Investigation of Goodness-of-fit in Ecological Data Modeling

July 2016

MAYUMI NAKA

A Thesis for the Degree of Ph.D. in Science

Investigation of Goodness-of-fit in Ecological Data Modeling

July 2016

Graduate School of Science and Technology
Keio University

MAYUMI NAKA

Preface

How to find a new value from data becomes very crucial in science and technology. One of the driving forces is, of course, the continued development of information technology. To understand the phenomena behind data, statistical modeling plays an important role and a model of probability distributions is one of the basic statistical models. This thesis is to make a contribution to the development of such a modeling with a probability distribution model through practice and theory. A focus is on the goodness-of-fit of distributions. We first present two case studies and build distributional models by considering theoretical aspects of the data as well as their statistical characteristics. Then, several theoretical properties of the goodness-of-fit test statistic for contaminated data are shown, which are inspired by the two case studies.

The first case study is the modeling of the weight of animals on seabed, which is discussed in Naka et al. (2012). It is shown that the gamma distribution, which is derived as the equilibrium distribution of the stochastic growth model, can be used for modeling the weight by using an extended version of the Cramér-von Mises statistic for independent but not identically distributed observations. Then the effects of trawling are investigated by comparing the weight distribution after trawling and the gamma distribution with the parameters estimated from the observations before trawling. This case study is a joint research in 2009-2011 with Ross Darnell, Charis Burridge, and Mick Haywood in CSIRO (Commonwealth Scientific and Industrial Research Organisation) .

The second case study is the modeling of the carapace length of banana prawns, which is observed in the survey for the assessment of the effect of freshwater flows in an estuary. A probability distribution model obtained for the carapace length of banana prawns is an asymmetric mixture distribution, which is derived by combining a growth model with temperature and salinity of water and a survival rate model. This case study is a joint research in 2012-2013 with Ian Halliday in Department of Employment, Economic Development and Innovation, Australia and Ross Darnell in CSIRO.

Theoretical results on the asymptotic behavior of the Cramér-von Mises goodness-of-fit test statistic for contaminated data are given to investigate the robustness of the statistic, which is studied in Naka and Shibata (2016). The asymptotic distributions of the Cramér-von Mises statistic for contaminated data are derived when parameters are known and when parameters are estimated by the minimum Cramér-von Mises distance method. The theoretical results together with the result of numerical experiments show that the Cramér-von Mises statistic is robust when the minimum distance estimator is used for the estimation of parameters.

Contents

Chapter 1

Introduction

A model of probability distributions is one of the basic statistical models to understand the phenomena behind data. Although it is not always the case, the probability distribution model plays an important role in analyzing data. The probability distribution model can describe stochastic mechanisms in the phenomena and the model is simple so that it is easy to understand from the model how the model describes the mechanisms. However, the probability distribution model has to be used with caution because it may lead us a wrong direction without examining the goodness-of-fit. In this chapter, we give a brief introduction of the methods to examine the goodness-of-fit of a distribution.

We assume in this chapter that $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with a distribution function $F(x)$ which is usually unknown. We denote $F(x, \boldsymbol{\theta})$ to be a distribution function of a model for the observations with a parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^\top \in \Theta \subset \mathbb{R}^m$.

We use hereby the goodness-of-fit of a distribution in view point of examining whether a model $F(x, \boldsymbol{\theta})$ fitted to the data gives us a reasonable approximation to the underlying distribution $F(x)$ and can be used for any further investigation of the phenomena. This usage is a little different from a general way of thinking for goodness-of-fit tests. The goodness-of-fit test is a statistical decision whether $F(x) = F(x, \boldsymbol{\theta})$ or not. However, there are many cases where such a decision is not real concern. Rather the main concern is often whether the fitted model can be used for further investigation. Therefore, the role of examining the goodness-of-fit would be to exclude cases where the use of the model will lead us an incorrect result. Otherwise we can continue the analysis based on the probability distribution model since the value of the analysis is determined not by such a statistical decision but by how persuasive the final result is. The two case studies presented in Chapter 2 and Chapter 3 are such cases. The aim of analysis in the case studies is to find out the effect of the environmental changes on animals on seabed or banana prawn. The probability distribution model is a main tool for such findings. Validation of the goodness-of-fit of the model plays an important role but not a goal.

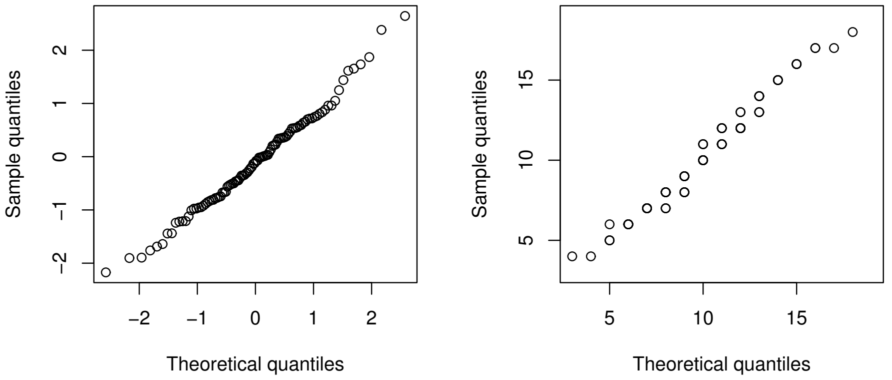
In the following, we first introduce two basic plots as graphical methods for checking goodness-of-fit and then we give a brief summary of goodness-of-fit tests for both continuous and discrete distributions.

1.1 Graphical methods for checking goodness-of-fit

Graphical methods for checking goodness-of-fit is useful at an early stage of building a probability distribution model. It assists us in finding out the discrepancy between the observations and the model. Let $x_1, x_2, \ldots, x_n$ be observations from distribution $F(x)$. One of the graphical methods for checking goodness-of-fit is a quantile-quantile or Q-Q plot (Chambers et al., 1983). A Q-Q plot is a plot of the n points

$$\left(F^{-1}\left(\frac{j-0.5}{n}, \boldsymbol{\theta}\right), x_{(j)}\right), \quad j = 1, 2, \ldots, n,$$

where $F^{-1}(\cdot, \boldsymbol{\theta})$ is the inverse function of $F(\cdot, \boldsymbol{\theta})$ and $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ are order statistics of the observations. Two examples of Q-Q plots are shown in Figure 1.1. Figure 1.1 (a) is a Q-Q plot for the standard normal distribution of a simulated random sample with size 100 from the standard normal distribution, and Figure 1.1 (b) is that for the Poisson distribution with mean 10 of a simulated random sample with size 100 from the Poisson distribution with mean 10.



(a) The standard normal distribution. (b) The Poisson distribution with mean 10.

