Chapter 1

Introduction

In many situations, data are only available in a grouped form. Typical continuous variables such as income, age, test scores and many more are for various reasons classified into a few class intervals. The implication is that the usual statistical techniques employed for continuous variables can no longer be applied in the usual sense. Often when researchers are confronted with grouped data, the underlying continuous nature of the variable is ignored and the data do not comply to the requirements of the statistical tests applied.

The maximum likelihood (ML) estimation procedure of *Matthews and Crowther (1995)* will be utilized to fit a continuous distribution to a grouped data set. This grouped data set may be a single frequency distribution or various frequency distributions that arise from a cross classification of several factors in a multifactor design. It will also be shown how to fit a bivariate normal distribution to a two-way contingency table where the two underlying continuous variables are jointly normally distributed.

This thesis is organized in three different parts, each playing a vital role in the explanation of analysing grouped data with the ML estimation of *Matthews and Crowther*. All the examples, applications and simulations are done with the SAS procedure IML, listed in the Appendix.

Part I

The ML estimation procedure of *Matthews and Crowther* is formulated. This procedure plays an integral role and is implemented in all three parts of the thesis. In Part I the exponential distribution is fitted to a grouped data set to explain the technique. Two different formulations of the constraints are employed in the ML estimation procedure and provide identical results. The justification of the method is further motivated by a simulation study. Similar to the exponential distribution, the estimation of the normal distribution is also explained in detail. Part I is summarized in Chapter 5 where a general method is outlined to fit continuous distributions to a grouped data set. Distributions such as the Weibull, the log-logistic and the Pareto distributions can be fitted very effectively by formulating the vector of constraints in terms of a linear model.

Part II

In Part II it is explained how to model a grouped response variable in a multifactor design. This multifactor design arise from a cross classification of the various factors or independent variables to be analysed. The cross classification of the factors results in a total of T cells, each containing a frequency distribution. Distribution fitting is done simultaneously to each of the T cells of the multifactor design. Distribution fitting is also done under the additional constraints that the parameters of the underlying continuous distributions satisfy a certain structure or design. The effect of the factors on the grouped response variable may be evaluated from this fitted design. Applications of a single-factor and a two-factor model are considered to demonstrate the versatility of the technique.

Part III

A two-way contingency table where the two variables have an underlying bivariate normal distribution is considered. The estimation of the bivariate normal distribution reveals the complete underlying continuous structure between the two variables. The ML estimate of the correlation coefficient ρ is used to great effect to describe the relationship between the two variables. Apart from an application a simulation study is also provided to support the method proposed.

Part I

Fitting distributions to grouped data

Chapter 2

The ML estimation procedure

In this chapter the ML estimation procedure of *Matthews and Crowther (1995)* is presented. This procedure is employed to find the ML estimates in the statistical analysis of grouped data. The formulation and explanation of the ML estimation procedure described in this chapter will be used throughout the thesis.

2.1 Formulation

Consider a total of n observations tabulated in a frequency distribution with k classes.

Table 2.1: General formulation of a frequency distribution.

Class Interval	Frequency
$(-\infty, x_1)$	f_1
$[x_1, x_2)$	f_2
$\vdots$	$\vdots$
$[x_{k-2}, x_{k-1})$	f_{k-1}
$[x_{k-1}, \infty)$	f_k

The observations in Table 2.1 originate from a continuous distribution and information concerning the distribution is now only available in grouped form. In Table 2.1 the first and last intervals of the frequency distribution may be open ended class intervals.

Denote the vector of the first $(k-1)$ frequencies in Table 2.1 by

$$\mathbf{f} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{k-1} \end{pmatrix} \tag{2.1}$$

with corresponding vector of upper class boundaries

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{k-1} \end{pmatrix} . \tag{2.2}$$

It is assumed that the vector $\mathbf{f}$ is a random vector with some discrete distribution such as Poisson, multinomial or product multinomial. Assume multinomial sampling and define

$$\mathbf{p}_0 = \frac{1}{n}\mathbf{f} \tag{2.3}$$

as the vector of relative frequencies. Let $\boldsymbol{\pi}_0$ denote the vector of probabilities, where the i-th element of $\boldsymbol{\pi}_0$ is the probability that an observation falls in the i-th class interval. Hence, the expected value of $\mathbf{p}_0$ is

$$\mathrm{E}(\mathbf{p}_0) = \boldsymbol{\pi}_0 \tag{2.4}$$

with covariance matrix

$$\mathrm{Cov}(\mathbf{p}_0) = \frac{1}{n}\left(\mathrm{diag}\left[\boldsymbol{\pi}_0\right] - \boldsymbol{\pi}_0\boldsymbol{\pi}_0'\right) = \mathbf{V}_0 \tag{2.5}$$

where $\mathrm{diag}\left[\boldsymbol{\pi}_0\right]$ is a diagonal matrix with the elements of $\boldsymbol{\pi}_0$ on the diagonal.

The vector of cumulative relative frequencies is denoted by

$$\mathbf{p} = \mathbf{C}\mathbf{p}_0 \tag{2.6}$$

where $\mathbf{C}$ is a triangular matrix such that

$$\mathbf{C} : (k-1) \times (k-1) = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix} . \tag{2.7}$$

The expected value of $\mathbf{p}$ is

$$\begin{aligned} \mathrm{E}(\mathbf{p}) &= \mathbf{C}\boldsymbol{\pi}_0 \\ &= \boldsymbol{\pi} \end{aligned} \tag{2.8}$$

with covariance matrix

$$\begin{aligned} \mathrm{Cov}(\mathbf{p}) &= \mathbf{C}\mathbf{V}_0\mathbf{C}' \\ &= \mathbf{C}\left\{\frac{1}{n}(\mathrm{diag}\left[\boldsymbol{\pi}_0\right] - \boldsymbol{\pi}_0\boldsymbol{\pi}_0')\right\}\mathbf{C}' \\ &= \frac{1}{n}\left\{\mathbf{C}\,\mathrm{diag}\left[\mathbf{C}^{-1}\boldsymbol{\pi}\right]\mathbf{C}' - \boldsymbol{\pi}\boldsymbol{\pi}'\right\} \\ &= \mathbf{V} . \end{aligned} \tag{2.9}$$

2.2 Estimation

The frequency vector $\mathbf{f}$ is distributed according to a multinomial distribution and consequently belongs to the exponential class. Since the vector of cumulative relative frequencies is a one-to-one transformation of $\mathbf{f}$, the random vector $\mathbf{p}$ may be implemented in the ML estimation procedure of *Matthews and Crowther (1995)* presented in Proposition 1. Utilizing the ML estimation, it is possible to find the ML estimate of $\boldsymbol{\pi}$, under the restriction that $\boldsymbol{\pi}$ satisfies the constraints defined in the ML estimation procedure.

The basic foundation of this research are given in the following two propositions. The proofs are given in Matthews and Crowther (1995).

Proposition 1 *(ML estimation procedure)*
Consider a random vector of cumulative relative frequencies $\mathbf{p}$*, which may be considered as a non-singular (one-to-one) transformation of the canonical vector of observations, belonging to the exponential family, with*

$$\mathrm{E}(\mathbf{p}) = \boldsymbol{\pi} \quad \textit{and} \ \mathrm{Cov}(\mathbf{p}) = \mathbf{V} \ .$$

The observed $\mathbf{p}$ *is the unrestricted ML estimate of* $\boldsymbol{\pi}$ *and the covariance matrix* $\mathbf{V}$ *may be a function of* $\boldsymbol{\pi}$*. Let* $\mathbf{g}(\boldsymbol{\pi})$ *be a continuous vector valued function of* $\boldsymbol{\pi}$*, for which the first order partial derivatives,*

$$\mathbf{G}_{\pi} = \frac{\partial \mathbf{g}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} \tag{2.10}$$

with respect to $\boldsymbol{\pi}$ *exist. The ML estimate of* $\boldsymbol{\pi}$*, subject to the constraints* $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$ *is obtained iteratively from*

$$\widehat{\boldsymbol{\pi}} = \mathbf{p} - (\mathbf{G}_{\pi}\mathbf{V})' \left(\mathbf{G}_{p}\mathbf{V}\mathbf{G}_{\pi}'\right)^{*} \mathbf{g}(\mathbf{p}) \tag{2.11}$$

where $\mathbf{G}_{p} = \left.\dfrac{\partial \mathbf{g}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}\right|_{\pi=p}$ *and* $\left(\mathbf{G}_{p}\mathbf{V}\mathbf{G}_{\pi}'\right)^{*}$ *is a generalized inverse of* $\left(\mathbf{G}_{p}\mathbf{V}\mathbf{G}_{\pi}'\right)$*.*

The iterative procedure implies a double iteration over $\mathbf{p}$ and $\boldsymbol{\pi}$. The procedure starts with the unrestricted ML estimate of $\boldsymbol{\pi}$, as the starting value for both $\mathbf{p}$ and $\boldsymbol{\pi}$. Convergence is first obtained over $\mathbf{p}$ using (2.11). The converged value of $\mathbf{p}$ is then used as the next value of $\boldsymbol{\pi}$, with convergence over $\mathbf{p}$ starting again at the observed $\mathbf{p}$. In this procedure $\mathbf{V}$ is recalculated for each new value of $\boldsymbol{\pi}$ in the iterative procedure. Convergence over $\boldsymbol{\pi}$ ultimately leads to $\widehat{\boldsymbol{\pi}}$, the restricted ML estimate of $\boldsymbol{\pi}$.

Proposition 2 *The asymptotic covariance matrix of* $\widehat{\boldsymbol{\pi}}$*, under* $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$*, is*

$$\mathrm{Cov}\left(\widehat{\boldsymbol{\pi}}\right) \approxeq \mathbf{V} - (\mathbf{G}_{\pi}\mathbf{V})' \left(\mathbf{G}_{\pi}\mathbf{V}\mathbf{G}_{\pi}'\right)^{*} (\mathbf{G}_{\pi}\mathbf{V}) \tag{2.12}$$

which is estimated by replacing $\boldsymbol{\pi}$ *by* $\widehat{\boldsymbol{\pi}}$*.*

In *Matthews and Crowther (1995)* it is assumed that the restrictions are linearly independent, but in *Matthews and Crowther (1998)*, it is shown that if the restrictions are linearly dependent, it leads to the generalized inverse, $\left(\mathbf{G}_{\pi}\mathbf{V}\mathbf{G}_{\pi}'\right)^{*}$, to be introduced in (2.11) and (2.12).

The objective is now to find the ML estimate of $\boldsymbol{\pi}$, under the constraints that the cumulative relative frequencies $\boldsymbol{\pi}$, equal the cumulative distribution curve, $F(\mathbf{x})$ at the upper class boundaries $\mathbf{x}$. This

implies that the ML estimate of $\boldsymbol{\pi}$ is to be obtained under the restriction

$$F(\mathbf{x}) = \boldsymbol{\pi} \tag{2.13}$$

which means that the vector of constraints in (2.11) may be formulated as

$$\mathbf{g}(\boldsymbol{\pi}) = F(\mathbf{x}) - \boldsymbol{\pi} = \mathbf{0} \ . \tag{2.14}$$

The set of constraints in Proposition 1 is essentially not unique and may be dependent. Any function say $\mathbf{g}_1(\boldsymbol{\pi})$, that implies the same constraints on $\boldsymbol{\pi}$ as $\mathbf{g}(\boldsymbol{\pi})$, may be used and will provide the same results. The objective now is to choose $\mathbf{g}(\boldsymbol{\pi})$ in such a way to simplify the calculation of derivatives and to streamline the estimation process. In some instances it is possible to find the ML estimate of $\boldsymbol{\pi}$ under constraints, by making use of traditional methods, but the procedure suggested in Proposition 1 provides an elegant and straightforward method for obtaining the ML estimates.

2.3 Goodness of fit

In order to test the deviation of the observed probabilities $\mathbf{p}$ from the restricted ML estimates $\widehat{\boldsymbol{\pi}}$, imposed by the constraints $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$, it is convenient to formulate and test the null hypothesis

$$\mathbf{H}_0 : \mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$$

by some goodness of fit statistic like the Pearson χ^2-statistic

$$\boldsymbol{\chi}^2 = \sum_{i=1}^{k} \frac{(p_i - \widehat{\pi}_i)^2}{\widehat{\pi}_i} \tag{2.15}$$

or the Wald statistic

$$\mathbf{W} = \mathbf{g}(\mathbf{p})'(\mathbf{G}_p \mathbf{V} \mathbf{G}_p')^* \mathbf{g}(\mathbf{p}) \ . \tag{2.16}$$

Both the Pearson and the Wald statistic have a χ^2-distribution with r degrees of freedom, where r is equal to the number of linear independent constraints in $\mathbf{g}(\boldsymbol{\pi})$.

Another useful measure, is the measure of discrepancy

$$\mathbf{D} = \mathbf{W}/n \tag{2.17}$$

which will provide more conservative results for large sample sizes. As a rule of thumb the observed and expected frequencies are considered to not deviate significantly from each other if the discrepancy is less than 0.05.

Chapter 3

The exponential distribution

To illustrate the underlying methodology of fitting a distribution via the ML estimation process described in Proposition 1, it will be shown how to fit an exponential distribution to the frequency data in Table 2.1.

The probability density function (pdf) of an exponential random variable with expected value μ is given by

$$f(x;\mu) = \frac{1}{\mu} e^{-x/\mu} \tag{3.1}$$

and the cumulative distribution function (cdf) is given by

$$F(x;\mu) = 1 - e^{-x/\mu} \ . \tag{3.2}$$

To fit an exponential distribution it is required (see 2.13) that

$$\mathbf{1} - \exp(-\theta \mathbf{x}) = \boldsymbol{\pi} \tag{3.3}$$

where $\mathbf{1} : (k-1) \times 1$ is a vector of ones, $\mathbf{x}$ is the vector of upper class boundaries and $\theta = \mu^{-1}$.

From this requirement (3.3) two alternative ways of performing the estimation procedure are described. In Sections 3.1 and 3.2 it will be shown that although the specifications of the two sets of constraints, $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$, seem completely different, the final results obtained are identical.

