 387-396.

Jensen, J.L. (1981). On the hyperboloid distribution. *Scand. J. Statist.* **8**, 193-206.

Jørgensen, B. (1982). *Statistical Properties of the Generalized Inverse Gaussian Distribution*. Lecture Notes in Statistics 9. New York: Springer-Verlag.

Jørgensen, B. (1983). Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika* **70**, 19-28.

Jørgensen, B. (1984). The delta algorithm and GLIM. *Int. Statist. Rev.* **52**, 283-300.

Jørgensen, B. (1986). Some properties of exponential dispersion models. *Scand. J. Statist.* **13**, 187-198.

Jørgensen, B. (1987a). Exponential dispersion models (with discussion). *J.R. Statist. Soc.* **B 49**, 127-162.

Jørgensen, B. (1987b). Small-dispersion asymptotics. *Rev. Bras. Prob. Estat.* **1**, 59-90.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley: New York.

Kemp, C.D. and Kemp, A.W. (1965). Some properties of the Hermite distribution. *Biometrika* **52**, 381-394.

Küchler, U. (1982). Exponential families of Markov processes - Part I. General results. *Math. Operationsforsch. Statist., Ser. Statistics* **13**, 57-69.

Lugannani, R. and Rice, S.O. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. *Adv. Appl. Prob.* **12**, 475-490

Mardia, K.V. (1972). *Statistics of Directional Data.* London: Academic Press.

McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11**, 59-67.

McCullagh, P. (1985). On the asymptotic distributions of Pearson's statistic in linear exponential-family models. *Int. Statist. Rev.* **53**, 61-67.

McCullagh, P. and Nelder, J.A. (1983). *Generalized Linear Models.* London: Champman and Hall.

Morris, C.N. (1981). *Models for positive data with good convolution properties.* Memo no. 8949, Rand Corporation, Santa Monica, California.

Morris, C.N. (1982). Natural exponential families with quadratic variance functions. *Ann. Statist.* **10**, 65-80.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *J.R. Statist. Soc.* **A 135**, 370-384.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Appl. Statist.* **29**, 15-24.

Pregibon, D. (1981). Logistic regression diagnostics. *Ann. Statist.* **9**, 705-724.

Ryan, B.F., Joiner, B.L. and Ryan, T.A. (1985). Minitab Handbook (2nd ed.) Boston: Duxburg Press.

Sinclair, C.D. (1982). GLIM for preference. In *GLIM 82: Proceedings of the International Conference on Generalized Linear Models* (ed. R. Gilchrist.). Lecture Notes in Statistics 14, New York: Springer-Verlag.

Siegel, A.F. (1985). Modelling data containing exact zeroes using zero degrees of freedom. *J.R. Statist. Soc.* **B 47**, 267-271.

Sweeting, T.J. (1981). Scale parameters: A Bayesian treatment. *J.R. Statist. Soc.* **B 43**, 333-338.

Sweeting, T.J. (1984). Approximate inference in location-scale regression models. *J. Amer. Statist. Assoc.* **79**, 847-852.

Tweedie, M.C.K. (1947). Functions of a statistical variate with given means, with special reference to Laplacian distributions. *Proc. Cambridge Phil. Soc.* **49**, 41-49.

Tweedie, M.C.K. (1957). Statistical properties of Inverse Gaussian distributions, I. *Ann. Math. Statist.* **28**, 362-372.

Tweedie, M.C.K. (1984). An index which distinghishes between some important exponential families. In *Statistics: Applications and New Directions.*

Proceedings of the Indian Statistical Institute Golden Jubilee International Conference. (Eds. J.K. Ghosh and J. Roy), pp. 579-604. Calcutta: Indian Statistical Institute.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* **61**, 439-447.

West, M. (1985). Generalized linear Models: scale parameters, outlier accomodation and prior distributions. *Bayesian Statistics* **2**, 531-558. Amsterdam: North-Holland.

Whitehead, J. (1980). Fitting Cox's regression model to survival data using GLIM. *Appl. Statist.* **29**, 268-275.

Impresso na Gráfica do
impa
pelo Sistema Xerox/10[illegible]