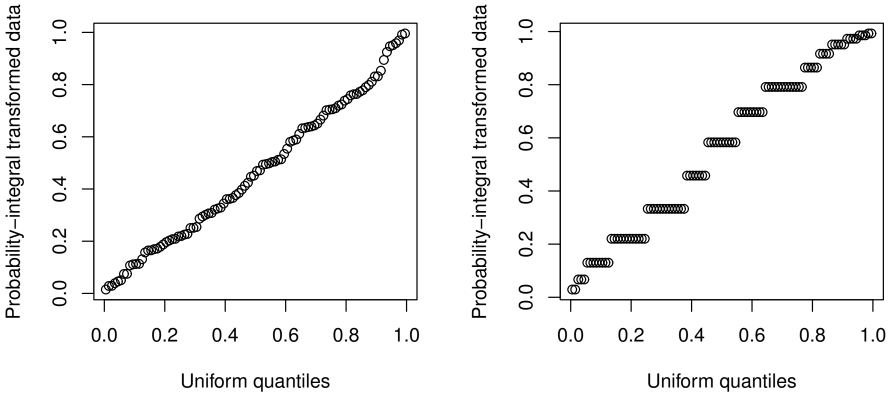
Figure 1.1: Examples of Q-Q plots.

If the observations follow the distribution $F(x, \boldsymbol{\theta})$, the points would tend to follow the line $y = x$. If $F(x, \boldsymbol{\theta})$ is a distribution function with scale and location parameters, it is enough to obtain $F^{-1}((j-0.5)/n, \boldsymbol{\theta})$, $j = 1, 2, \ldots, n$, with any values of the parameters and see whether the points follow the line $y = ax + b$ or not. This is an advantage of Q-Q plots. As seen from Figure 1.1 (b), the points are overlapped because the distribution function is not continuous so that a Q-Q plot is not suitable for discrete distributions.

A similar graphical method is a probability-probability or P-P plot (Gan and Koehler, 1990, Holmgren, 1995), which is also called the “Universal Q-Q plot” by Luceño (2007). A P-P plot is a plot of the n points

$$\left(\frac{j-0.5}{n}, F(x_{(j)}, \boldsymbol{\theta}) \right), \quad j = 1, 2, \ldots, n.$$

P-P plots for the same data as in Figure 1.1 are given in Figure 1.2. It can be compared from two figures how Q-Q plots and P-P plots look different for the same observations. As same as a Q-Q plot, the points on a P-P plot would tend to follow the line $y = x$ if the observations follow the distribution $F(x, \boldsymbol{\theta})$. Also a P-P plot is not suitable for discrete distributions because it is not clear whether the



(a) The standard normal distribution. (b) The Poisson distribution with mean 10.

Figure 1.2: Examples of P-P plots.

points follow the line $y = x$ or not since the distribution function is discontinuous. An advantage of P-P plots is that it is applicable to independent but not identically distributed observations, which we will deal with in Chapter 2.

1.2 Goodness-of-fit test statistics for continuous distributions

Goodness-of-fit test is a statistical test to examine whether a model fits well to the observations. If we focus on the goodness-of-fit of a distribution, the aim of the test is to test the null hypothesis $\mathrm{H}_0 : \ F(x) = F(x, \boldsymbol{\theta})$, where the observations are from a distribution function $F(x)$ and $F(x, \boldsymbol{\theta})$ is a distribution function of a parametric continuous distribution.

Although there are many tests for specific distributions, such as the Shapiro-Wilk test for the normal distribution (Shapiro and Wilk, 1965), here we consider goodness-of-fit tests based on the empirical distribution function. Let

$F_n(x)$ be the empirical distribution

$$F_n(x) = \sum_{j=1}^{n} 1_{X_j \le x},$$

where

$$1_{X_j \le x} = \begin{cases} 1 & X_j \le x \\ 0 & X_j > x \end{cases} .$$

Goodness-of-fit test statistics based on the empirical distribution function are defined as a distance between $F_n(x)$ and $F(x, \boldsymbol{\theta})$. In the following, we will introduce some basic test statistics.

The Kolmogorov-Smirnov type statistic is a statistic based on the supremum distance between $F_n(x)$ and $F(x, \boldsymbol{\theta})$. The well-known form of the Kolmogorov-Smirnov statistic is

$$D_n(\boldsymbol{\theta}) = \sup_{-\infty \le x \le \infty} \sqrt{n} \left| F_n(x) - F(x, \boldsymbol{\theta}) \right|$$

and one-sided versions are

$$D_n^+(\boldsymbol{\theta}) = \sup_{-\infty \le x \le \infty} \sqrt{n} \left\{ F_n(x) - F(x, \boldsymbol{\theta}) \right\}, \quad D_n^-(\boldsymbol{\theta}) = \sup_{-\infty \le x \le \infty} \sqrt{n} \left\{ F(x, \boldsymbol{\theta}) - F_n(x) \right\}.$$

These statistics would be the most often used for testing goodness-of-fit and have been investigated their properties by many researchers. It can be seen from the definitions that these statistics are basically computed from one observation.

On the other hand, the Cramér-von Mises type statistic is a statistic based on the L^2 norm between $F_n(x)$ and $F(x, \boldsymbol{\theta})$ such as

$$n \int_{-\infty}^{\infty} \left\{ F_n(x) - F(x, \boldsymbol{\theta}) \right\}^2 g(x) dx$$

with a weight function $g(x)$. In contrast to the Kolmogorov-Smirnov type statistic, this test statistic is computed from all observations. One of the well-known this type of statistics is the Cramér-von Mises statistic, which is defined as

$$W_n^2(\boldsymbol{\theta}) = n \int_{-\infty}^{\infty} \left\{ F_n(x) - F(x, \boldsymbol{\theta}) \right\}^2 dF(x, \boldsymbol{\theta}). \tag{1.1}$$

Another special statistic is the Anderson-Darling statistic, which is defined as

$$A_n^2(\boldsymbol{\theta}) = n\int_{-\infty}^{\infty} \frac{\{F_n(x) - F(x,\boldsymbol{\theta})\}^2}{F(x,\boldsymbol{\theta})\{1-F(x,\boldsymbol{\theta})\}} dF(x,\boldsymbol{\theta}).$$

Since $n\mathrm{E}\{F_n(x) - F(x,\boldsymbol{\theta})\}^2 = F(x,\boldsymbol{\theta})\{1-F(x,\boldsymbol{\theta})\}$ if the observations are from distribution $F(x,\boldsymbol{\theta})$, the weight $g(x)$ used in this statistic is the reciprocal of the variance. For these test statistics, details are in Durbin (1973).