3.1 Direct set of constraints

A direct set of constraints in (2.11) follows from (3.3) with

$$\mathbf{g}(\boldsymbol{\pi}) = \{\mathbf{1} - \exp(-\theta\mathbf{x})\} - \boldsymbol{\pi} \ . \tag{3.4}$$

The parameter θ is expressed in terms of the cumulative probabilities in (3.3) and hence

$$\theta = -\frac{\mathbf{x}' \ln(\mathbf{1} - \boldsymbol{\pi})}{\mathbf{x}'\mathbf{x}} \ . \tag{3.5}$$

The chain rule for matrix differentiation is employed in (3.6) to obtain the following matrix of partial derivatives

$$\begin{aligned}
\mathbf{G}_{\pi} &= \frac{\partial \mathbf{g}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} \\
&= \frac{\partial \left(\{\mathbf{1} - \exp(-\theta\mathbf{x})\} - \boldsymbol{\pi}\right)}{\partial \boldsymbol{\pi}} \\
&= -\frac{\partial \exp(-\theta\mathbf{x})}{\partial \boldsymbol{\pi}} - \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\pi}} \\
&= -\frac{\partial \exp(-\theta\mathbf{x})}{\partial \theta} \cdot \frac{\partial \theta}{\partial \boldsymbol{\pi}} - \mathbf{I} \qquad (3.6) \\
&= -\frac{\partial \begin{pmatrix} \exp(-\theta x_1) \\ \exp(-\theta x_2) \\ \vdots \\ \exp(-\theta x_{k-1}) \end{pmatrix}}{\partial \theta} \cdot \left(-\frac{\mathbf{x}'}{\mathbf{x}'\mathbf{x}}\right) \cdot \frac{\partial \begin{pmatrix} \ln(1-\pi_1) \\ \ln(1-\pi_2) \\ \vdots \\ \ln(1-\pi_{k-1}) \end{pmatrix}}{\partial \boldsymbol{\pi}} - \mathbf{I} \\
&= \begin{pmatrix} \exp(-\theta x_1) \cdot x_1 \\ \exp(-\theta x_2) \cdot x_2 \\ \vdots \\ \exp(-\theta x_{k-1}) \cdot x_{k-1} \end{pmatrix} \cdot \left(-\frac{\mathbf{x}'}{\mathbf{x}'\mathbf{x}}\right) \cdot \operatorname{diag} \begin{bmatrix} -(1-\pi_1)^{-1} \\ -(1-\pi_2)^{-1} \\ \vdots \\ -(1-\pi_{k-1})^{-1} \end{bmatrix} - \mathbf{I} \\
&= -\left(\operatorname{diag}\left[\exp(-\theta\mathbf{x})\right]\right) \cdot \mathbf{P}_x \cdot \mathbf{D}_{\pi} - \mathbf{I} \qquad (3.7)
\end{aligned}$$

where

$$\mathbf{P}_x = \mathbf{x}\left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}' \tag{3.8}$$

is the projection matrix of $\mathbf{x}$ and

$$\begin{aligned}\mathbf{D}_{\pi} &= \frac{\partial \ln(\mathbf{1}-\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} \\ &= -\left(\operatorname{diag}\left[\mathbf{1}-\boldsymbol{\pi}\right]\right)^{-1} .\end{aligned} \tag{3.9}$$

The estimation procedure in Proposition 1 utilizes a double iteration over $\boldsymbol{\pi}$ and $\mathbf{p}$ starting with the observed vector of cumulative relative frequencies as the initial values for both convergencies over $\boldsymbol{\pi}$ and $\mathbf{p}$. The iterative procedure may be summarised as follows:

$\mathbf{p}^{\dagger} =$ observed cumulative relative frequencies

$\mathbf{p} = \mathbf{p}^{\dagger}$

$\mathbf{P}_x = \mathbf{x}\left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}'$

DO OVER $\boldsymbol{\pi}$

$\boldsymbol{\pi} = \mathbf{p}$

$\mathbf{V} = \frac{1}{n}\left\{\mathbf{C}\operatorname{diag}\left[\mathbf{C}^{-1}\boldsymbol{\pi}\right]\mathbf{C}' - \boldsymbol{\pi}\boldsymbol{\pi}'\right\}$

$\theta_{\pi} = -\dfrac{\mathbf{x}'\ln(\mathbf{1}-\boldsymbol{\pi})}{\mathbf{x}'\mathbf{x}}$

$\mathbf{D}_{\pi} = -\left(\operatorname{diag}\left[\mathbf{1}-\boldsymbol{\pi}\right]\right)^{-1}$

$\mathbf{G}_{\pi} = -\left(\operatorname{diag}\left[\exp\left(-\theta_{\pi}\mathbf{x}\right)\right]\right)\cdot\mathbf{P}_x\cdot\mathbf{D}_{\pi} - \mathbf{I}$

$\mathbf{p} = \mathbf{p}^{\dagger}$

DO OVER $\mathbf{p}$

$\theta_p = -\dfrac{\mathbf{x}'\ln(\mathbf{1}-\mathbf{p})}{\mathbf{x}'\mathbf{x}}$

$\mathbf{D}_{\pi} = -\left(\operatorname{diag}\left[\mathbf{1}-\boldsymbol{\pi}\right]\right)^{-1}$

$\mathbf{G}_p = -\left(\operatorname{diag}\left[\exp\left(-\theta_p\mathbf{x}\right)\right]\right)\cdot\mathbf{P}_x\cdot\mathbf{D}_p - \mathbf{I}$

$\mathbf{g}(\mathbf{p}) = \left\{\mathbf{1}-\exp\left(-\theta_p\mathbf{x}\right)\right\} - \mathbf{p}$

$\mathbf{p} = \mathbf{p} - \left(\mathbf{G}_{\pi}\mathbf{V}\right)'\left(\mathbf{G}_{\pi}\mathbf{V}\mathbf{G}_p\right)^{*}\mathbf{g}\left(\mathbf{p}\right)$

END

END

From the above it follows that convergence is first obtained over $\mathbf{p}$ where the parameter θ_p, the vector of constraints $\mathbf{g}(\mathbf{p})$ and the matrix of partial derivatives $\mathbf{G}_p$ are all functions of $\mathbf{p}$. Convergence over $\mathbf{p}$ leads to the next value of $\boldsymbol{\pi}$ with convergence over $\mathbf{p}$ starting again at the observed vector of cumulative relative frequencies namely $\mathbf{p}^{\dagger}$. The values of $\mathbf{V}$, θ_{π} and $\mathbf{G}_{\pi}$ are recalculated for every value of $\boldsymbol{\pi}$ when iterating over $\boldsymbol{\pi}$. Convergence over $\boldsymbol{\pi}$ ultimately leads to $\widehat{\boldsymbol{\pi}}$, the restricted ML estimate of $\boldsymbol{\pi}$ under $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$ with corresponding ML estimator

$$\widehat{\theta} = -\frac{\mathbf{x}' \ln(\mathbf{1} - \widehat{\boldsymbol{\pi}})}{\mathbf{x}'\mathbf{x}} \tag{3.10}$$

and hence the ML estimator for the exponential distribution

$$\widehat{\mu} = \frac{1}{\widehat{\theta}} = -\left(\frac{\mathbf{x}' \ln(\mathbf{1} - \widehat{\boldsymbol{\pi}})}{\mathbf{x}'\mathbf{x}}\right)^{-1} \tag{3.11}$$

follows. The iterative process is illustrated in Example 3.1.

Example 3.1

Consider $n = 100$ observations simulated from an exponential distribution with expected value $\mu = \theta^{-1} = 50$. The grouped data set is shown in Table 3.1.

Table 3.1: Simulated data set from an exponential distribution.

Class interval	Frequency
$[0, 12.5)$	17
$[12.5, 25)$	14
$[25, 50)$	31
$[50, 100)$	26
$[100, \infty)$	12

Table 3.2 shows the various values of $\boldsymbol{\pi}$ and $\mathbf{p}$ in the double iteration process, with corresponding values for $\mu = \theta^{-1}$. The results in Table 3.2 can be calculated directly, or can be obtained using the SAS program *EXP1.SAS* listed in Appendix A.1.

Table 3.2: Double iteration process.

Iteration over $\boldsymbol{\pi}$		Iteration over $\mathbf{p}$			
		$j=1$	$j=2$	$j=3$	$j=4$
$i=1$	$\begin{pmatrix}0.1700\\0.3100\\0.6200\\0.8800\end{pmatrix}$ $\mu_\pi = 48.83$	$\begin{pmatrix}0.1700\\0.3100\\0.6200\\0.8800\end{pmatrix}$ $\mu_p = 48.83$	$\begin{pmatrix}0.2373\\0.4184\\0.6620\\0.8862\end{pmatrix}$ $\mu_p = 46.03$	$\begin{pmatrix}0.2380\\0.4194\\0.6629\\0.8863\end{pmatrix}$ $\mu_p = 45.99$	$\begin{pmatrix}0.2380\\0.4194\\0.6629\\0.8863\end{pmatrix}$ $\mu_p = 45.99$
$i=2$	$\begin{pmatrix}0.2380\\0.4194\\0.6629\\0.8863\end{pmatrix}$ $\mu_\pi = 45.99$	$\begin{pmatrix}0.1700\\0.3100\\0.6200\\0.8800\end{pmatrix}$ $\mu_p = 48.83$	$\begin{pmatrix}0.2137\\0.3820\\0.6186\\0.8563\end{pmatrix}$ $\mu_p = 51.63$	$\begin{pmatrix}0.2147\\0.3833\\0.6197\\0.8553\end{pmatrix}$ $\mu_p = 51.72$	$\begin{pmatrix}0.2147\\0.3833\\0.6197\\0.8553\end{pmatrix}$ $\mu_p = 51.72$
$i=3$	$\begin{pmatrix}0.2147\\0.3833\\0.6197\\0.8553\end{pmatrix}$ $\mu_\pi = 51.72$	$\begin{pmatrix}0.1700\\0.3100\\0.6200\\0.8800\end{pmatrix}$ $\mu_p = 48.83$	$\begin{pmatrix}0.2143\\0.3829\\0.6196\\0.8570\end{pmatrix}$ $\mu_p = 51.49$	$\begin{pmatrix}0.2152\\0.3841\\0.6207\\0.8561\end{pmatrix}$ $\mu_p = 51.57$	$\begin{pmatrix}0.2152\\0.3841\\0.6207\\0.8561\end{pmatrix}$ $\mu_p = 51.57$
$i=4$	$\begin{pmatrix}0.2152\\0.3841\\0.6207\\0.8561\end{pmatrix}$ $\mu_\pi = 51.57$	$\begin{pmatrix}0.1700\\0.3100\\0.6200\\0.8800\end{pmatrix}$ $\mu_p = 48.83$	$\begin{pmatrix}0.2143\\0.3828\\0.6196\\0.8570\end{pmatrix}$ $\mu_p = 51.49$	$\begin{pmatrix}0.2152\\0.3841\\0.6207\\0.8561\end{pmatrix}$ $\mu_p = 51.58$	$\begin{pmatrix}0.2152\\0.3841\\0.6207\\0.8561\end{pmatrix}$ $\mu_p = 51.58$
$i=5$	$\begin{pmatrix}0.2152\\0.3841\\0.6207\\0.8561\end{pmatrix}$ $\mu_\pi = 51.58$	$\begin{pmatrix}0.1700\\0.3100\\0.6200\\0.8800\end{pmatrix}$ $\mu_p = 48.83$	$\begin{pmatrix}0.2143\\0.3828\\0.6196\\0.8570\end{pmatrix}$ $\mu_p = 51.49$	$\begin{pmatrix}0.2152\\0.3841\\0.6207\\0.8561\end{pmatrix}$ $\mu_p = 51.58$	$\begin{pmatrix}0.2152\\0.3841\\0.6207\\0.8561\end{pmatrix}$ $\mu_p = 51.58$

The procedure starts with the unrestricted ML estimate of $\boldsymbol{\pi}$

$$\boldsymbol{\pi} = \mathbf{p} = \begin{pmatrix} 0.1700 \\ 0.3100 \\ 0.6200 \\ 0.8800 \end{pmatrix}$$

(the observed vector of cumulative relative frequencies) and after convergence the restricted ML estimate of $\boldsymbol{\pi}$

$$\widehat{\boldsymbol{\pi}} = \begin{pmatrix} 0.2152 \\ 0.3841 \\ 0.6207 \\ 0.8561 \end{pmatrix}$$

is obtained. The elements of $\widehat{\boldsymbol{\pi}}$ follow a cumulative exponential curve at the upper class boundaries and hence the ML estimate

$$\widehat{\mu} = -\left(\frac{\mathbf{x}' \ln(\mathbf{1} - \widehat{\boldsymbol{\pi}})}{\mathbf{x}'\mathbf{x}}\right)^{-1} = 51.58$$

follows. The estimated exponential distribution is therefore

$$f(x) = \frac{1}{51.58} \exp(-\frac{x}{51.58})$$

and is shown in Figure 3.1, together with the observed frequency distribution (blue line) and estimated frequency distribution (red line).

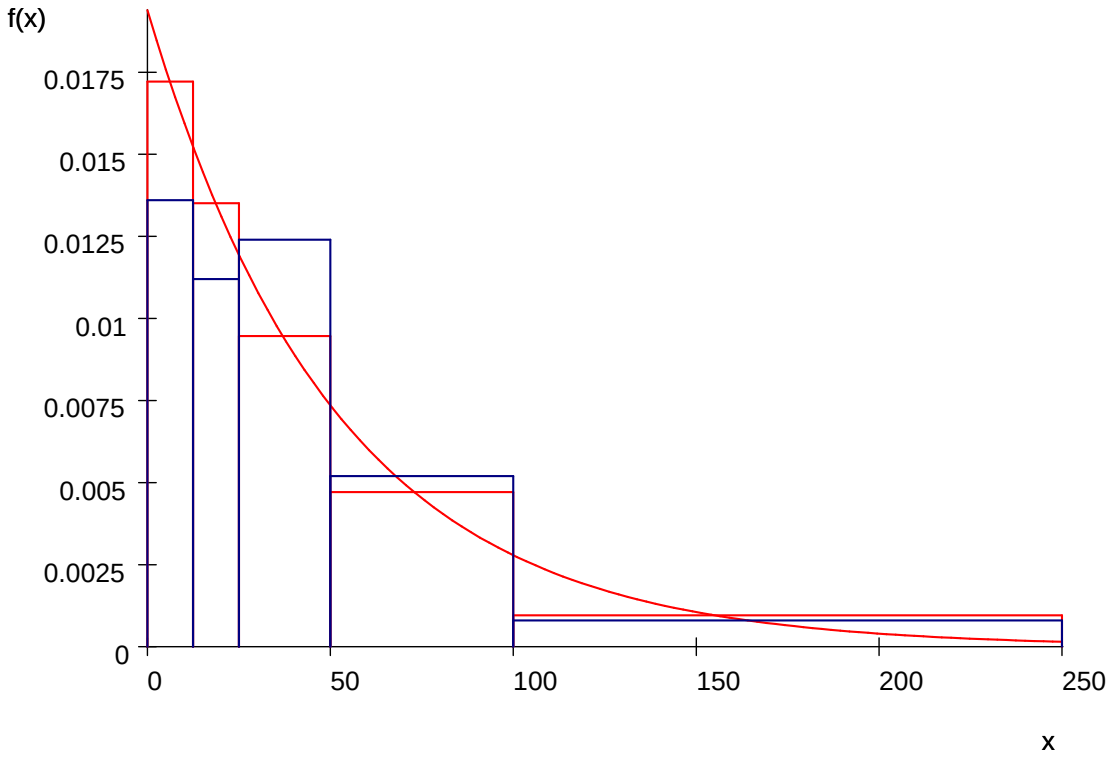

Figure 3.1: The estimated exponential distribution with the observed and estimated frequency distribution.

3.2 Constraints in terms of a linear model

An alternative formulation of the vector of constraints may be developed. The linear model

$$\ln(\mathbf{1} - \boldsymbol{\pi}) = -\theta\mathbf{x} \tag{3.12}$$

follows from the requirement (3.3) implying that $\ln(\mathbf{1} - \boldsymbol{\pi})$ is a scalar multiple of the upper class boundaries, $\mathbf{x}$. Or equivalently, $\ln(\mathbf{1} - \boldsymbol{\pi})$ is in the vector space generated by $\mathbf{x}$.