One of the problems in using these test statistics is that the asymptotic distributions of the statistics depend on the model $F(x,\boldsymbol{\theta})$ in case that the parameters are unknown and estimated from a sample, while the distributions of the test statistics do not depend on $F(x,\boldsymbol{\theta})$, that is, the tests are distribution-free, in case that the parameters are known. To overcome this problem, various methods have been proposed. For example, Khmaladze (1981) showed that the process $\sqrt{n}\{F_n(x) - K(x,F_n(x),\boldsymbol{\theta})\}$ converges to the standard wiener process so that distribution-free test is possible by using the following transformation $K(x,F_n(x),\boldsymbol{\theta})$, known as "Khmaladze transform":

$$K(x,F_n(x),\boldsymbol{\theta}) = \int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\min(x,y)} q(z,\boldsymbol{\theta})^{\top} C^{-1}(z,\boldsymbol{\theta}) dF(z,\boldsymbol{\theta})\right\} q(y,\boldsymbol{\theta}) dF_n(y),$$

where

$$q(x,\boldsymbol{\theta})^{\top} = \left(1, \frac{\partial}{\partial\boldsymbol{\theta}} \log f(x,\boldsymbol{\theta})\right),$$

$$C(z,\boldsymbol{\theta}) = \int_{z}^{\infty} q(x,\boldsymbol{\theta}) q(x,\boldsymbol{\theta})^{\top} f(x,\boldsymbol{\theta}) dx,$$

and $f(x,\boldsymbol{\theta})$ is the probability density function of $F(x,\boldsymbol{\theta})$. Applications of the transformation for testing exponentiality are given in Khmaladze et al. (2007) and Haywood and Khmaladze (2008).

Although many variations of the goodness-of-fit test statistics have been introduced, we focus on the Cramér-von Mises statistic in this thesis. This is because the Cramér-von Mises statistic is simple and asymptotically equal to the sum of the squared distances between points and a line $y = x$ on a P-P plot so that it is easy to understand. The relation between the statistic and the P-P plot can be seen from the fact that the Cramér-von Mises statistic can be expressed as

$$W_n^2(\boldsymbol{\theta}) = \sum_{j=1}^{n} \left\{ F(X_{(j)}, \boldsymbol{\theta}) - \frac{j-0.5}{n} \right\}^2 + \frac{1}{12n}, \tag{1.2}$$

where $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ are order statistics of the random variables. The equivalence of the two representations (1.1) and (1.2) can be shown as follows. From (1.1) and the definition of the empirical distribution function, we have

$$\begin{aligned} W_n^2(\boldsymbol{\theta}) = & \int_{-\infty}^{\infty} \left\{ \frac{1}{n} \left(\sum_{j=1}^{n} 1_{\, X_j \leq x} \right)^2 - 2 \left(\sum_{j=1}^{n} 1_{\, X_j \leq x} \right) F(x, \boldsymbol{\theta}) \right\} dF(x, \boldsymbol{\theta}) \\ & + n \int_{-\infty}^{\infty} F^2(x, \boldsymbol{\theta}) dF(x, \boldsymbol{\theta}). \end{aligned}$$

Since the first term on the right hand side is equal to

$$\begin{aligned} & \sum_{j=1}^{n} \int_{X_j}^{\infty} \left\{ \frac{1}{n} - 2F(x, \boldsymbol{\theta}) \right\} dF(x, \boldsymbol{\theta}) + \frac{2}{n} \sum_{j=2}^{n} (j-1) \int_{X_{(j)}}^{\infty} 1 dF(x, \boldsymbol{\theta}) \\ = & \sum_{j=1}^{n} \left[\frac{1}{n} - \frac{F\left(X_j, \boldsymbol{\theta}\right)}{n} - \left\{ 1 - F^2\left(X_j, \boldsymbol{\theta}\right) \right\} \right] + \frac{2}{n} \sum_{j=2}^{n} (j-1) \left\{ 1 - F\left(X_{(j)}, \boldsymbol{\theta}\right) \right\} \\ = & \sum_{j=1}^{n} \left\{ F^2\left(X_{(j)}, \boldsymbol{\theta}\right) - \frac{2(j-0.5) F\left(X_{(j)}, \boldsymbol{\theta}\right)}{n} \right\} \\ = & \sum_{j=1}^{n} \left\{ F\left(X_{(j)}, \boldsymbol{\theta}\right) - \frac{j-0.5}{n} \right\}^2 - \frac{n}{3} + \frac{1}{12n} \end{aligned}$$

and the second term is equal to $n \int_0^1 t^2 dt = n/3$, the representation (1.2) is derived.

Asymptotic distribution of the Cramér-von Mises statistic

For the goodness-of-fit test, the distribution of the test statistic is essential to calculate the p-value. If $F(x, \boldsymbol{\theta})$ is a distribution function of a continuous distribution, it is known that the asymptotic distribution of the Cramér-von Mises statistic is given by a distribution of a weighted sum of chi-squared random variables with 1 degree of freedom, such that

$$\sum_{j=1}^{\infty} \lambda_j Z_j^2,$$

where $Z_j,\ j = 1,2,\ldots,$ follows the standard normal distribution, regardless of parameters being known or estimated (Darling, 1955, Shorack and Wellner, 1986).

When the parameters are known, the weights for the chi-squared random variables are given as the eigenvalues of the integral equation

$$\lambda f(u) = \int_0^1 \rho_0(u,v) f(v) dv$$

with the kernel function

$$\rho_0(u,v) = \min(u,v) - uv.$$

Here λ is an eigenvalue of the integral equation and $f(u)$ is an eigenfunction corresponding to λ. It can be seen from the kernel function that the asymptotic distribution of the statistic does not depend on the model $F(x,\boldsymbol{\theta})$.

However, the asymptotic distribution depends on the model $F(x,\boldsymbol{\theta})$ and the estimator if the parameters are necessary to be estimated from observations. Let $\tilde{\boldsymbol{\theta}}$ be an estimator of the parameter $\boldsymbol{\theta}$ in the model $F(x,\boldsymbol{\theta})$. Then under some regularity conditions, the asymptotic distribution is given as a distribution of a weighted sum of chi-squared random variables with 1 degree of freedom, where the weights are eigenvalues of the integral equation with the kernel function

$$\rho(u,v) = \rho_0(u,v) - \boldsymbol{g}(u,\boldsymbol{\theta})^\top \boldsymbol{h}(v) - \boldsymbol{h}(u)^\top \boldsymbol{g}(v,\boldsymbol{\theta}) + \boldsymbol{g}(u,\boldsymbol{\theta})^\top \Sigma\, \boldsymbol{g}(v,\boldsymbol{\theta}). \tag{1.3}$$

Here

$$\boldsymbol{g}(u,\boldsymbol{\theta}) = \left(\frac{\partial}{\partial \theta_j} F(x,\boldsymbol{\theta}) \bigg|_{x=F^{-1}(u,\boldsymbol{\theta})} ;\ 1 \le j \le m \right),$$

$$\boldsymbol{h}(u) = \lim_{n\to\infty} \mathrm{E}\left[\sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^{n} 1_{F(X_i,\boldsymbol{\theta})<u} - u \right\} \left\{ \sqrt{n} \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta} \right) \right\} \right],$$

and

$$\Sigma = \lim_{n\to\infty} n \, \mathrm{E}\left\{ \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^{\top} \right\}.$$

For a general estimator, no simpler form of the function (1.3) is obtained for other than the maximum likelihood estimator. If the maximum likelihood estimator is employed, (1.3) becomes

$$\rho(u,v) = \rho_0(u,v) - \boldsymbol{g}(u,\boldsymbol{\theta})^{\top} I(\boldsymbol{\theta})^{-1} \boldsymbol{g}(v,\boldsymbol{\theta}),$$

where $I(\boldsymbol{\theta})$ is the Fisher information matrix. Since the maximum likelihood estimator is widely used and has a simple form of $\rho(u,v)$, much works have been done for the Cramér-von Mises statistic when the parameters are estimated by the maximum likelihood method. It is shown in Sukhatme (1972) that the asymptotic distribution does not depend on the unknown parameters for location-scale family, that is, for parametric distribution family with location and scale parameters. Moreover, Martynov (2010) showed that the asymptotic distribution does not depend on the unknown parameters for parametric distribution family with power and scale parameters, for example the Weibull and the Pareto distributions. The critical points obtained for various significance levels are tabulated, for example the normal, the gamma, and the logistic distributions in D'Agostino and Stephens (1986), the exponential distribution with scale and location parameters in Spinelli and Stephens (1987), the Weibull distribution in Lockhart and Stephens (1994), the Laplace distribution in Puig and Stephens (2000), the generalized Pareto distribution in Choulakian and Stephens (2001), and the hyperbolic distribution in Puig and Stephens (2001). An R package "fgof" for calculating goodness-of-fit test statistics and p-values for some distributions by a fast weighted bootstrap is developed by Kojadinovic and Yan (2012).

For the minimum distance estimator, which is an estimator of the parameters chosen to minimize the Cramér-von Mises statistic, there are some results for location-scale family of distributions. This is because the asymptotic distribution is independent of the unknown parameters. In fact, when the minimum distance estimator is used, Boos (1981) showed a representation of the asymptotic distribution for location-scale family. Koul and DeWet (1983) gave a similar method of the evaluation in the case of regression. However, their results are still not simple enough for calculation, and the generalization over location-scale family does not seem to be easy. In Chapter 4, we focus on the minimum distance estimator and describe a practical procedure for obtaining the asymptotic distribution of the statistic by a new approach.

1.3 Goodness-of-fit test statistics for discrete distributions

Let $x_1, x_2, \ldots, x_n$ be independent and identically distributed observations following a discrete distribution with K cells labeled $1, 2, \ldots, K$ and with probability p_j of falling into cell j, $j = 1, 2, \ldots, K$. We wish to check $p_j = p_j(\boldsymbol{\theta})$ with a parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^\top \in \Theta \subset \mathbb{R}^m$. Let o_j be the observed number of the observation and $e_j = np_j(\boldsymbol{\theta})$ be the expected number in cell j, $j = 1, 2, \ldots, K$.

One of the most used goodness-of-fit tests for discrete distributions would be Pearson's chi-squared test and the test statistic is given as

$$\chi^2(\boldsymbol{\theta}) = \sum_{j=1}^{K} \frac{\{o_j - e_j(\boldsymbol{\theta})\}^2}{e_j(\boldsymbol{\theta})}.$$

If the parameters are estimated properly, then the asymptotic distribution of the test statistic is distribution-free, that is, the test statistic does not depend on the probabilities $p_j(\boldsymbol{\theta})$, $j = 1, 2, \ldots$. Pearson's chi-squared test is also widely used for data other than ordered or not ordered categorical data. For example, if the observations follow a discrete distribution taking nonnegative integer, such as the Poisson distribution and the geometric distribution, it is enough to sum up the tail part to do the test. Also the test can be used for rounded or grouped data, which occurs often in practice.

One of the major disadvantages of Pearson's chi-squared test is its sensitivity to cell selection. If the number of the cells is countable and infinite, for example in case of the Poisson distribution, there are some choices of values that are included into the last cell. In addition, it is required to be enough observations in each cell because the validation of the test is shown as a limiting result. For example, it is noted in Cochran (1954) that the number of observations in each cell should not be less than 5. The cell selection is required to satisfy these needs, however, the result of the test tends to be changed by the selection. A lot of goodness-of-fit tests are introduced to overcome this disadvantage.

The tests based on the empirical distribution function can be applied for cases other than non-ordered categorical data. Let $c_1, c_2, \ldots, c_K$ be the increasing values corresponded for each cell $1, 2, \ldots, K$. Then the distribution function $F(c_k, \boldsymbol{\theta})$ and the empirical distribution function $F_n(c_k)$ can be defined as

$$F(c_k, \boldsymbol{\theta}) = \frac{1}{n} \sum_{j=1}^{k} e_j(\boldsymbol{\theta}), \quad F_n(c_k) = \frac{1}{n} \sum_{j=1}^{k} o_j$$

for $k = 1, 2, \ldots, K$, respectively, so that the test statistics based on the empirical distribution function, such as the Kolmogorov-Smirnov and the Cramér-von Mises type statistics, are derived directly for discrete distributions. It is often said that the tests based on the empirical distribution functions are more powerful than Pearson's chi-squared test because they concern the order of the cells while Pearson's chi-squared test does not.

The Kolmogorov-Smirnov statistic for discrete distributions has been studied for long time as same as for continuous distributions. The discrete distribution version of the Kolmogorov-Smirnov statistic $D_n(\boldsymbol{\theta})$ is

$$D_n^{(d)}(\boldsymbol{\theta}) = \sup_{1 \le k \le K} \sqrt{n} \left| F_n(c_k) - F(c_k, \boldsymbol{\theta}) \right| = \sup_{1 \le k \le K} \frac{1}{\sqrt{n}} \left| \sum_{j=1}^{k} o_j - \sum_{j=1}^{k} e_j(\boldsymbol{\theta}) \right|.$$

Conover (1972) gave a method of finding the critical value for the Kolmogorov-Smirnov test for discrete distributions. Horn (1977) compared 5 goodness-of-fit tests for discrete distributions, including Pearson's chi-squared test and the Kolmogorov-Smirnov test, and suggested the Kolmogorov-Smirnov test for small sample size and ordered categorical data. The asymptotic distribution of the statistic is derived by Wood and Altavela (1978) when the parameters are known.