Since $\mathbf{Q}_x = \mathbf{I} - \mathbf{x}\left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}'$ is the projection matrix of the vector space orthogonal to $\mathbf{x}$, the vector of constraints, $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$, may now be expressed in terms of a new $\mathbf{g}(\boldsymbol{\pi})$ namely

$$\mathbf{g}(\boldsymbol{\pi}) = \mathbf{Q}_x \ln(\mathbf{1} - \boldsymbol{\pi}) \ . \tag{3.13}$$

The rationale behind the constraints (3.13) is that $\ln(\mathbf{1}-\boldsymbol{\pi})$ is an element of the vector space of $\mathbf{x}$ if and only if $\ln(\mathbf{1}-\boldsymbol{\pi})$ is orthogonal to the error space of $\mathbf{x}$ (i.e. vector space orthogonal to $\mathbf{x}$) in which case $\mathbf{Q}_x \ln(\mathbf{1}-\boldsymbol{\pi}) = \mathbf{0}$. The vector of constraints (3.13) consists out of $(k-2)$ linear independent functions, since

$$\begin{aligned}\operatorname{rank}(\mathbf{Q}_x) &= \operatorname{rank}(\mathbf{I}) - \operatorname{rank}\left(\mathbf{x}\left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}'\right) \\ &= (k-1) - \operatorname{rank}(\mathbf{x}) \\ &= (k-1) - 1\end{aligned}$$

The matrix of partial derivatives is now much simpler than the previous formulation (3.7) with

$$\begin{aligned}\mathbf{G}_\pi &= \frac{\partial \mathbf{g}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} \\ &= \frac{\partial}{\partial \boldsymbol{\pi}}\left\{\mathbf{Q}_x \ln(\mathbf{1}-\boldsymbol{\pi})\right\} \qquad (3.14) \\ &= \mathbf{Q}_x \mathbf{D}_\pi \qquad (3.15)\end{aligned}$$

where $\mathbf{D}_\pi = -\left(\operatorname{diag}\left[\mathbf{1}-\boldsymbol{\pi}\right]\right)^{-1}$ (previously derived in (3.9)).

The restricted ML estimate of $\boldsymbol{\pi}$ namely $\widehat{\boldsymbol{\pi}}$ is obtained after convergence of the iterative procedure and leads to the ML estimators

$$\widehat{\theta} = -\frac{\mathbf{x}' \ln(\mathbf{1}-\widehat{\boldsymbol{\pi}})}{\mathbf{x}'\mathbf{x}}$$

and

$$\widehat{\mu} = \frac{1}{\widehat{\theta}}\,.$$

Using the multivariate delta theorem (see *Bishop, Fienberg and Holland (1975) p.492*) the asymptotic variance of $\widehat{\theta}$ follows

$$\begin{aligned}\operatorname{Var}(\widehat{\theta}) &\approxeq \left\{\frac{\partial \theta}{\partial \boldsymbol{\pi}}\right\} \operatorname{Cov}(\widehat{\boldsymbol{\pi}}) \left\{\frac{\partial \theta}{\partial \boldsymbol{\pi}}\right\}' \\ &= \left\{\frac{\mathbf{x}'}{\mathbf{x}'\mathbf{x}}\mathbf{D}_\pi\right\} \operatorname{Cov}(\widehat{\boldsymbol{\pi}}) \left\{\frac{\mathbf{x}'}{\mathbf{x}'\mathbf{x}}\mathbf{D}_\pi\right\}' \qquad (3.16)\end{aligned}$$

where $\operatorname{Cov}(\widehat{\boldsymbol{\pi}})$ is given in Proposition 2 (2.12).

Applying the multivariate delta theorem again it follows that

$$\begin{aligned} \text{Var}(\widehat{\mu}) &\approxeq \left\{\frac{\partial \mu}{\partial \theta}\right\}^2 \text{Var}(\widehat{\theta}) \\ &= \frac{1}{\theta^4}\,\text{Var}(\widehat{\theta}) \end{aligned} \tag{3.17}$$

and hence

$$\widehat{\mu} \approxeq \text{N}\left(\mu, \frac{1}{\theta^4}\,\text{Var}(\widehat{\theta})\right) \ . \tag{3.18}$$

Example 3.2

In this example the estimation of the exponential distribution to the simulated frequency distribution in Table 3.1 is revisited. The vector of constraints (3.13) is now formulated in terms of a linear model. The results are exactly the same as in the previous formulation (3.4), although the intermediate iterations differ. The restricted ML estimate of $\boldsymbol{\pi}$ is tabulated in Table 3.3.

The restricted and unrestricted ML estimate of $(-\ln(\mathbf{1}-\boldsymbol{\pi}))$ are tabulated in Table 3.3.

Table 3.3: The restricted and unrestricted ML estimates.

Upper class boundaries	Unrestricted MLE		Restricted MLE	
	$\mathbf{p}$	$-\ln(\mathbf{1}-\mathbf{p})$	$\widehat{\boldsymbol{\pi}}$	$-\ln(\mathbf{1}-\widehat{\boldsymbol{\pi}})$
12.5	0.1700	0.18633	0.21522	0.24235
25	0.3100	0.37106	0.38412	0.48471
50	0.6200	0.96758	0.62069	0.96941
100	0.8800	2.1203	0.85613	1.9388

According to (3.12) the plot of $\ln(\mathbf{1}-\widehat{\boldsymbol{\pi}})$ against $\mathbf{x}$ should follow a straight line. In Figure 3.2 the unrestricted ML estimates are indicated with blue dots, while the restricted ML estimates are indicated with red circles. The circles follow the straight line

$$y = 0.019388x$$

implying that $\widehat{\theta} = 0.019388$ and consequently $\widehat{\mu} = 0.019388^{-1} = 51.578$.

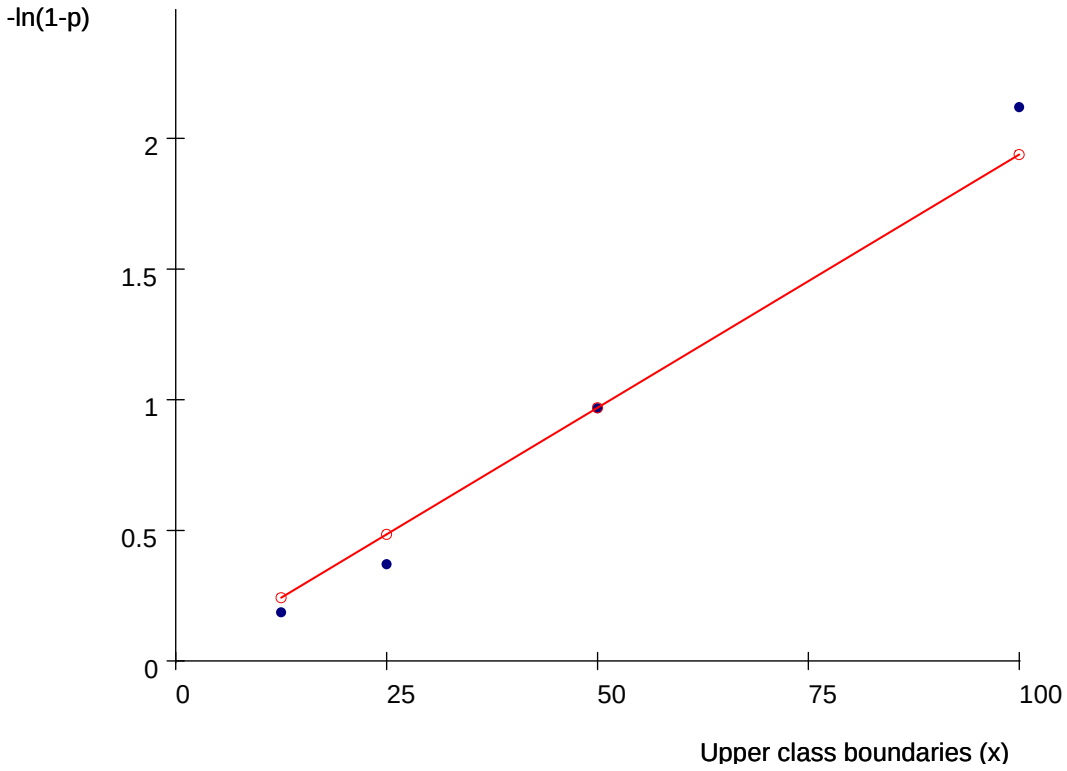

Figure 3.2: The values of $-\ln(\mathbf{1}-\widehat{\boldsymbol{\pi}})$ follow a straight line.

Other relevant statistics are summarised in Table 3.4.

Table 3.4: ML estimates and goodness of fit statistics.

MLE		Goodness of fit			
Estimate	Std. error	Statistic	Value	df	prob
$\widehat{\mu}=51.578$	$\widehat{\sigma}_{\widehat{\mu}}=5.654$	Pearson	4.376	3	0.2236
		Wald	4.295	3	0.2313

As can be expected, the Pearson and Wald statistics indicate an adequate fit.

For a 95% confidence interval for μ

$$\widehat{\mu} \pm 1.96\,(\widehat{\sigma}_{\widehat{\mu}})$$

the margin of error is $1.96\,(5.654)=11.082$, resulting in the confidence interval

$$(40.496\ ,\ 62.660)\ .$$

The SAS program *EXP2.SAS* listed in Appendix A.2 estimates the exponential distribution utilising the vector of constraints as a linear model.

3.3 Simulation study

In this study 1000 samples were simulated from an exponential distribution with expected value $\mu = 50$. Each sample consisted of 100 observations and were classified into the 5 class intervals of Table 3.1. Since the data was simulated from an exponential distribution with expected value $\mu = 50$ the true population value for $\boldsymbol{\pi}$ follows from

$$\boldsymbol{\pi} = \mathbf{1} - \exp\left(-\frac{\mathbf{x}}{50}\right) = \begin{pmatrix} 0.2212 \\ 0.3935 \\ 0.6321 \\ 0.8647 \end{pmatrix}$$

which implies that the standard error for $\widehat{\mu}$ is

$$\begin{aligned} \sigma_{\widehat{\mu}} &\approxeq \sqrt{50^4 \operatorname{Var}(\widehat{\theta})} \\ &= 5.458 \end{aligned}$$

($\operatorname{Var}(\widehat{\theta})$ derived in (3.16)). This compares well with the standard deviation of 5 of the mean of an ungrouped sample of 100 observations from this exponential distribution.

The ML estimate for μ as well as its estimated standard error were calculated for each of the 1000 generated frequency distributions. The true theoretical values as well as the descriptive statistics for the ML estimates are summarised in Table 3.5.

Table 3.5: Simulation results for the exponential distribution.

MLE	Theoretical Value	Mean	Std. deviation	P_5	Median	P_{95}
$\widehat{\mu}$	50	50.127	5.727	41.381	49.676	59.956
$\widehat{\sigma}_{\widehat{\mu}}$	5.458	5.487	0.716	4.421	5.418	6.732

From Table 3.5 it follows that the mean and median of the ML estimates are relatively close to the theoretical values. Further it is known that approximately 90% of the $\widehat{\mu}$-values should be within 1.645 standard deviations from $\mu = 50$ i.e. $1.645\sigma_{\widehat{\mu}} = 8.978$. This is in accordance with the fifth and the ninety-fifth percentile of the $\widehat{\mu}$-values tabulated in Table 3.5. The standard deviation of the $\widehat{\mu}$-values is also quite close to the standard error $\sigma_{\widehat{\mu}}$.

In Table 3.6 the percentiles of the estimated 1000 Pearson and Wald statistics are tabulated. The critical values of a χ^2-distribution with 3 degrees of freedom is also shown in Table 3.6.

Table 3.6: Percentiles of the Pearson and Wald statistic.

	Percentiles						
	P_5	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	P_{95}
Pearson	0.4370	0.6794	1.2829	2.5617	4.3586	6.5765	8.1324
Wald	0.3529	0.6299	1.2751	2.5533	4.2654	6.4586	8.0703
	Critical values of a χ^2-distribution with 3 degrees of freedom.						
	$\chi^2_{0.05}$	$\chi^2_{0.10}$	$\chi^2_{0.25}$	$\chi^2_{0.50}$	$\chi^2_{0.75}$	$\chi^2_{0.90}$	$\chi^2_{0.95}$
$\chi^2(3)$	0.3518	0.5844	1.2125	2.366	4.1083	6.2514	7.8147

From Table 3.6 it is clear that the empirical percentiles of the Pearson and Wald statistics correspond very well to the theoretical percentiles of a χ^2-distribution with 3 degrees of freedom.

The simulation study was done with the SAS program *EXPSIM.SAS* listed in Appendix A.3.

Chapter 4

The normal distribution

Analogous to the exponential distribution described in Chapter 3 the normal distribution with pdf

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \tag{4.1}$$

will be fitted to grouped data using a direct set of constraints and also constraints specified in terms of a linear model. The mean and variance of the normal distribution are μ and σ^2 respectively.

By means of standardisation, $z = \frac{x-\mu}{\sigma}$, the standard normal distribution with pdf

$$\phi(z) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}z^2\right\} \tag{4.2}$$

is obtained. The cdf of the standard normal distribution is denoted by $\mathbf{\Phi}(z)$.

To fit a normal distribution to the frequency data in Table 2.1 it is required that

$$\mathbf{\Phi}\left(\frac{\mathbf{x}-\mu\mathbf{1}}{\sigma}\right) = \boldsymbol{\pi} \tag{4.3}$$

where $\mathbf{\Phi}(\cdot)$ is the (vector valued) cdf of the standard normal distribution, $\mathbf{1}$ is the $(k-1)$ vector of ones and $\mathbf{x}$ is the vector of upper class boundaries defined in (2.2).

4.1 Direct set of constraints

To fit a normal distribution to grouped data a direct set of constraints, $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$, with

$$\mathbf{g}(\boldsymbol{\pi}) = \boldsymbol{\Phi}(\mathbf{z}) - \boldsymbol{\pi} \tag{4.4}$$

follows from (4.3). The vector of standardised upper class boundaries in (4.4) is a function of the parameters to be estimated namely

$$\begin{aligned} \mathbf{z} &= \left(\frac{\mathbf{x} - \mu\mathbf{1}}{\sigma}\right) \\ &= \begin{pmatrix} \mathbf{x} & -\mathbf{1} \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma} \\ \frac{\mu}{\sigma} \end{pmatrix} \\ &= \mathbf{X}\boldsymbol{\alpha} \end{aligned} \tag{4.5}$$

with

$$\mathbf{X} = \begin{pmatrix} \mathbf{x} & -\mathbf{1} \end{pmatrix} \tag{4.6}$$

and

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma} \\ \frac{\mu}{\sigma} \end{pmatrix} \tag{4.7}$$

the vector of so-called natural parameters.