The Cramér-von Mises statistic for discrete distributions is also investigated. The discrete distribution version of the Cramér-von Mises statistic $W_n^2(\boldsymbol{\theta})$ is

$$\begin{aligned} W_n^{(d)2}(\boldsymbol{\theta}) &= n\sum_{k=1}^{K}\{F_n(c_k)-F(c_k,\boldsymbol{\theta})\}^2\{F(c_k,\boldsymbol{\theta})-F(c_{k-1},\boldsymbol{\theta})\} \\ &= \frac{1}{n}\sum_{k=1}^{K}\left\{\sum_{j=1}^{k}o_j-\sum_{j=1}^{k}e_j(\boldsymbol{\theta})\right\}^2 p_k(\boldsymbol{\theta}). \end{aligned} \qquad (1.4)$$

Choulakian et al. (1994) introduced $W_n^{(d)2}(\boldsymbol{\theta})$ as well as the Anderson-Darling type

$$A_n^{(d)2}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{k=1}^{K}\frac{\left\{\sum_{j=1}^{k}o_j-\sum_{j=1}^{k}e_j(\boldsymbol{\theta})\right\}^2 p_k(\boldsymbol{\theta})}{H_k(1-H_k)},$$

where $H_k = \sum_{j=1}^{k}e_j(\boldsymbol{\theta})/n$, and gave tables for tests for the discrete uniform distribution. Spinelli and Stephens (1997) developed their result to the Poisson distribution when mean parameter is unknown. Spinelli (2001) gave slightly different definitions for these statistics and showed that the statistics give powerful tests for exponentiality with grouped data. The asymptotic distributions of the new version of the statistics when the parameters are estimated by the maximum likelihood method are given by Lockhart et al. (2007). They also showed from Monte Carlo simulations that the percentage points converge to the asymptotic points quickly and these tests are more powerful than Pearson's chi-squared test when the probabilities in the cells are in a steadily increasing pattern.

There are many studies comparing various tests and introducing new goodness-of-fit tests. For testing the Poisson distribution, extensive comparisons among a variety of tests as well as simulation results for power studies are given by Gürtler and Henze (2000) and Karlis and Xekalaki (2000). Power studies for the uniform null in 10 cells are given by Steele and Chaseling (2006). As an example of the new goodness-of-fit tests, Székely and Rizzo (2004) proposed M-test for the Poisson distribution, which is based on a characterization by mean distances.

Asymptotic distribution of the Cramér-von Mises statistic for discrete distributions

We first note that the other version of the Cramér-von Mises statistic for discrete distributions introduced by Spinelli (2001) is

$$\frac{1}{n}\sum_{k=1}^{K}\left\{\sum_{j=1}^{k} o_j - \sum_{j=1}^{k} e_j(\boldsymbol{\theta})\right\}^2 \frac{p_k(\boldsymbol{\theta}) + p_{k+1}(\boldsymbol{\theta})}{2}, \tag{1.5}$$

where $p_{K+1}(\boldsymbol{\theta}) = p_1(\boldsymbol{\theta})$. It is explained that the reason for this modification is that the distribution of the test statistic is identical for a new random variable $Y = -X$ for a negative exponential distribution, so that the test becomes symmetric. However, in this thesis we focus on the statistic $W_n^{(d)2}(\boldsymbol{\theta})$ because we apply the statistic to truncated data, which is assumed to follow a continuous distribution before the truncation, in the second case study presented in Chapter 3 and it is not necessary to be the test symmetric in that situation.

The asymptotic distribution of the statistic $W_n^{(d)2}(\boldsymbol{\theta})$ is given as a distribution of a weighted sum of chi-squared random variables with 1 degree of freedom, as same as the statistic $W_n^2(\boldsymbol{\theta})$ for continuous distributions (Choulakian et al., 1994). Let

$$\boldsymbol{y} = \left(\frac{1}{\sqrt{n}}\left(\sum_{j=1}^{k} o_j - \sum_{j=1}^{k} e_j(\boldsymbol{\theta})\right); 1 \leq k \leq K\right)$$

and $\Sigma_y = \mathrm{E}\left(\boldsymbol{y}\boldsymbol{y}^\top\right)$. Then we have

$$W_n^{(d)2}(\boldsymbol{\theta}) = \boldsymbol{y}^\top P(\boldsymbol{\theta})\boldsymbol{y} = \left(\Sigma_y^{-\frac{1}{2}}\boldsymbol{y}\right)^\top \Sigma_y^{\frac{1}{2}} P(\boldsymbol{\theta}) \Sigma_y^{\frac{1}{2}} \left(\Sigma_y^{-\frac{1}{2}}\boldsymbol{y}\right),$$

where $P(\boldsymbol{\theta}) = \mathrm{diag}(\boldsymbol{p}(\boldsymbol{\theta}))$ and $\boldsymbol{p}(\boldsymbol{\theta}) = (p_1(\boldsymbol{\theta}), p_2(\boldsymbol{\theta}), \ldots, p_K(\boldsymbol{\theta}))^\top$. The distribution of $\Sigma_y^{-\frac{1}{2}}\boldsymbol{y}$ converges to the multivariate normal distribution with mean $\mathbf{0}$ and variance I, so that the asymptotic distribution of $W_n^{(d)2}(\boldsymbol{\theta})$ is given by

$$\sum_{j=1}^{K-1} \lambda_j V_j^2,$$

where V_j follows the standard normal distribution and λ_j is an eigenvalue of $\Sigma_y^{\frac{1}{2}} P(\boldsymbol{\theta}) \Sigma_y^{\frac{1}{2}}$, $j = 1, 2, \ldots, K-1$.

When the parameters are estimated by the maximum likelihood method, Lockhart et al. (2007) showed that the asymptotic distribution is also given as a distribution of a weighted sum of chi-squared random variables with 1 degree of freedom. Although their result is for the statistic (1.5), it can be easily modified for $W_n^{(d)2}(\boldsymbol{\theta})$. This is because the statistic (1.5) can be written as $\boldsymbol{y}^\top P'(\boldsymbol{\theta})\boldsymbol{y}$, where

$$P'(\boldsymbol{\theta}) = \mathrm{diag}\left(\frac{p_1(\boldsymbol{\theta}) + p_2(\boldsymbol{\theta})}{2}, \frac{p_2(\boldsymbol{\theta}) + p_3(\boldsymbol{\theta})}{2}, \ldots, \frac{p_K(\boldsymbol{\theta}) + p_{K+1}(\boldsymbol{\theta})}{2} \right).$$

Applying their result to $W_n^{(d)2}(\boldsymbol{\theta})$, the weights for the chi-squared random variables are given by the eigenvalues of $\Sigma_{\tilde{y}} P(\boldsymbol{\theta})$, where

$$\Sigma_{\tilde{y}} = \Sigma_y P(\boldsymbol{\theta}) - AZ(\boldsymbol{\theta}) \left\{ Z(\boldsymbol{\theta})^\top P(\boldsymbol{\theta}) Z(\boldsymbol{\theta}) \right\}^{-1} Z(\boldsymbol{\theta})^\top A^\top P(\boldsymbol{\theta})$$

with a $K \times m$ matrix

$$Z(\boldsymbol{\theta}) = \left(\frac{\partial}{\partial \theta_k} p_j(\boldsymbol{\theta}); 1 \le j \le K, 1 \le k \le m \right)$$

and a $K \times K$ matrix

$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}.$$

As a reference, we note another approach to deriving the asymptotic distribution of the Cramér-von Mises statistic for discrete distributions. The asymptotic distribution can be derived by using the fact that the empirical process $\sqrt{n}\{F_n(x) - F(x, \boldsymbol{\theta})\}$ converges to a Gaussian process even for discrete distributions. In case of known parameters, the convergence is shown in Theorem 16.4 in Billingsley (1968) and Wood and Altavela (1978) gave the asymptotic distribution of the Kolmogorov-Smirnov test statistic for discrete distributions by using this approach. The convergence of the empirical process when the parameters are estimated is shown by Burke et al. (1979). Henze (1996) extended the results to the case under triangular arrays to establish the validity of parametric bootstrap and also gave simulation results for the Poisson and the geometric distribution.