Under normality (see 4.3)

$$\begin{aligned} \boldsymbol{\Phi}^{-1}(\boldsymbol{\pi}) &= \left(\frac{\mathbf{x} - \mu\mathbf{1}}{\sigma}\right) \\ &= \mathbf{X}\boldsymbol{\alpha} \end{aligned} \tag{4.8}$$

which leads to the expression

$$\boldsymbol{\alpha} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}(\boldsymbol{\pi}) \ . \tag{4.9}$$

The parameters of the normal distribution are now specified in terms of the cumulative relative frequencies $\boldsymbol{\pi}$. Hence, from (4.5) and (4.9) the standardised upper class boundaries may be expressed as

$$\mathbf{z} = \mathbf{P}_X\boldsymbol{\Phi}^{-1}(\boldsymbol{\pi}) \tag{4.10}$$

where

$$\mathbf{P}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{4.11}$$

is the projection matrix generated by the columns of $\mathbf{X}$. This implies that, under normality the vector $\mathbf{z}$ is the projection of $\boldsymbol{\Phi}^{-1}(\boldsymbol{\pi})$ on the vector space of $\mathbf{X}$.

From the chain rule for matrix differentiation, employed in (4.12), it follows that the matrix of partial derivatives is

$$\begin{aligned}
\mathbf{G}_{\pi} &= \frac{\partial \mathbf{g}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} \\
&= \frac{\partial \boldsymbol{\Phi}(\mathbf{z})}{\partial \boldsymbol{\pi}} - \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\pi}} \\
&= \frac{\partial \boldsymbol{\Phi}(\mathbf{z})}{\partial \mathbf{z}} \cdot \frac{\partial \mathbf{z}}{\partial \boldsymbol{\pi}} - \mathbf{I} \qquad (4.12) \\
&= \operatorname{diag}[\boldsymbol{\phi}(\mathbf{z})] \cdot \mathbf{P}_X \cdot \mathbf{D}_{\pi} - \mathbf{I} \qquad (4.13)
\end{aligned}$$

where

$$\mathbf{D}_{\pi} = \frac{\partial \boldsymbol{\Phi}^{-1}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} \; . \qquad (4.14)$$

To solve (4.14) set $\boldsymbol{\nu} = \boldsymbol{\Phi}^{-1}(\boldsymbol{\pi})$ then $\boldsymbol{\Phi}(\boldsymbol{\nu}) = \boldsymbol{\pi}$ and hence

$$\begin{aligned}
\mathbf{D}_{\pi} &= \frac{\partial \boldsymbol{\nu}}{\partial \boldsymbol{\pi}} \\
&= \left(\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\nu}} \right)^{-1} \\
&= \left(\frac{\partial \boldsymbol{\Phi}(\boldsymbol{\nu})}{\partial \boldsymbol{\nu}} \right)^{-1} \\
&= \left(\operatorname{diag}[\boldsymbol{\phi}(\boldsymbol{\nu})] \right)^{-1} \\
&= \left(\operatorname{diag}\left[\boldsymbol{\phi}\left(\boldsymbol{\Phi}^{-1}(\boldsymbol{\pi}) \right) \right] \right)^{-1} \qquad (4.15)
\end{aligned}$$

with $\boldsymbol{\phi}(\cdot)$ the vector valued pdf of the standard normal distribution.

The vector of constraints (4.4) and the matrix of partial derivatives (4.13) may be implemented in the ML estimation procedure, where the restricted ML estimate $\widehat{\boldsymbol{\pi}}$ is obtained iteratively in a double iterative procedure.

The iterative procedure may be summarized as follows:

$\mathbf{p}^{\dagger}$ = observed cumulative relative frequencies

$\mathbf{p} = \mathbf{p}^{\dagger}$

$\mathbf{P}_X = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$

DO OVER $\boldsymbol{\pi}$

$\boldsymbol{\pi} = \mathbf{p}$

$\mathbf{V} = \frac{1}{n}\left\{\mathbf{C}\,\mathrm{diag}\left[\mathbf{C}^{-1}\boldsymbol{\pi}\right]\mathbf{C}' - \boldsymbol{\pi}\boldsymbol{\pi}'\right\}$

$\mathbf{D}_{\pi} = \left(\mathrm{diag}\left[\boldsymbol{\phi}\left(\boldsymbol{\Phi}^{-1}\left(\boldsymbol{\pi}\right)\right)\right]\right)^{-1}$

$\mathbf{z}_{\pi} = \mathbf{P}_X\boldsymbol{\Phi}^{-1}\left(\boldsymbol{\pi}\right)$

$\mathbf{G}_{\pi} = \mathrm{diag}\left[\boldsymbol{\phi}\left(\mathbf{z}_{\pi}\right)\right]\cdot\mathbf{P}_X\cdot\mathbf{D}_{\pi} - \mathbf{I}$

$\mathbf{p} = \mathbf{p}^{\dagger}$

DO OVER $\mathbf{p}$

$\mathbf{D}_p = \left(\mathrm{diag}\left[\boldsymbol{\phi}\left(\boldsymbol{\Phi}^{-1}\left(\mathbf{p}\right)\right)\right]\right)^{-1}$

$\mathbf{z}_p = \mathbf{P}_X\boldsymbol{\Phi}^{-1}\left(\mathbf{p}\right)$

$\mathbf{G}_p = \mathrm{diag}\left[\boldsymbol{\phi}\left(\mathbf{z}_p\right)\right]\cdot\mathbf{P}_X\cdot\mathbf{D}_p - \mathbf{I}$

$\mathbf{g}(\mathbf{p}) = \boldsymbol{\Phi}\left(\mathbf{z}_p\right) - \mathbf{p}$

$\mathbf{p} = \mathbf{p} - \left(\mathbf{G}_{\pi}\mathbf{V}\right)'\left(\mathbf{G}_{\pi}\mathbf{V}\mathbf{G}_p\right)^{*}\mathbf{g}\left(\mathbf{p}\right)$

END

END

For convergence over $\mathbf{p}$ the vector of upper class boundaries $\mathbf{z}_p$, the matrix of partial derivatives $\mathbf{G}_p$ and the vector of constraints $\mathbf{g}(\mathbf{p})$ are all functions of $\mathbf{p}$. Utilizing

$$\mathbf{p} = \mathbf{p} - \left(\mathbf{G}_{\pi}\mathbf{V}\right)'\left(\mathbf{G}_{\pi}\mathbf{V}\mathbf{G}_p\right)^{*}\mathbf{g}\left(\mathbf{p}\right)$$

convergence is obtained over $\mathbf{p}$ resulting in a new value for $\boldsymbol{\pi}$. For convergence over $\boldsymbol{\pi}$ the covariance matrix $\mathbf{V}$, vector of upper class boundaries $\mathbf{z}_{\pi}$ and the matrix of partial derivatives $\mathbf{G}_{\pi}$ are all functions of $\boldsymbol{\pi}$. Convergence over $\boldsymbol{\pi}$ leads to the restricted ML estimate $\widehat{\boldsymbol{\pi}}$ which follows a cumulative

normal distribution curve at the upper class boundaries $\mathbf{x}$. From the restricted ML estimate $\widehat{\boldsymbol{\pi}}$ the ML estimator

$$\widehat{\boldsymbol{\alpha}} = \begin{pmatrix} \widehat{\alpha}_1 \\ \widehat{\alpha}_2 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\boldsymbol{\Phi}^{-1}\,(\widehat{\boldsymbol{\pi}}) \tag{4.16}$$

follows and consequently the ML estimators for the normal distribution are

$$\widehat{\mu} = \frac{\widehat{\alpha}_2}{\widehat{\alpha}_1} \tag{4.17}$$

and

$$\widehat{\sigma} = \frac{1}{\widehat{\alpha}_1}\ . \tag{4.18}$$

See (4.7) for the formulation of the parameters.

Example 4.1

The normal distribution will now be fitted to 100 observations simulated from a normal population with mean 58 and standard deviation 15. The data is shown in Table 4.1.

Table 4.1: Simulated data set from a normal distribution.

Class Interval	Frequency
$[0, 40)$	9
$[40, 50)$	26
$[50, 60)$	24
$[60, 75)$	27
$[75, 100)$	14

The various values for $\mathbf{p}$ and $\boldsymbol{\pi}$ in the double iteration process are calculated with the SAS program *NORM1.SAS* (listed in Appendix A.4) and tabulated in Table 4.2. The corresponding values for μ and σ are also listed in Table 4.2.

Table 4.2: Double iteration process.

Iteration over $\boldsymbol{\pi}$		Iteration over $\mathbf{p}$			
		$j=1$	$j=2$	$j=3$	$j=4$
$i=1$	$\begin{pmatrix}0.0900\\0.3500\\0.5900\\0.8600\end{pmatrix}$	$\begin{pmatrix}0.0900\\0.3500\\0.5900\\0.8600\end{pmatrix}$	$\begin{pmatrix}0.0950\\0.2721\\0.5375\\0.8734\end{pmatrix}$	$\begin{pmatrix}0.0951\\0.2713\\0.5367\\0.8736\end{pmatrix}$	$\begin{pmatrix}0.0951\\0.2713\\0.5367\\0.8736\end{pmatrix}$
	$\mu_\pi = 57.79$	$\mu_p = 57.79$	$\mu_p = 58.68$	$\mu_p = 58.69$	$\mu_p = 58.69$
	$\sigma_\pi = 14.76$	$\sigma_p = 14.76$	$\sigma_p = 14.27$	$\sigma_p = 14.26$	$\sigma_p = 14.26$
$i=2$	$\begin{pmatrix}0.0951\\0.2713\\0.5367\\0.8736\end{pmatrix}$	$\begin{pmatrix}0.0900\\0.3500\\0.5900\\0.8600\end{pmatrix}$	$\begin{pmatrix}0.1196\\0.3094\\0.5670\\0.8791\end{pmatrix}$	$\begin{pmatrix}0.1206\\0.3078\\0.5667\\0.8796\end{pmatrix}$	$\begin{pmatrix}0.1206\\0.3078\\0.5667\\0.8796\end{pmatrix}$
	$\mu_\pi = 58.69$	$\mu_p = 57.79$	$\mu_p = 57.50$	$\mu_p = 57.49$	$\mu_p = 57.49$
	$\sigma_\pi = 14.26$	$\sigma_p = 14.76$	$\sigma_p = 14.92$	$\sigma_p = 14.92$	$\sigma_p = 14.92$
$i=3$	$\begin{pmatrix}0.1206\\0.3078\\0.5667\\0.8796\end{pmatrix}$	$\begin{pmatrix}0.0900\\0.3500\\0.5900\\0.8600\end{pmatrix}$	$\begin{pmatrix}0.1188\\0.3084\\0.5666\\0.8794\end{pmatrix}$	$\begin{pmatrix}0.1197\\0.3068\\0.5663\\0.8799\end{pmatrix}$	$\begin{pmatrix}0.1197\\0.3068\\0.5663\\0.8799\end{pmatrix}$
	$\mu_\pi = 57.49$	$\mu_p = 57.79$	$\mu_p = 57.52$	$\mu_p = 57.52$	$\mu_p = 57.52$
	$\sigma_\pi = 14.92$	$\sigma_p = 14.76$	$\sigma_p = 14.88$	$\sigma_p = 14.89$	$\sigma_p = 14.89$
$i=4$	$\begin{pmatrix}0.1197\\0.3068\\0.5663\\0.8799\end{pmatrix}$	$\begin{pmatrix}0.0900\\0.3500\\0.5900\\0.8600\end{pmatrix}$	$\begin{pmatrix}0.1188\\0.3084\\0.5666\\0.8794\end{pmatrix}$	$\begin{pmatrix}0.1197\\0.3068\\0.5663\\0.8799\end{pmatrix}$	$\begin{pmatrix}0.1197\\0.3068\\0.5663\\0.8799\end{pmatrix}$
	$\mu_\pi = 57.52$	$\mu_p = 57.79$	$\mu_p = 57.52$	$\mu_p = 57.52$	$\mu_p = 57.52$
	$\sigma_\pi = 14.89$	$\sigma_p = 14.76$	$\sigma_p = 14.89$	$\sigma_p = 14.89$	$\sigma_p = 14.89$

From Table 4.2 it can be seen that the ML procedure converges extremely fast. The procedure starts off with the unrestricted ML estimate for $\boldsymbol{\pi}$ (observed cumulative relative frequencies)

$$\boldsymbol{\pi} = \mathbf{p} = \begin{pmatrix} 0.0900 \\ 0.3500 \\ 0.5900 \\ 0.8600 \end{pmatrix}$$

and converges ultimately to the restricted ML estimate of $\boldsymbol{\pi}$

$$\widehat{\boldsymbol{\pi}} = \begin{pmatrix} 0.1197 \\ 0.3068 \\ 0.5663 \\ 0.8799 \end{pmatrix} .$$

The elements of $\widehat{\boldsymbol{\pi}}$ follow a cumulative normal distribution curve at the upper class boundaries of $\mathbf{x}$ and hence the ML estimates of the natural parameters follows from (4.16) with

$$\widehat{\boldsymbol{\alpha}} = \begin{pmatrix} \widehat{\alpha}_1 \\ \widehat{\alpha}_2 \end{pmatrix} = \begin{pmatrix} 0.06717 \\ 3.86338 \end{pmatrix} .$$

From (4.17) and (4.18) the ML estimates for the normal distribution are

$$\widehat{\mu} = 57.52 \quad \text{and} \quad \widehat{\sigma} = 14.89 \ .$$

The estimated normal distribution is shown in Figure 4.1, together with the observed frequency distribution (blue line) and estimated frequency distribution (red line).

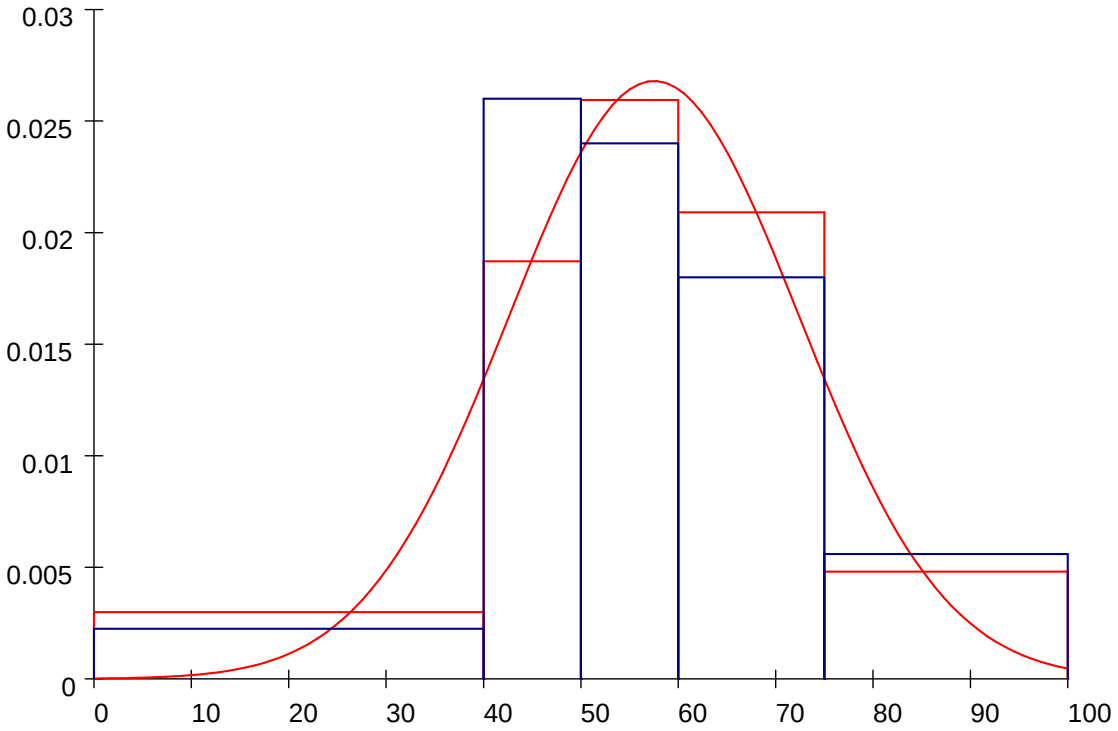

Figure 4.1: The estimated normal distribution with the observed and estimated frequency distribution.

4.2 Constraints in terms of a linear model

In the previous section a normal distribution was fitted to a grouped data set utilizing a direct set of constraints. In this section the constraints will be formulated in terms of a linear model.

From (4.3) it is possible to formulate the linear model

$$\begin{aligned} \boldsymbol{\Phi}^{-1}(\boldsymbol{\pi}) &= \left(\frac{\mathbf{x}-\mu\mathbf{1}}{\sigma}\right) \\ &= \mathbf{X}\boldsymbol{\alpha} \end{aligned} \tag{4.19}$$

where

$$\mathbf{X} = \begin{pmatrix} \mathbf{x} & -\mathbf{1} \end{pmatrix} \tag{4.20}$$

is the design matrix and

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma} \\ \frac{\mu}{\sigma} \end{pmatrix} \tag{4.21}$$

is the vector of natural parameters.

The linear model (4.19) implies the vector of constraints

$$\mathbf{g}(\boldsymbol{\pi}) = \mathbf{Q}_X \boldsymbol{\Phi}^{-1}(\boldsymbol{\pi}) = \mathbf{0} \tag{4.22}$$

to be imposed in the ML estimation procedure, where

$$\mathbf{Q}_X = \mathbf{I} - \mathbf{P}_X \tag{4.23}$$

is the projection matrix orthogonal to $\mathbf{X}$ and $\mathbf{P}_X$ is previously defined in (4.11). According to (4.22) the vector of cumulative probabilities will be fitted such that $\boldsymbol{\Phi}^{-1}(\boldsymbol{\pi})$ is orthogonal to the error space of $\mathbf{X}$ or equivalently such that $\boldsymbol{\Phi}^{-1}(\boldsymbol{\pi})$ is in the vector space of $\mathbf{X}$.

The matrix of partial derivatives follows

$$\begin{aligned} \mathbf{G}_\pi &= \frac{\partial \mathbf{Q}_X \boldsymbol{\Phi}^{-1}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} \\ &= \mathbf{Q}_X \mathbf{D}_\pi \end{aligned} \tag{4.24}$$

where $\mathbf{D}_\pi = \left(\operatorname{diag}\left[\boldsymbol{\phi}\left(\boldsymbol{\Phi}^{-1}(\boldsymbol{\pi})\right)\right]\right)^{-1}$ is already derived in (4.15).

Employing the vector of constraints (4.22) and the matrix of partial derivatives (4.24) in the ML estimation procedure the restricted ML estimate, $\widehat{\boldsymbol{\pi}}$, is obtained. It follows from (4.19) that the ML estimator of $\boldsymbol{\alpha}$ is

$$\widehat{\boldsymbol{\alpha}} = \begin{pmatrix} \widehat{\alpha}_1 \\ \widehat{\alpha}_2 \end{pmatrix} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}' \boldsymbol{\Phi}^{-1}(\widehat{\boldsymbol{\pi}}) \tag{4.25}$$

with asymptotic covariance matrix

$$\begin{aligned} \operatorname{Cov}(\widehat{\boldsymbol{\alpha}}) &\approxeq \left(\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{\pi}}\right) \operatorname{Cov}(\widehat{\boldsymbol{\pi}}) \left(\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{\pi}}\right)' \\ &= \left\{\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}_\pi\right\} \operatorname{Cov}(\widehat{\boldsymbol{\pi}}) \left\{\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}_\pi\right\}' . \end{aligned} \tag{4.26}$$

The ML estimators

$$\widehat{\mu} = \frac{\widehat{\alpha}_2}{\widehat{\alpha}_1} \quad \text{and} \quad \widehat{\sigma} = \frac{1}{\widehat{\alpha}_1} \tag{4.27}$$

follows from (4.25) and (4.21).

Let

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \sigma \end{pmatrix} = \begin{pmatrix} \frac{\alpha_2}{\alpha_1} \\ \frac{1}{\alpha_1} \end{pmatrix} \tag{4.28}$$

denote the vector of original parameters for the normal distribution. To find the asymptotic distribution for the ML estimate $\widehat{\boldsymbol{\beta}}$, the multivariate δ-theorem is once again implemented and hence

$$\widehat{\boldsymbol{\beta}} \approx \mathrm{N}\left(\boldsymbol{\beta}, \mathrm{Cov}(\widehat{\boldsymbol{\beta}})\right) \tag{4.29}$$

$$= \mathrm{N}\left(\begin{pmatrix} \mu \\ \sigma \end{pmatrix}, \mathbf{B}\,\mathrm{Cov}(\widehat{\boldsymbol{\alpha}})\mathbf{B}'\right) \tag{4.30}$$

where

$$\mathbf{B} = \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\alpha}} = \frac{\partial \begin{pmatrix} \mu \\ \sigma \end{pmatrix}}{\partial \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}} = \begin{pmatrix} -\frac{\alpha_2}{\alpha_1^2} & \frac{1}{\alpha_1} \\ -\frac{1}{\alpha_1^2} & 0 \end{pmatrix} . \tag{4.31}$$

Example 4.2

The normal distribution will now be fitted to the frequency distribution tabulated in Table 4.1, now employing the vector of constraints as a linear model (4.22). By making use of the SAS program *NORM2.SAS* in Appendix A.5, the ML estimation procedure yields exactly the same restricted ML estimate for $\boldsymbol{\pi}$, as in Example 4.1, namely

$$\widehat{\boldsymbol{\pi}} = \begin{pmatrix} 0.1197 \\ 0.3068 \\ 0.5663 \\ 0.8799 \end{pmatrix}$$

although the intermediate iterations differ. The elements of

$$\boldsymbol{\Phi}^{-1}(\widehat{\boldsymbol{\pi}}) = \begin{pmatrix} -1.17652 \\ -0.50480 \\ 0.16691 \\ 1.17448 \end{pmatrix}$$

are the estimates of the inverse normal probabilities (standardised upper class boundaries) and

$$\begin{aligned} \mathbf{P}_X &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \begin{pmatrix} 0.64486 & 0.40187 & 0.15888 & -0.20561 \\ 0.40187 & 0.30841 & 0.21495 & 0.07477 \\ 0.15888 & 0.21495 & 0.27103 & 0.35514 \\ -0.20561 & 0.07477 & 0.35514 & 0.77570 \end{pmatrix} \end{aligned}$$

is the projection matrix generated by the columns of $\mathbf{X}$. Multiplying these two matrices lead to

$$\mathbf{P}_X\boldsymbol{\Phi}^{-1}(\widehat{\boldsymbol{\pi}}) = \boldsymbol{\Phi}^{-1}(\widehat{\boldsymbol{\pi}})$$

which means that $\boldsymbol{\Phi}^{-1}(\widehat{\boldsymbol{\pi}})$ is in the vector space of $\mathbf{X}$ and consequently $\boldsymbol{\Phi}^{-1}(\widehat{\boldsymbol{\pi}})$ is a linear combination of the columns of $\mathbf{X}$ in (4.20). It is also clear that

$$\mathbf{Q}_X\boldsymbol{\Phi}^{-1}(\widehat{\boldsymbol{\pi}}) = \mathbf{0}$$

indicating that $\boldsymbol{\Phi}^{-1}(\widehat{\boldsymbol{\pi}})$ is orthogonal to the error space of $\mathbf{X}$. (See 4.22 and 4.23.)

The ML estimates and goodness of fit statistics are summarized in Table 4.3

Table 4.3: ML estimates and goodness of fit statistics for the normal distribution.

MLE		Goodness of fit			
Estimate	Std. error	Statistic	Value	df	prob
$\widehat{\mu} = 57.515$	$\widehat{\sigma}_{\widehat{\mu}} = 1.556$	Pearson	4.654	2	0.0976
$\widehat{\sigma} = 14.887$	$\widehat{\sigma}_{\widehat{\sigma}} = 1.327$	Wald	4.855	2	0.1455

According to the goodness of fit statistics summarized in Table 4.3, the null hypothesis of an adequate fit is not rejected at a 5% level of significance. The adequate fit is further illustrated in Figure 4.1.

The estimated standard errors $\widehat{\sigma}_{\widehat{\mu}}$ and $\widehat{\sigma}_{\widehat{\sigma}}$ in Table 4.3 follows from the estimated covariance matrix

$$\widehat{\mathrm{Cov}}\left(\widehat{\boldsymbol{\beta}}\right) = \widehat{\mathrm{Cov}}\begin{pmatrix}\widehat{\mu}\\ \widehat{\sigma}\end{pmatrix} = \begin{pmatrix}2.4219 & 0.0353\\ 0.0353 & 1.7622\end{pmatrix}$$

which is estimated by substituting the restricted ML estimate $\widehat{\boldsymbol{\pi}}$ in $\mathrm{Cov}\left(\widehat{\boldsymbol{\pi}}\right)$.

The 95% confidence intervals for μ and σ are tabulated in Table 4.4.

Table 4.4: 95% confidence intervals for μ and σ.

Parameter	Margin of error	Confidence interval
μ	$1.96\,(1.556) = 3.049$	$(54.951, 61.049)$
σ	$1.96\,(1.327) = 2.601$	$(12.286, 17.488)$

From the confidence intervals reported in Table 4.4 the population parameters μ and σ do not differ significantly from the theoretical values 58 and 15.

4.3 Simulation study

Similar to the simulation study done for the exponential distribution in the previous chapter, 1000 samples were simulated, each containing 100 observations. These samples were all simulated from a normal population with mean $\mu = 58$ and standard deviation $\sigma = 15$. The descriptive statistics for the 1000 sample means and sample standard deviations of the ungrouped data sets are summarised in Table 4.5.

Table 4.5: Descriptive statistics for sample statistics of ungrouped data sets.

Statistic	Mean	Std. deviation	P_5	Median	P_{95}
$\bar{x}$	57.993	1.489	55.582	57.919	60.446
s	14.902	1.078	13.244	14.881	16.673

Evaluating the sample statistics for the ungrouped data sets, the mean and median are very close to the theoretical values. The standard deviation of $\bar{x}$ is close to the standard error of $\bar{x}$, i.e.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = 1.5 \ .$$

The 1000 simulated data sets were all classified into the same set of class intervals as that of Table 4.1. The normal distribution was fitted to each of the 1000 generated frequency distributions and the descriptive statistics for the ML estimates are tabulated in Table 4.6.

Table 4.6: Simulation results for the normal distribution.

MLE	Theoretical Value	Mean	Std. deviation	P_5	Median	P_{95}
$\widehat{\mu}$	58.000	57.993	1.548	55.512	57.945	60.598
$\widehat{\sigma}_{\widehat{\mu}}$	1.569	1.562	0.146	1.343	1.550	1.826
$\widehat{\sigma}$	15.000	14.915	1.384	12.797	14.823	17.376
$\widehat{\sigma}_{\widehat{\sigma}}$	1.341	1.340	0.171	1.091	1.320	1.653

In the case of a normal distribution with $\mu = 58$ and $\sigma = 15$ the theoretical value for $\boldsymbol{\pi}$ is

$$\boldsymbol{\pi} = \boldsymbol{\Phi}\left(\frac{\mathbf{x} - 58\,(\mathbf{1})}{15}\right) = \boldsymbol{\Phi}\begin{pmatrix} -1.2000 \\ -0.5333 \\ 0.1333 \\ 1.1333 \end{pmatrix} = \begin{pmatrix} 0.11507 \\ 0.29690 \\ 0.55304 \\ 0.87146 \end{pmatrix}$$

leading to the asymptotic covariance matrix

$$\mathrm{Cov}\begin{pmatrix} \widehat{\mu} \\ \widehat{\sigma} \end{pmatrix} \approxeq \begin{pmatrix} 2.46085 & 0.05201 \\ 0.05201 & 1.79748 \end{pmatrix}$$

and yielding the standard errors $\sigma_{\widehat{\mu}} = 1.569$ and $\sigma_{\widehat{\sigma}} = 1.341$ tabulated in Table 4.6. In view of the fact that the standard error for a random sample from a $\mathrm{N}(58, 15^2)$ distribution is $\frac{15}{\sqrt{100}} = 1.5$, not much accuracy has been lost by using a grouped sample in the estimation of μ. As is evident from Table 4.6 the mean and median of each of the ML estimates compare extremely well with the theoretical values (approximate in the case of $\sigma_{\widehat{\mu}}$ and $\sigma_{\widehat{\sigma}}$). It is also interesting to note that standard deviations for $\widehat{\mu}$ and $\widehat{\sigma}$ are close to the standard errors $\sigma_{\widehat{\mu}}$ and $\sigma_{\widehat{\sigma}}$. To evaluate the fifth and the ninety fifth percentiles the margin of error for the 90% confidence intervals are summarised in Table 4.7.

Table 4.7: 90% margin of error for the ML estimators of the normal distribution.

Estimate	Std. Error	Margin of Error
$\widehat{\mu}$	$\sigma_{\widehat{\mu}}$	$1.645\sigma_{\widehat{\mu}} = 2.581$
$\widehat{\sigma}$	$\sigma_{\widehat{\sigma}}$	$1.645\sigma_{\widehat{\sigma}} = 2.206$

It is known that approximately 90% of the $\widehat{\mu}$-values should be in the interval $(55.419, 60.581)$, while 90% of the $\widehat{\sigma}$-values should be in the interval $(12.794, 17.206)$. This compares well with the simulated values in Table 4.6.

The goodness of fit statistics were calculated for each of the 1000 fitted normal distributions. From Table 4.8 it follows that the Pearson and Wald statistics correspond very well to that of a χ^2-distribution with 2 degrees of freedom.

Table 4.8: Percentiles of the Pearson and Wald statistic.

	Percentiles						
	P_5	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	P_{95}
Pearson	0.1291	0.2355	0.5945	1.3728	2.7147	4.6345	5.8393
Wald	0.1066	0.2054	0.5925	1.3742	2.7591	4.6721	6.1128
	Percentiles of a χ^2-distribution with 2 degrees of freedom.						
	$\chi^2_{0.05}$	$\chi^2_{0.10}$	$\chi^2_{0.25}$	$\chi^2_{0.50}$	$\chi^2_{0.75}$	$\chi^2_{0.90}$	$\chi^2_{0.95}$
$\chi^2(2)$	0.1026	0.2107	0.5754	1.3863	2.7726	4.6052	5.9915

Chapter 5

The Weibull, log-logistic and Pareto distributions

In this chapter it will be shown how to fit the Weibull, log-logistic and Pareto distributions to a grouped data set. Estimation will be done by constructing the vector of constraints in terms of a linear model. This method is preferred due to the simplicity and the overall generalization of the technique. This generalization is outlined in 3 easy steps where the estimation of the exponential and normal distributions are also considered.