Chapter 2

Trawling effect on the weight of animals on seabed

In this chapter, we discuss on the modeling the weight of animals on seabed as the first case study. The trawling effect on the weight of animals on seabed is verified by using the derived probability distribution model.

2.1 Introduction

Effects of various methods of harvesting the sea have been investigated in many articles on marine ecology. Collie et al. (2000) carried out a meta-analysis of 39 published fishing impact studies to draw general conclusions. Bishop et al. (2000) investigated the impact of technology on vessel performance in a trawl fishery during 1988-96 by using a generalized estimating equation. Burridge et al. (2003) investigated the trawl-depletion rate for benthic fauna in an area closed to commercial trawling.

In this chapter we investigate the effect of trawling through changes of weight distribution of animals on seabed, which is modeled by the equilibrium distribution of a stochastic growth model. The stochastic growth model is frequently used for modeling population size (Russo et al., 2009) or size of plants (Rupšys, 2007) or animals (Tovar-Ávila et al., 2009). We show that the gamma distribution, which is the equilibrium distribution of the stochastic growth model, is useful for modeling the weight distribution when no effective ecological disturbance exists. This result allows us to detect any effective disturbance by departure from the gamma distribution with the parameters estimated from the sample before the disturbance. The reason why we focus on individual weights of animals as an index of disturbance in this analysis is that it is sensitive to any ecological disturbances and easy to measure compared to their size. An advantage of our approach is that it makes possible to draw a whole picture of the current status of each species on seabed before and after trawling without introducing any particular estimating equation or indexes, such as Shannon's index and Simpson's index for biodiversity (Kaiser and Spencer, 1996).

The data used in this analysis were obtained in the project "Quantifying the effects of trawling on seabed fauna in the Northern Prawn Fishery" by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in Australia, which will be explained in detail in Section 2.2. Analyses of the data were already reported in Haywood et al. (2005). Together with calculating various fundamental statistics and drawing many graphs and maps, they tried finding out the effect of trawling by an application of a simple depletion and recovery model, however, it does not seem successful enough. There are several reasons why their analysis was not successful enough. One is that it is a class by class analysis, using popular descriptive statistics and plots, however, class by class analysis seems to be too coarse to verify any effect of ecological disturbances from our preliminary analysis. Instead we verify such disturbances by species by species analysis. Another reason is that their analysis is based on the whole weight of each species caught, normalized by the dredge area. The biomass density is a useful abundance measure for each survey area from the view point of fishery but not so for the detection of ecological disturbances. Changes of individual weights would be more useful for detecting ecological disturbance. For these reasons we examine the change of the weight in this analysis by fitting the model for individual weights of animals on seabed.

The model for individual weights is derived as we described above, however, the observations are only the total weight and the number of catches for each species. Therefore, we have to deal with not identically distributed variables. It does not cause any serious problem in parameter estimation but requires a modification of the goodness-of-fit test statistic because most of the tests are proposed for independent and identically distributed observations. We evaluate p-values of the extended version of the Cramér-von Mises statistic by executing computer simulations around the maximum likelihood estimate since even the asymptotic distribution is unknown.

This chapter is organized as follows. In Section 2.2, we give descriptions of the data. The stochastic growth model is introduced in Section 2.3 and the gamma distribution is derived as the equilibrium distribution for the weight. The models for the distributions of individual animal weights for each cases are determined by the gamma distributions with parameters estimated from the samples observed before trawling in Section 2.4. The effect of trawling is investigated through changes of the weight distribution in Section 2.5. Comparisons of the methods to investigate the effect of trawling between our approach and using simple mean tests, Welch's t-test and Student's t-test, are given in Section 2.6.

2.2 Seabed fauna data in Northern Prawn Fishery

The data are from the Fisheries Research and Development Corporation (FRDC) funded Project 2002/102, “Quantifying the effects of trawling on seabed fauna in the Northern Prawn Fishery” (NPF) in Australia. The project was originally identified as a high priority research area by the Northern Prawn Fishery Management Advisory Committee (NORMAC) because under the Environmental Protection and Biodiversity Conservation Act (EPBC Act), Australian fisheries are required to demonstrate their environmental sustainability. Industry offered special funding to support the research and the FRDC was asked to manage the project. CSIRO agreed to carry out the work, develop the scope of the work and the experimental design, and contribute to the funding.

Trawlers in the NPF tend to concentrate their fishing on areas of higher prawn density. Also intensive trawling of small areas is a feature of the tiger and endeavor prawn fishery. Around 20% of the catch is prawns but the rest are other animals collected from the seabed. The names of such by-catch animals are shown in Table A.1 in the Appendix A, although those are caught not by trawling but by experimental dredges.

Table 2.1: Part of the data.

Region	Plot	Treatment	Time	Scientific name	Count	Weight (g)
East	12	4	Before	*Retiflustra cornea*	1	0.25
East	12	4	Before	*Melaxinaea vitrea*	1	9.16
East	12	4	Before	Tubeworm OPNO 006	14	2.28
East	12	4	Before	Neritidae OPNO 142	0	0
East	12	4	Before	*Leucosia whitei*	1	1.49

The data consist of 207,726 records obtained by the experimental dredge survey and Table 2.1 shows a part of the data, which is for an explanation of the data structure. The first column labeled “Region” indicates the region where the experiment was performed. The survey area shown in Figure 2.1 is roughly divided into two regions, East (East of Mornington Island) and West (West of Mornington Island); three experimental plots, which are small areas for experiment, are set in each region, “Plot” 3, 5 and 6 in the West and “Plot” 9, 10, and 12 in the East. Geographical features of the seabed of the East and West are different; the East is deeper but the West is rougher and harder acoustically (Haywood et al., 2005). Such a difference suggests the need of separate analyses for the East and West regions. At each plot, three levels of experimental trawlings (“Treatment”) were repeated three times. The three levels are the intensities of trawling, 0, 4, and 20, and the number of repeated trawlings on each plot. The trawled seabed was dredged immediately after, 6, 12, and 18 months after trawling as well as before trawling, indicated by the variable “Time.” “Scientific name” is the name of the species caught by each dredge and “Count” is the number of individuals of each species caught in each dredge. “Weight” is the total dry weight of each case in grams.

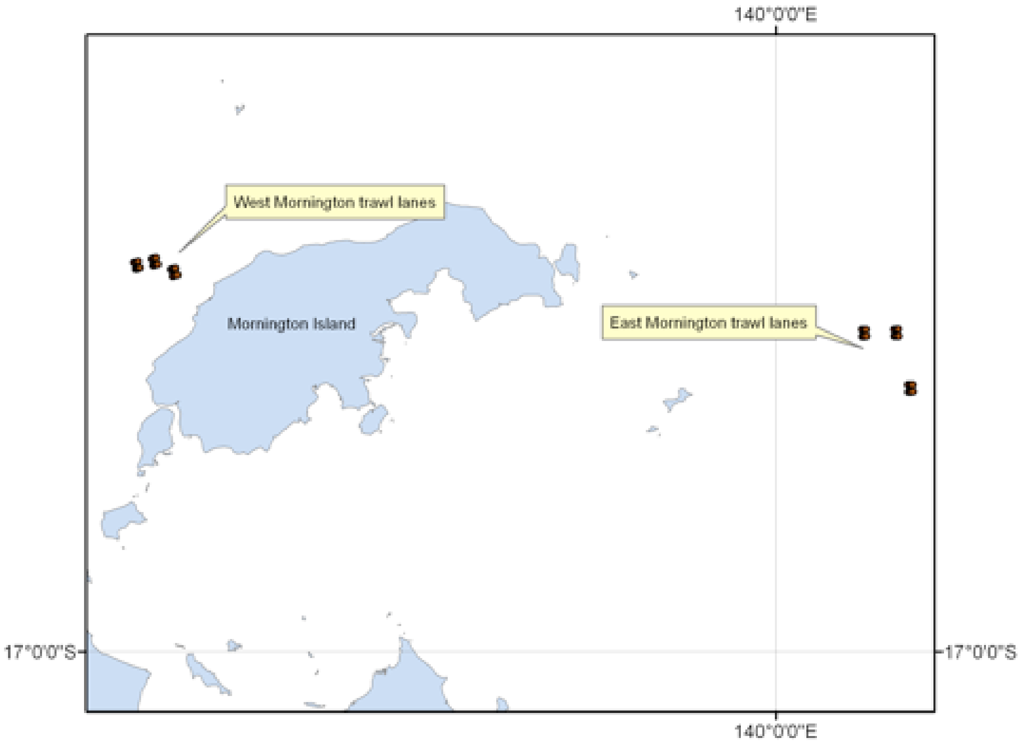


Figure 2.1: Two regions near Mornington Island used for the experimental survey (Haywood et al., 2005).

Although the primary aim of the survey was to investigate the effect of different levels of the trawl intensity and the recovery time, we will concentrate on whether the effect of trawling is significant, since the number of effective observations is not large enough for a detailed analysis because of the large number of empty catches. Therefore, in this analysis the treatment levels 4 and 20 are combined, and the weights recorded immediately after trawling are used for the analysis in contrast with the weights before trawling. We may satisfy ourselves if the effect of trawling were verified in a systematic manner. For this analysis only 16 classes of species were considered as there were too many zero catches for the other classes. Furthermore, 5 classes out of 16 are not appropriate for this analysis with the following reasons. Demospongiae is difficult to count since they are colonial, Pisces can easily escape from dredging and trawling, and Phaeophyta, Liliopsida, and Chlorophyta are fragile plants difficult to collect intact. Consequently, 11 classes remained for our investigation: Hydorozoa, Anthozoa, Gymnolaemata, Polychaeta, Bivalvia, Gastropoda, Asteroidea, Ophiuroidea, Echinoidea, Crustacea, and Ascidiacea. We also removed those species with observations of less than 5, while the maximum number of observations for each case is 27 before trawling and 18 after trawling. As a result 76 species remained in those classes for the analysis, although 778 species were observed in this survey.

2.3 Probability distribution model of the weight of animals on seabed

2.3.1 Stochastic growth model and its equilibrium distribution

Richards (1969) showed that many of the deterministic growth models are given by modifying the relative growth rate $(1/x)dx/dt$ as

$$\frac{dx_t}{x_t} = f(x_t)dt,$$

where $f(x_t)$ is a function of x_t at time t. An example of the deterministic models is the logistic growth model

$$dx_t = \rho x_t(\kappa - x_t)dt, \tag{2.1}$$

where κ is a growth limit and ρ is a rate of growth. This model is one of the well known models for population growth (Davidson, 1938, Smith, 1963), probably first proposed by Verhulst (1838). This model has since been used for describing many other aspects of growth other than the growth related to the population, for example, Marubini et al. (1972) analyzed the growth of boys' and girls' heights.

We introduce a stochastic growth model for individual weights of animals on seabed for each species at time t,

$$dX_t = rX_t\left(1-\frac{X_t}{k}\right)dt+\sigma X_t dB_t, \tag{2.2}$$

where r is the growth rate, k is the growth limit, and B_t is a standard Brownian motion. Model (2.2) is a stochastic version of (2.1), but unlike (2.1) it reflects the animal growth with some random fluctuations. In this respect the stochastic growth model (2.2) is a better model for the weight. To understand the relation between these two models, the derivation given by May (1973) might be helpful. May (1973) assumed that the growth limit κ randomly fluctuates as $\kappa = k+\gamma(t)$ reflecting environmental changes for plants and animals, where $\gamma(t) = \kappa - k$ represents the fluctuation of the growth limit around k at time t. If $\gamma(t)dt$ is given as

$$\gamma(t)dt = (\kappa - k)\,dt = \sigma_0 dB_t,$$

then (2.1) becomes

$$dX_t = (\rho k)X_t\left(1-\frac{X_t}{k}\right)dt+(\rho\sigma_0)X_t dB_t,$$

which is equal to the model (2.2) when $r = \rho\kappa$ and $\sigma = \rho\sigma_0$. The source of random fluctuation in this derivation is the growth limit κ. In other words, individual difference comes from different values of κ in this model.

In this analysis, we are interested in investigating the distributional change of weight X_t rather than tracing the growth of individual weights. Let $p(t,x)$ be the probability density function of X_t. Then, as shown in the following, $p(t,x)$ converges to the equilibrium distribution $p(x)$ as time t goes on, provided that $2r > \sigma^2$ (May, 1973).

To use the equilibrium distribution $p(x)$ as a model for the weight of animals on seabed, we consider the equilibrium distribution as follows. Let $p(t,x)$ is the distribution of the weight in the population at time t, not the age of the animals. If we assume that all individuals in the population grow as the stochastic differential equation (2.2), then the distribution $p(t,x)$ gets close to the equilibrium distribution $p(x)$ as time goes on. Here we note that the convergence of $p(t,x)$ to $p(x)$ does not depend on the distribution of $p(t,x)$, which is shown in the following. Assuming that the lifetime of the animals is long compared to the time that $p(t,x)$ can be approximated by $p(x)$, then it would be reasonable to use the equilibrium distribution $p(x)$ as a model for the weight of animals on seabed where no disturbance from outside was made for an adequate period.