5.1 The Weibull distribution

The pdf of the Weibull distribution is

$$f(x;\kappa,\theta) = \frac{\kappa}{\theta^{\kappa}} x^{\kappa-1} \exp\left[-\left(\frac{x}{\theta}\right)^{\kappa}\right] \tag{5.1}$$

with cdf

$$F(x;\kappa,\theta) = 1 - \exp\left[-\left(\frac{x}{\theta}\right)^{\kappa}\right] \ . \tag{5.2}$$

The parameter κ is a shape parameter with θ the so-called scale parameter. The three basic shapes of the Weibull distribution are illustrated in Figure 5.1.

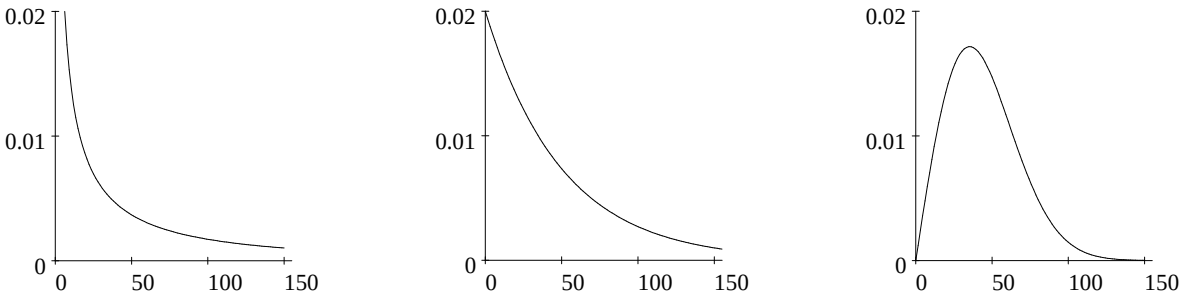

Figure 5.1: Weibull distributions with $\theta = 50$.

The mean and variance of the Weibull distribution are

$$\mu = \theta \left[\Gamma \left(1 + \frac{1}{\kappa} \right) \right] \tag{5.3}$$

and

$$\sigma^2 = \theta^2 \left[\Gamma \left(1 + \frac{2}{\kappa} \right) - \Gamma^2 \left(1 + \frac{1}{\kappa} \right) \right] \tag{5.4}$$

respectively.

To fit a Weibull distribution it is required that

$$\boldsymbol{\pi} = \mathbf{1} - \exp \left[- \left(\frac{\mathbf{x}}{\theta} \right)^{\kappa} \right] \tag{5.5}$$

which implies that

$$\ln \left(\mathbf{1} - \boldsymbol{\pi} \right) = - \left(\frac{\mathbf{x}}{\theta} \right)^{\kappa} . \tag{5.6}$$

Taking the natural logarithm of (5.6) yields the linear model

$$\begin{aligned} \ln \left[- \ln \left(\mathbf{1} - \boldsymbol{\pi} \right) \right] &= \kappa \ln \mathbf{x} - \left(\kappa \ln \theta \right) \mathbf{1} \\ &= \begin{pmatrix} \ln \mathbf{x} & -\mathbf{1} \end{pmatrix} \begin{pmatrix} \kappa \\ \kappa \ln \theta \end{pmatrix} \\ &= \mathbf{X} \boldsymbol{\alpha} \end{aligned} \tag{5.7}$$

where

$$\mathbf{X} = \begin{pmatrix} \ln \mathbf{x} & -\mathbf{1} \end{pmatrix} \tag{5.8}$$

is the design matrix and

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} \kappa \\ \kappa \ln \theta \end{pmatrix} \tag{5.9}$$

is the vector of natural parameters.

The vector of constraints

$$\mathbf{g}(\boldsymbol{\pi}) = \mathbf{Q}_X \ln \left[-\ln \left(\mathbf{1} - \boldsymbol{\pi} \right) \right] = \mathbf{0} \tag{5.10}$$

follows from (5.7) where $\mathbf{Q}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the projection matrix orthogonal to $\mathbf{X}$. The matrix of partial derivatives becomes

$$\begin{aligned} \mathbf{G}_\pi &= \frac{\partial \mathbf{g}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} \\ &= \frac{\partial \left\{ \mathbf{Q}_X \ln \left[-\ln \left(\mathbf{1} - \boldsymbol{\pi} \right) \right] \right\}}{\partial \boldsymbol{\pi}} \\ &= \mathbf{Q}_X \mathbf{D}_\pi \end{aligned} \tag{5.11}$$

where

$$\begin{aligned} \mathbf{D}_\pi &= \frac{\partial \ln \left[-\ln \left(\mathbf{1} - \boldsymbol{\pi} \right) \right]}{\partial \boldsymbol{\pi}} \\ &= \left\{ \operatorname{diag} \left[-\ln \left(\mathbf{1} - \boldsymbol{\pi} \right) \right] \right\}^{-1} \frac{\partial}{\partial \boldsymbol{\pi}} \left\{ -\ln \left(\mathbf{1} - \boldsymbol{\pi} \right) \right\} \\ &= -\left\{ \operatorname{diag} \left[\ln \left(\mathbf{1} - \boldsymbol{\pi} \right) \right] \right\}^{-1} \left\{ \operatorname{diag} \left[\mathbf{1} - \boldsymbol{\pi} \right] \right\}^{-1} . \end{aligned} \tag{5.12}$$

The restricted ML estimate $\widehat{\boldsymbol{\pi}}$ is estimated such that $\ln \left[-\ln \left(\mathbf{1} - \widehat{\boldsymbol{\pi}} \right) \right]$ is a linear combination of $\mathbf{X}$ leading to the ML estimator

$$\widehat{\boldsymbol{\alpha}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \ln \left[-\ln \left(\mathbf{1} - \widehat{\boldsymbol{\pi}} \right) \right] \tag{5.13}$$

with asymptotic covariance matrix

$$\operatorname{Cov}(\widehat{\boldsymbol{\alpha}}) \approxeq \left\{ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi \right\} \operatorname{Cov}(\widehat{\boldsymbol{\pi}}) \left\{ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi \right\}' . \tag{5.14}$$

The parameters of the Weibull distribution are

$$\boldsymbol{\beta} = \begin{pmatrix} \kappa \\ \theta \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \exp\left(\frac{\alpha_2}{\alpha_1} \right) \end{pmatrix} . \tag{5.15}$$

Hence, the ML estimator for $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\kappa} \\ \widehat{\theta} \end{pmatrix} = \begin{pmatrix} \widehat{\alpha}_1 \\ \exp\left(\frac{\widehat{\alpha}_2}{\widehat{\alpha}_1}\right) \end{pmatrix} \tag{5.16}$$

with asymptotic covariance matrix

$$\operatorname{Cov}(\widehat{\boldsymbol{\beta}}) \approxeq \mathbf{B} \operatorname{Cov}(\widehat{\boldsymbol{\alpha}}) \mathbf{B}' \tag{5.17}$$

where

$$\begin{aligned} \mathbf{B} &= \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\alpha}} \\ &= \frac{\partial \begin{pmatrix} \kappa \\ \theta \end{pmatrix}}{\partial \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}} \\ &= \begin{pmatrix} 1 & 0 \\ -\frac{\alpha_2}{\alpha_1^2} \exp\left(\frac{\alpha_2}{\alpha_1}\right) & \frac{1}{\alpha_1} \exp\left(\frac{\alpha_2}{\alpha_1}\right) \end{pmatrix} . \end{aligned} \tag{5.18}$$

According to the multivariate delta theorem the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ is

$$\widehat{\boldsymbol{\beta}} \approxeq \mathrm{N}\left(\boldsymbol{\beta}, \mathbf{B} \operatorname{Cov}(\widehat{\boldsymbol{\alpha}}) \mathbf{B}'\right) .$$

5.2 The log-logistic distribution

The log-logistic distribution is defined in a manner analogous to the definition of the lognormal distribution. If $\log(x)$ follows a logistic distribution then x is said to follow a log-logistic distribution.

The pdf of the log-logistic distribution is

$$f(x; \kappa, \theta) = \frac{e^{\theta} \kappa x^{\kappa - 1}}{(1 + e^{\theta} x^{\kappa})^2} \tag{5.19}$$

with cdf

$$F(x; \kappa, \theta) = \frac{e^{\theta} x^{\kappa}}{1 + e^{\theta} x^{\kappa}} . \tag{5.20}$$

Setting $F(x;\kappa,\theta)=\pi$ it follows that

$$\frac{e^{\theta}x^{\kappa}}{1+e^{\theta}x^{\kappa}}=\pi$$

and therefore

$$\begin{aligned}\frac{\pi}{1-\pi} &= \frac{\left(e^{\theta}x^{\kappa}\right)/\left(1+e^{\theta}x^{\kappa}\right)}{\left(1+e^{\theta}x^{\kappa}-e^{\theta}x^{\kappa}\right)/\left(1+e^{\theta}x^{\kappa}\right)} \\ &= e^{\theta}x^{\kappa}\ . \end{aligned} \tag{5.21}$$

The mean and variance are given by

$$\mu=\exp\left(-\frac{\theta}{\kappa}\right)\left[\Gamma\left(1+\frac{1}{\kappa}\right)\Gamma\left(1-\frac{1}{\kappa}\right)\right] \tag{5.22}$$

and

$$\sigma^2=\exp\left(-\frac{2\theta}{\kappa}\right)\left[\Gamma\left(1+\frac{2}{\kappa}\right)\Gamma\left(1-\frac{2}{\kappa}\right)-\Gamma^2\left(1+\frac{1}{\kappa}\right)\Gamma^2\left(1-\frac{1}{\kappa}\right)\right] \tag{5.23}$$

respectively.

Implementing $\boldsymbol{\pi}=F(\mathbf{x})$, it follows from (5.21) that

$$\frac{\boldsymbol{\pi}}{\mathbf{1}-\boldsymbol{\pi}}=e^{\theta}\mathbf{x}^{\kappa}$$

resulting in the linear model

$$\begin{aligned}\ln\left(\frac{\boldsymbol{\pi}}{\mathbf{1}-\boldsymbol{\pi}}\right) &= \kappa\ln\mathbf{x}+\theta\mathbf{1} \\ &= \begin{pmatrix}\ln\mathbf{x} & \mathbf{1}\end{pmatrix}\begin{pmatrix}\kappa\\ \theta\end{pmatrix} \\ &= \mathbf{X}\boldsymbol{\alpha}\end{aligned} \tag{5.24}$$

where

$$\mathbf{X}=\begin{pmatrix}\ln\mathbf{x} & \mathbf{1}\end{pmatrix} \tag{5.25}$$

and

$$\boldsymbol{\alpha}=\begin{pmatrix}\kappa\\ \theta\end{pmatrix}\ . \tag{5.26}$$

The constraints formulated in terms of a linear model is

$$\mathbf{g}(\boldsymbol{\pi}) = \mathbf{Q}_X \ln\left(\frac{\boldsymbol{\pi}}{\mathbf{1}-\boldsymbol{\pi}}\right) = \mathbf{0} \tag{5.27}$$

with matrix of partial derivatives

$$\begin{aligned}\mathbf{G}_\pi &= \frac{\partial \mathbf{g}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} \\ &= \frac{\partial \mathbf{Q}_X \ln\left(\frac{\boldsymbol{\pi}}{\mathbf{1}-\boldsymbol{\pi}}\right)}{\partial \boldsymbol{\pi}} \\ &= \mathbf{Q}_X \mathbf{D}_\pi\end{aligned}$$

where $\mathbf{Q}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and

$$\begin{aligned}\mathbf{D}_\pi &= \frac{\partial}{\partial \boldsymbol{\pi}}\left\{\ln\left(\frac{\boldsymbol{\pi}}{\mathbf{1}-\boldsymbol{\pi}}\right)\right\} \\ &= \frac{\partial}{\partial \boldsymbol{\pi}}\left\{\ln(\boldsymbol{\pi}) - \ln(\mathbf{1}-\boldsymbol{\pi})\right\} \\ &= \left\{\mathrm{diag}[\boldsymbol{\pi}]\right\}^{-1} + \left\{\mathrm{diag}[\mathbf{1}-\boldsymbol{\pi}]\right\}^{-1} \quad .\end{aligned} \tag{5.28}$$

In the ML estimation procedure $\widehat{\boldsymbol{\pi}}$ is estimated such that $\ln\left(\frac{\widehat{\boldsymbol{\pi}}}{\mathbf{1}-\widehat{\boldsymbol{\pi}}}\right)$ is in the vector space of $\mathbf{X}$. The ML estimator $\widehat{\boldsymbol{\alpha}}$ follows from the linear model (5.24)

$$\widehat{\boldsymbol{\alpha}} = \begin{pmatrix}\widehat{\kappa} \\ \widehat{\theta}\end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \ln\left(\frac{\widehat{\boldsymbol{\pi}}}{\mathbf{1}-\widehat{\boldsymbol{\pi}}}\right) \tag{5.29}$$

with asymptotic covariance matrix

$$\mathrm{Cov}(\widehat{\boldsymbol{\alpha}}) = \mathrm{Cov}\begin{pmatrix}\widehat{\kappa} \\ \widehat{\theta}\end{pmatrix} = \left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi\right\}\mathrm{Cov}(\widehat{\boldsymbol{\alpha}})\left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi\right\}' \tag{5.30}$$

where $\mathbf{D}_\pi$ is derived in (5.28). As in the case of the Weibull distribution, the ML estimators of the log-logistic are approximately normally distributed.

5.3 The Pareto distribution

The Pareto distribution has been successfully used to model the income of a population (Johnson & Kotz (1970)). The pdf and cdf of the Pareto distribution are

$$f(x,\kappa,\theta) = \kappa\theta^{\kappa}x^{-(\kappa+1)} \tag{5.31}$$

and

$$F(x) = 1 - \left(\frac{x}{\theta}\right)^{-\kappa} \tag{5.32}$$

for $x > \theta,\ \theta > 0$ and $\kappa > 0$.

The mean and variance for the Pareto distribution are given by

$$\mu = \frac{\kappa\theta}{\kappa - 1} \qquad \kappa > 1 \tag{5.33}$$

and

$$\sigma^2 = \frac{\kappa\theta^2}{(\kappa-1)^2(\kappa-2)} \qquad \kappa > 2 \tag{5.34}$$

respectively.

To fit a Pareto distribution it is required that

$$\boldsymbol{\pi} = \mathbf{1} - \left(\frac{\mathbf{x}}{\theta}\right)^{-\kappa} . \tag{5.35}$$

Taking the natural logarithm of (5.35) leads to

$$\begin{aligned}
\ln(\mathbf{1}-\boldsymbol{\pi}) &= -\kappa \ln\left(\frac{\mathbf{x}}{\theta}\right) \\
&= -\kappa(\ln\mathbf{x} - \ln\theta) \\
&= \begin{pmatrix} -\ln\mathbf{x} & \mathbf{1} \end{pmatrix} \begin{pmatrix} \kappa \\ \kappa\ln\theta \end{pmatrix} \\
&= \mathbf{X}\boldsymbol{\alpha}
\end{aligned} \tag{5.36}$$

where

$$\mathbf{X} = \begin{pmatrix} -\ln\mathbf{x} & \mathbf{1} \end{pmatrix} \tag{5.37}$$

is the design matrix and

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} \kappa \\ \kappa\ln\theta \end{pmatrix} \tag{5.38}$$

is the vector of natural parameters.

Hence, the vector of constraints may be written as

$$\mathbf{g}(\boldsymbol{\pi}) = \mathbf{Q}_X \ln(\mathbf{1} - \boldsymbol{\pi}) = \mathbf{0} \tag{5.39}$$

where $\mathbf{Q}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. This implies that the restricted ML estimate $\widehat{\boldsymbol{\pi}}$ will be fitted such that $\ln(\mathbf{1} - \boldsymbol{\pi})$ is orthogonal to the error space of $\mathbf{X}$ with matrix of partial derivatives

$$\begin{aligned} \mathbf{G}_\pi &= \mathbf{Q}_X \frac{\partial \mathbf{g}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} \\ &= \frac{\partial \mathbf{Q}_X \ln(\mathbf{1} - \boldsymbol{\pi})}{\partial \boldsymbol{\pi}} \\ &= \mathbf{Q}_X \mathbf{D}_\pi \end{aligned}$$

where

$$\begin{aligned} \mathbf{D}_\pi &= \frac{\partial \ln(\mathbf{1} - \boldsymbol{\pi})}{\partial \boldsymbol{\pi}} \\ &= -\{\mathrm{diag}[\mathbf{1} - \boldsymbol{\pi}]\}^{-1} \quad . \end{aligned} \tag{5.40}$$

The ML estimator for $\boldsymbol{\alpha}$ follows

$$\widehat{\boldsymbol{\alpha}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \ln(\mathbf{1} - \widehat{\boldsymbol{\pi}}) \tag{5.41}$$

with asymptotic covariance matrix

$$\mathrm{Cov}(\widehat{\boldsymbol{\alpha}}) = \left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi\right\} \mathrm{Cov}(\widehat{\boldsymbol{\pi}}) \left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi\right\}' \quad . \tag{5.42}$$

Define the vector of parameters for the Pareto distribution

$$\boldsymbol{\beta} = \begin{pmatrix} \kappa \\ \theta \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \exp\left(\dfrac{\alpha_2}{\alpha_1}\right) \end{pmatrix} \quad . \tag{5.43}$$

(The parameterization follows from (5.38).)

Therefore the ML estimates for κ and θ are

$$\widehat{\kappa} = \widehat{\alpha}_1 \quad \text{and} \quad \widehat{\theta} = \exp\left(\frac{\widehat{\alpha}_2}{\widehat{\alpha}_1}\right)$$

implying that

$$\widehat{\boldsymbol{\beta}} \approxeq \mathrm{N}\left(\begin{pmatrix}\kappa\\ \theta\end{pmatrix}, \mathbf{B}\,\mathrm{Cov}(\widehat{\boldsymbol{\alpha}})\mathbf{B}'\right)$$

where

$$\mathbf{B} = \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\alpha}} = \begin{pmatrix} 1 & 0 \\ -\frac{\alpha_2}{\alpha_1^2}\cdot\exp\left(\frac{\alpha_2}{\alpha_1}\right) & \frac{1}{\alpha_1}\cdot\exp\left(\frac{\alpha_2}{\alpha_1}\right)\end{pmatrix}.$$

5.4 Generalization

In this section a short summary of fitting the distributions, tabulated in Table 5.1 will be given.

Table 5.1: Characteristics of distributions considered.

	PDF and CDF	Mean and Variance
Exponential	$f(x;\mu) = \frac{1}{\mu}e^{-x/\mu}$ $F(x;\mu) = 1 - e^{-x/\mu}$	μ $\sigma^2 = \mu^2$
Normal	$f(x;\mu,\sigma^2) = \boldsymbol{\phi}\left(\frac{x-\mu}{\sigma}\right)$ $F(x;\mu,\sigma^2) = \boldsymbol{\Phi}\left(\frac{x-\mu}{\sigma}\right)$	μ σ^2
Weibull	$f(x;\kappa,\theta) = \frac{\kappa}{\theta^\kappa}x^{\kappa-1}\exp\left[-\left(\frac{x}{\theta}\right)^\kappa\right]$ $F(x;\kappa,\theta) = 1 - \exp\left[-\left(\frac{x}{\theta}\right)^\kappa\right]$	$\mu = \theta\left[\Gamma\left(1+\frac{1}{\kappa}\right)\right]$ $\sigma^2 = \theta^2\left[\Gamma\left(1+\frac{2}{\kappa}\right) - \Gamma^2\left(1+\frac{1}{\kappa}\right)\right]$
Log-logistic	$f(x;\kappa,\theta) = \frac{e^\theta \kappa x^{\kappa-1}}{(1+e^\theta x^\kappa)^2}$ $F(x;\kappa,\theta) = \frac{e^\theta x^\kappa}{1+e^\theta x^\kappa}$	$\mu = \exp\left(-\frac{\theta}{\kappa}\right)\left[\Gamma\left(1+\frac{1}{\kappa}\right)\Gamma\left(1-\frac{1}{\kappa}\right)\right]$ $\sigma^2 = \exp\left(-\frac{2\theta}{\kappa}\right)\left[\Gamma\left(1+\frac{2}{\kappa}\right)\Gamma\left(1-\frac{2}{\kappa}\right)\right.$ $\left.-\Gamma^2\left(1+\frac{1}{\kappa}\right)\Gamma^2\left(1-\frac{1}{\kappa}\right)\right]$
Pareto	$f(x;\kappa,\theta) = \kappa\theta^\kappa x^{-(\kappa+1)}$ $F(x;\kappa,\theta) = 1 - \left(\frac{x}{\theta}\right)^{-\kappa}$	$\mu = \frac{\kappa\theta}{\kappa-1}$ $\sigma^2 = \frac{\kappa\theta^2}{(\kappa-1)^2(\kappa-2)}$

In the case of the distributions $F(x;\boldsymbol{\beta})$, specified in Table 5.1, the requirement

$$F(\mathbf{x};\boldsymbol{\beta}) = \boldsymbol{\pi} \tag{5.44}$$

where $F(\mathbf{x};\boldsymbol{\beta})$ denotes the distribution function at the upper class boundaries $\mathbf{x}$ with parameter vector $\boldsymbol{\beta}$, may be transformed into the linear model

$$\mathbf{h}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\alpha} \tag{5.45}$$

which implies that the estimation procedure may be performed in the three steps outlined below.

Step 1: The vector of constraints is given by

$$\mathbf{g}(\boldsymbol{\pi}) = \mathbf{Q}_X\mathbf{h}(\boldsymbol{\pi}) = \mathbf{0} \tag{5.46}$$

with matrix of partial derivatives

$$\mathbf{G}_\pi = \mathbf{Q}_X\mathbf{D}_\pi \tag{5.47}$$

where $\mathbf{Q}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{D}_\pi = \dfrac{\partial \mathbf{h}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}$.

Step 2: The ML estimate of $\boldsymbol{\alpha}$ follows as

$$\widehat{\boldsymbol{\alpha}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{h}(\widehat{\boldsymbol{\pi}}) \tag{5.48}$$

with asymptotic covariance matrix

$$\operatorname{Cov}(\widehat{\boldsymbol{\alpha}}) \approxeq \left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi\right\} \operatorname{Cov}(\widehat{\boldsymbol{\pi}}) \left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi\right\}' . \tag{5.49}$$

Step 3: The ML estimates of the original parameters namely $\widehat{\boldsymbol{\beta}}$, are obtained from $\widehat{\boldsymbol{\alpha}}$ with

$$\operatorname{Cov}\left(\widehat{\boldsymbol{\beta}}\right) \approxeq \mathbf{B} \operatorname{Cov}(\widehat{\boldsymbol{\alpha}}) \mathbf{B}' \tag{5.50}$$

where $\mathbf{B} = \dfrac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\alpha}}$. From the multivariate delta theorem, it follows that

$$\widehat{\boldsymbol{\beta}} \approxeq \mathrm{N}\left(\boldsymbol{\beta}, \mathbf{B} \operatorname{Cov}(\widehat{\boldsymbol{\alpha}})\mathbf{B}'\right) . \tag{5.51}$$

To fit the various continuous distributions in Table 5.1 to grouped data by means of the three steps listed above, a summary of the constraints and derivatives are given in Table 5.2(A) & Table 5.2(B).

Table 5.2(A): Constraints

	$\boldsymbol{\beta}$	$\mathbf{h}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\alpha}$		
		$\mathbf{h}(\boldsymbol{\pi})$	$\mathbf{X}$	$\boldsymbol{\alpha}$
Exponential	$\mu = \frac{1}{\alpha}$	$\ln(\mathbf{1}-\boldsymbol{\pi})$	$(-\mathbf{x})$	$\frac{1}{\mu}$
Normal	$\begin{pmatrix}\mu\\ \sigma\end{pmatrix} = \begin{pmatrix}\frac{\alpha_2}{\alpha_1}\\ \frac{1}{\alpha_1}\end{pmatrix}$	$\boldsymbol{\Phi}^{-1}(\boldsymbol{\pi})$	$\begin{pmatrix}\mathbf{x} & -\mathbf{1}\end{pmatrix}$	$\begin{pmatrix}\frac{1}{\sigma}\\ \frac{\mu}{\sigma}\end{pmatrix}$
Weibull	$\begin{pmatrix}\kappa\\ \theta\end{pmatrix} = \begin{pmatrix}\alpha_1\\ e^{\frac{\alpha_2}{\alpha_1}}\end{pmatrix}$	$\ln[-\ln(\mathbf{1}-\boldsymbol{\pi})]$	$\begin{pmatrix}\ln\mathbf{x} & -\mathbf{1}\end{pmatrix}$	$\begin{pmatrix}\kappa\\ \kappa\ln\theta\end{pmatrix}$
Log-logistic	$\begin{pmatrix}\kappa\\ \theta\end{pmatrix} = \begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}$	$\ln\left(\frac{\boldsymbol{\pi}}{\mathbf{1}-\boldsymbol{\pi}}\right)$	$\begin{pmatrix}\ln\mathbf{x} & \mathbf{1}\end{pmatrix}$	$\begin{pmatrix}\kappa\\ \theta\end{pmatrix}$
Pareto	$\begin{pmatrix}\kappa\\ \theta\end{pmatrix} = \begin{pmatrix}\alpha_1\\ e^{\frac{\alpha_2}{\alpha_1}}\end{pmatrix}$	$\ln(\mathbf{1}-\boldsymbol{\pi})$	$\begin{pmatrix}-\ln\mathbf{x} & \mathbf{1}\end{pmatrix}$	$\begin{pmatrix}\kappa\\ \kappa\ln\theta\end{pmatrix}$

Table 5.2(B): Derivatives

	$\mathbf{D} = \frac{\partial \mathbf{h}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}$	$\mathbf{B} = \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\alpha}}$
Exponential	$-(\operatorname{diag}[\mathbf{1}-\boldsymbol{\pi}])^{-1}$	$-\frac{1}{\alpha^2}$
Normal	$(\operatorname{diag}[\boldsymbol{\phi}(\boldsymbol{\Phi}^{-1}(\boldsymbol{\pi}))])^{-1}$	$\begin{pmatrix}-\frac{\alpha_2}{\alpha_1^2} & \frac{1}{\alpha_1}\\ -\frac{1}{\alpha_1^2} & 0\end{pmatrix}$
Weibull	$-(\operatorname{diag}[\ln(\mathbf{1}-\boldsymbol{\pi})])^{-1}(\operatorname{diag}[\mathbf{1}-\boldsymbol{\pi}])^{-1}$	$\begin{pmatrix}1 & 0\\ -\frac{\alpha_2}{\alpha_1^2}\cdot e^{\frac{\alpha_2}{\alpha_1}} & \frac{1}{\alpha_1}\cdot e^{\frac{\alpha_2}{\alpha_1}}\end{pmatrix}$
Log-logistic	$(\operatorname{diag}[\boldsymbol{\pi}])^{-1} + (\operatorname{diag}[\mathbf{1}-\boldsymbol{\pi}])^{-1}$	$\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$
Pareto	$-(\operatorname{diag}[\mathbf{1}-\boldsymbol{\pi}])^{-1}$	$\begin{pmatrix}1 & 0\\ -\frac{\alpha_2}{\alpha_1^2}\cdot e^{\frac{\alpha_2}{\alpha_1}} & \frac{1}{\alpha_1}\cdot e^{\frac{\alpha_2}{\alpha_1}}\end{pmatrix}$

Example 5.1

A typical example was taken from a data set with $n = 206$ insurance policies. The annual income (in R1000) of the policy holders is reported in Table 5.3.

Table 5.3: Income of a group of insurance policy holders.

Income (in R1000)	$[0, 40)$	$[40, 75)$	$[75, 125)$	$[125, 175)$	$[175, \infty)$
Frequency	9	37	67	63	30

For this example the normal, Weibull and log-logistic distributions are fitted and the results are given in Table 5.4.

Table 5.4: Estimates of parameters and test statistics

	MLE					Wald			Discrep-
	$\widehat{\boldsymbol{\beta}}$	Estimate	Std. Error	$\widehat{\mu}$	$\widehat{\sigma}$	Statistic	df	prob	ancy
Normal	$\widehat{\mu}$	118.4	3.7604						
	$\widehat{\sigma}$	51.4	3.0834	118.4	51.4	3.980	2	0.1367	0.019
Weibull	$\widehat{\kappa}$	2.4647	0.1675						
	$\widehat{\theta}$	134.44	4.2552	119.2	51.7	1.293	2	0.5240	0.006
Log-	$\widehat{\kappa}$	3.3337	0.2293						
logistic	$\widehat{\theta}$	−15.710	1.0883	129.7	88.0	8.731	2	0.0127	0.042

According to the Wald statistic the Weibull distribution provided the best fit, followed by the normal distribution. The distributions are illustrated in Figure 5.2. In constructing the histogram, it is assumed that the income of all the policy holders in the sample is less than R500 000. The distributions were all fitted with the SAS program *FIT.SAS* listed in Appendix A.

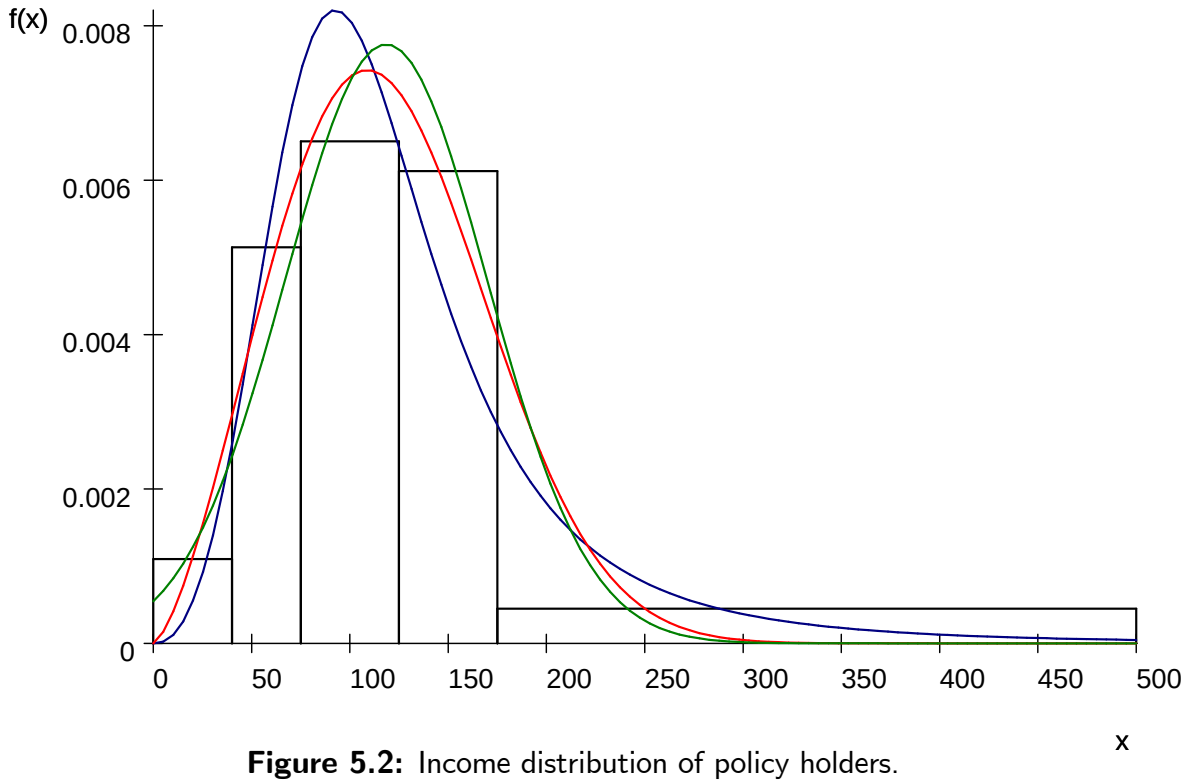

Figure 5.2: Income distribution of policy holders.

Normal: *Green*	Weibull: *Red*	Log-logistic: *Blue*

Part II

Linear models for grouped data

Chapter 6

Multifactor design

Consider any single-factor or multifactor design resulting in a cross classification of T different cells to be analysed. The response vector in each cell is a frequency distribution of an underlying continuous response variable, categorised in k class intervals. The focus is to model the behavior of this grouped response variable over the T cells to evaluate the effect of the explanatory variables on the dependent variable. The basic formulation of the grouped response variable, to be modeled over the T cells of the multifactor design is summarised in Table 6.1.

Table 6.1: Grouped data in a multifactor design.

	Class interval				
Cells	$(-\infty, x_1)$	$[x_1, x_2)$	$\cdots$	$[x_{k-2}, x_{k-1})$	$[x_{k-1}, \infty)$
1	f_{11}	f_{12}	$\cdots$	$f_{1,k-1}$	f_{1k}
2	f_{21}	f_{22}	$\cdots$	$f_{2,k-1}$	f_{2k}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
T	f_{T1}	f_{T2}	$\cdots$	$f_{T,k-1}$	f_{Tk}

6.1 Formulation

Considering the frequencies tabulated in Table 6.1, let

$$\mathbf{F} = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1,k-1} \\ f_{21} & f_{22} & \cdots & f_{2,k-1} \\ \vdots & \vdots & \cdots & \vdots \\ f_{T1} & f_{T2} & \cdots & f_{T,k-1} \end{pmatrix} = \begin{pmatrix} \mathbf{f}_1' \\ \mathbf{f}_2' \\ \vdots \\ \mathbf{f}_T' \end{pmatrix} : T \times (k-1) \tag{6.1}$$

be the matrix where the rows of $\mathbf{F}$ denote the T cells of the multifactor design and the columns of $\mathbf{F}$ denote the first $(k-1)$ class intervals of the grouped response variable. Similarly to the estimation of distribution functions done in Part I, only the first $(k-1)$ class intervals need to be considered for each cell.

Define

$$\text{vec}\,(\mathbf{F}) = \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_T \end{pmatrix} : T\,(k-1) \times 1 \tag{6.2}$$

as the so-called concatenated frequency vector where the T **rows** of $\mathbf{F}$ in (6.1) are stacked row by row in a single column vector. The frequency vector for the t-th cell in (6.2) is

$$\mathbf{f}_t = \begin{pmatrix} f_{t1} \\ f_{t2} \\ \vdots \\ f_{t,k-1} \end{pmatrix} \qquad t = 1, 2 \cdots T \tag{6.3}$$

and consists of the first $(k-1)$ frequencies with corresponding vector of upper class boundaries

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{k-1} \end{pmatrix}. \tag{6.4}$$

Note: The definition of $\mathrm{vec}\,(\mathbf{F})$ (6.2) differs from the standard definition where the **columns** of $\mathbf{F}$ (6.1) are stacked as a single column vector. (See *Muirhead (1972) (p.17)*). However, by stacking the rows below each other coincides with the definition of the COLVEC function in SAS which is used extensively in this thesis for the computer programming of applications of grouped data in a multifactor design.

It is assumed that the vector $\mathbf{f}$ is a product multinomial vector with fixed subtotals

$$\mathbf{n} = \begin{pmatrix} n_1 \\ n_2 \\ \vdots \\ n_T \end{pmatrix} \tag{6.5}$$

allocated to each of the T cells.

Define

$$\mathbf{p}_0 = \begin{pmatrix} \mathbf{p}_{01} \\ \mathbf{p}_{02} \\ \vdots \\ \mathbf{p}_{0T} \end{pmatrix} = \begin{pmatrix} \frac{1}{n_1}\mathbf{f}_1 \\ \frac{1}{n_2}\mathbf{f}_2 \\ \vdots \\ \frac{1}{n_T}\mathbf{f}_T \end{pmatrix} = \left((\mathrm{diag}\,(\mathbf{n}))^{-1} \otimes \mathbf{I}_{k-1}\right) \cdot \mathbf{f} \tag{6.6}$$

as the concatenated vector of relative frequencies for the T cells. Hence, let

$$\mathrm{E}(\mathbf{p}_0) = \begin{pmatrix} \boldsymbol{\pi}_{01} \\ \boldsymbol{\pi}_{02} \\ \vdots \\ \boldsymbol{\pi}_{0T} \end{pmatrix} = \boldsymbol{\pi}_0$$

then

$$\mathrm{Cov}(\mathbf{p}_0) = \begin{pmatrix} \mathbf{V}_{01} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{02} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V}_{0T} \end{pmatrix} = \mathbf{V}_0 \tag{6.7}$$

where

$$\text{Cov}\left(\mathbf{p}_{0t}\right) = \frac{1}{n_t}\left(\text{diag}\left(\boldsymbol{\pi}_{0t}\right) - \boldsymbol{\pi}_{0t}\boldsymbol{\pi}_{0t}'\right) = \mathbf{V}_{0t} \qquad , \; t = 1, \cdots, T \tag{6.8}$$

is the covariance matrix for the vector of relative frequencies for the t-th cell.

Following (6.7) and (6.8) the covariance matrix of $\mathbf{p}_0$ may be expressed in terms of Kronecker products

$$\mathbf{V}_0 = \left\{\left(\text{diag}\left[\mathbf{n}\right]\right)^{-1} \otimes \mathbf{I}_{k-1}\right\} \cdot \left\{\text{diag}\left[\boldsymbol{\pi}_0\right] - \text{diag}\left[\boldsymbol{\pi}_0\right]\left(\mathbf{I}_T \otimes \left(\mathbf{1}_{k-1}\mathbf{1}_{k-1}'\right)\right)\text{diag}\left[\boldsymbol{\pi}_0\right]\right\} \tag{6.9}$$

where $\mathbf{1}_{k-1}$ is a $(k-1)$ vector of ones.

Define the concatenated vector of cumulative relative frequencies

$$\mathbf{p} = \begin{pmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_T \end{pmatrix} = \begin{pmatrix} \mathbf{C}\mathbf{p}_{01} \\ \mathbf{C}\mathbf{p}_{02} \\ \vdots \\ \mathbf{C}\mathbf{p}_{0T} \end{pmatrix} = \left(\mathbf{I}_T \otimes \mathbf{C}\right)\mathbf{p}_0 \tag{6.10}$$

where

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} : (k-1) \times (k-1) \; . \tag{6.11}$$

In (6.10) $\mathbf{p}_t = \mathbf{C}\mathbf{p}_{0t}$ for $t = 1, 2, \cdots, T$ is the cumulative relative frequency vector for the t-th cell in the multifactor design.

The random vector $\mathbf{p}$ consists of the cumulative relative frequencies from T independent multinomial populations, therefore let

$$\text{E}(\mathbf{p}) = \begin{pmatrix} \boldsymbol{\pi}_1 \\ \boldsymbol{\pi}_2 \\ \vdots \\ \boldsymbol{\pi}_T \end{pmatrix} = \boldsymbol{\pi} \tag{6.12}$$

where

$$\text{E}(\mathbf{p}_t) = \boldsymbol{\pi}_t \qquad , \; t = 1, \cdots, T$$

is the expected value for the vector of cumulative relative frequencies for the t-th cell and

$$\mathrm{Cov}(\mathbf{p}) = \begin{pmatrix} \mathbf{V}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{V}_T \end{pmatrix} = \mathbf{V} \tag{6.13}$$

where

$$\begin{aligned} \mathrm{Cov}(\mathbf{p}_t) &= \frac{1}{n_t}\left\{\mathbf{C}\,\mathrm{diag}\left(\mathbf{C}^{-1}\boldsymbol{\pi}_t\right)\mathbf{C}' - \boldsymbol{\pi}_t\boldsymbol{\pi}_t'\right\} \\ &= \mathbf{V}_t \qquad ,\ t = 1, \cdots, T \end{aligned} \tag{6.14}$$

is the covariance matrix for the vector of cumulative relative frequencies for the t-th cell.

From (6.10) it follows that the covariance matrix of $\mathbf{p}$ may also be expressed by

$$\mathbf{V} = (\mathbf{I}_T \otimes \mathbf{C})\,\mathbf{V}_0\,(\mathbf{I}_T \otimes \mathbf{C})' \tag{6.15}$$

where $\mathbf{V}_0$ is the covariance matrix of $\mathbf{p}_0$ in (6.9).

Note: For simplicity the class boundaries $\mathbf{x}$ are assumed to be constant over the different cells. The extension to the situation where this is not the case, can be done in a straight forward way.

6.2 Estimation

The ML estimation procedure entails that distribution fitting be done under the restriction that the cumulative relative frequencies equal the cumulative distribution curve at the upper class boundaries, for every cell in the multifactor design, i.e.

$$\begin{pmatrix} F_1(\mathbf{x}, \boldsymbol{\beta}_1) \\ F_2(\mathbf{x}, \boldsymbol{\beta}_2) \\ \vdots \\ F_T(\mathbf{x}, \boldsymbol{\beta}_T) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\pi}_1 \\ \boldsymbol{\pi}_2 \\ \vdots \\ \boldsymbol{\pi}_T \end{pmatrix} \tag{6.16}$$

with

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_T \end{pmatrix} \tag{6.17}$$

the concatenated vector of original parameters to be estimated.

Utilizing the ML estimation procedure, the vector of constraints to be imposed is

$$\mathbf{g}(\boldsymbol{\pi}) = \begin{pmatrix} F_1(\mathbf{x}, \boldsymbol{\beta}_1) \\ F_2(\mathbf{x}, \boldsymbol{\beta}_2) \\ \vdots \\ F_T(\mathbf{x}, \boldsymbol{\beta}_T) \end{pmatrix} - \begin{pmatrix} \boldsymbol{\pi}_1 \\ \boldsymbol{\pi}_2 \\ \vdots \\ \boldsymbol{\pi}_T \end{pmatrix} = \mathbf{0}. \tag{6.18}$$

In the case where (6.16) may be transformed into the linear model

$$\mathbf{h}(\boldsymbol{\pi}) = \begin{pmatrix} \mathbf{X}\boldsymbol{\alpha}_1 \\ \mathbf{X}\boldsymbol{\alpha}_2 \\ \vdots \\ \mathbf{X}\boldsymbol{\alpha}_2 \end{pmatrix} = (\mathbf{I}_T \otimes \mathbf{X})\,\boldsymbol{\alpha} \tag{6.19}$$

with

$$\boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_T \end{pmatrix} \tag{6.20}$$

a simultaneous distribution fitting for the T frequency distributions is outlined in the following three steps.

Step 1: The restricted ML estimate $\widehat{\boldsymbol{\pi}}$ is obtained by implementing the vector of constraints, $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$, with

$$\mathbf{g}(\boldsymbol{\pi}) = (\mathbf{I}_T \otimes \mathbf{Q}_X)\, \mathbf{h}(\boldsymbol{\pi}) \tag{6.21}$$

and matrix of partial derivatives

$$\mathbf{G}_\pi = (\mathbf{I}_T \otimes \mathbf{Q}_X)\, \mathbf{D}_\pi \tag{6.22}$$

where $\mathbf{Q}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{D}_\pi = \dfrac{\partial \mathbf{h}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}$ in the ML estimation process.

Step 2: The ML estimate of $\boldsymbol{\alpha}$ follows as

$$\widehat{\boldsymbol{\alpha}} = \left(\mathbf{I}_T \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\right) \mathbf{h}(\widehat{\boldsymbol{\pi}}) \tag{6.23}$$

with asymptotic covariance matrix

$$\mathrm{Cov}\,(\widehat{\boldsymbol{\alpha}}) \approxeq \left\{\mathbf{I}_T \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi\right\} \mathrm{Cov}\,(\widehat{\boldsymbol{\pi}}) \left\{\mathbf{I}_T \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi\right\}' \ . \tag{6.24}$$

Step 3: The ML estimates of the original parameters namely $\widehat{\boldsymbol{\beta}}$, are obtained from $\widehat{\boldsymbol{\alpha}}$ with

$$\mathrm{Cov}\left(\widehat{\boldsymbol{\beta}}\right) \approxeq \mathbf{B}\, \mathrm{Cov}\,(\widehat{\boldsymbol{\alpha}})\, \mathbf{B}' \tag{6.25}$$

where $\mathbf{B} = \dfrac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\alpha}}$. From the multivariate delta theorem, it follows that

$$\widehat{\boldsymbol{\beta}} \approxeq \mathrm{N}\left(\boldsymbol{\beta}, \mathbf{B}\, \mathrm{Cov}(\widehat{\boldsymbol{\alpha}})\mathbf{B}'\right) \ . \tag{6.26}$$

It follows from (6.23) that each of the T estimated distribution functions will have its own set of parameter estimates characterising the shape and locality of the distribution. Certain parameter structures may now be defined which may be incorporated to evaluate the effect of the factor(s) on the respons variable in any multiway design.