Convergence to the equilibrium distribution

As has been shown by May (1973), the equilibrium distribution of X_t is the gamma distribution $\mathrm{G_A}(\nu,\alpha)$ with the probability density function

$$f(x,\boldsymbol{\theta}) = \frac{1}{\alpha\Gamma(\nu)}\left(\frac{x}{\alpha}\right)^{\nu-1}\exp\left(-\frac{x}{\alpha}\right),$$

where the shape parameter $\nu = 2r/\sigma^2 - 1 > 0$ and the scale parameter $\alpha = \sigma^2 k/2r > 0$. As a reference, we give a simple proof for the convergence to the equilibrium distribution as time tends to infinity with necessary conditions and an application for other growth models.

The equilibrium distribution is derived in a general frame work when X_t satisfies

$$dX_t = a(X_t)dt + b(X_t)dB_t \tag{2.3}$$

for some real functions $a(x) \neq 0$ and $b(x) \neq 0$. We hereafter assume that this is the Ito type stochastic differential equation. It is well known that the probability density function $p(t,x)$ for X_t satisfies the Kolmogorov forward equation,

$$\frac{\partial p(t,x)}{\partial t} = -\frac{\partial}{\partial x}\{a(x)p(t,x)\} + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left\{b^2(x)p(t,x)\right\}, \tag{2.4}$$

see for example Goel and Richter-Dyn (1974). If the equilibrium distribution $p(x) = \lim_{t\to\infty} p(t,x)$ exists, then it satisfies the equation

$$0 = -\frac{d}{dx}\{a(x)p(x)\} + \frac{1}{2}\frac{d^2}{dx^2}\left\{b^2(x)p(x)\right\}. \tag{2.5}$$

Theorem 1 gives us an explicit expression of $p(x)$ under some assumptions.

Theorem 1. *Assume that the equilibrium distribution $p(x)$ exists and satisfies*

$$\lim_{x\to\infty} p(x)a(x) = 0$$

and

$$\lim_{x\to\infty} \frac{d}{dx}\left\{b^2(x)p(x)\right\} = 0.$$

Then the solution of (2.5) can be written as

$$p(x) = \frac{C}{b^2(x)} \exp\left(\int^x \frac{2a(u)}{b^2(u)} du\right),$$

where C is a constant.

Proof. By integrating the both sides of (2.5), we have

$$a(x)p(x) - \frac{1}{2}\frac{d}{dx}\left\{b^2(x)p(x)\right\} + \tilde{C} = 0,$$

for a constant $\tilde{C}$. It is shown that $\tilde{C}$ is 0 because other terms tend to 0 as x tends to infinity from the assumptions. The result then easily follows from the fact that $q(x) = b^2(x)p(x)$ is the solution of

$$q'(x) - \frac{2a(x)}{b^2(x)} q(x) = 0.$$

□

In the analysis of the trawling data, we only use the equilibrium distribution of the model (2.2), however, we note that the following theorem shows that various types of distributions appear as equilibrium distributions of X_t from other growth models. Table 2.2 is a list of such distributions given in Rupšys (2007) for the case of the growth model

$$dX_t = rX_t^a \left\{1 - \left(\frac{X_t}{k}\right)^b\right\} dt + \sigma X_t dB_t. \tag{2.6}$$

We hereafter assume that $a, b > 0$ for model (2.6). The following theorem gives us an organized view of these distributions concentrated on $x > 0$.

Table 2.2: Some examples of equilibrium distributions for the growth model (2.6).

Law	Parameters(a,b)	$p(x)$
Verhulst	(1, 1)	$Cx^{2\left(r\sigma^{-2}-1\right)}\exp\left(-2rxk^{-1}\sigma^{-2}\right)$
Gompertz	(1, $b\to 0$)	$Cx^{-2}\exp\left(-r\left(\log\frac{k}{x}\right)^{2}\sigma^{-2}\right)$
Mitscherlich	(0,1)	$Cx^{-2\left(rk^{-1}\sigma^{-2}+1\right)}\exp\left(-2rx^{-1}\sigma^{-2}\right)$
Bertalanffy	$\left(\frac{2}{3},\frac{1}{3}\right)$	$Cx^{-2\left(rk^{-\frac{1}{3}}\sigma^{-2}+1\right)}\exp\left(-6rx^{-\frac{1}{3}}\sigma^{-2}\right)$
Richards	$(1,b\geq -1)$	$C\left(\frac{x}{k}\right)^{2\left(r\sigma^{-2}-1\right)}\exp\left(-2r\left(\frac{x}{k}\right)^{\beta}\beta^{-1}\sigma^{-2}\right)$

Theorem 2. *The equilibrium distribution $p(x)$ for (2.6) is a power transformed gamma distribution if and only if one of the following conditions is satisfied.*

1. *If $a=1$ and $2r>\sigma^2$, then X^b follows the gamma distribution with shape and scale parameters*

$$\nu=\frac{1}{b}\left(\frac{2r}{\sigma^2}-1\right),\quad \alpha=\frac{b\sigma^2k^b}{2r}.$$

2. *If $a+b=1$, then X^{-b} follows the gamma distribution with shape and scale parameters*

$$\nu=\frac{1}{b}\left(\frac{2r}{\sigma^2k^b}+1\right),\quad \alpha=\frac{b\sigma^2}{2r}.$$

Proof. Using Theorem 1 to the stochastic differential equation (2.6), we have

$$p(x)=\frac{C}{\sigma^2}x^{-2}\exp\left(\frac{2r}{\sigma^2}\int^x u^{a-2}-\frac{1}{k^b}u^{a+b-2}du\right).$$

If $a=1$, $p(x)$ becomes

$$p(x)=\frac{C}{\sigma^2}x^{-2+\frac{2r}{\sigma^2}}\exp\left(-\frac{2r}{b\sigma^2k^b}x^b\right),$$

which can be a probability density function of a power transformed gamma distribution if $2r>\sigma^2$ because the shape parameter is necessary to be $\nu>0$.

Therefore, a probability density function of $Y = X^b$ becomes

$$p_y(y) = \frac{C}{b\sigma^2} y^{\frac{2r\sigma^{-2}-1}{b}-1} \exp\left(-\frac{2r}{b\sigma^2 k^b} y\right),$$

which is the density function of the gamma distribution with shape and scale parameters

$$\nu = \frac{1}{b}\left(\frac{2r}{\sigma^2} - 1\right), \quad \alpha = \frac{b\sigma^2 k^b}{2r},$$

when $a = 1$ and $2r > \sigma^2$ are satisfied.

On the other hand, if $a+b = 1$, then

$$p(x) = \frac{C}{\sigma^2} x^{-2-\frac{2r}{\sigma^2 k^b}} \exp\left(-\frac{2r}{b\sigma^2} x^{-b}\right).$$

A probability density function of $Y = X^{-b}$ becomes

$$p_y(y) = \frac{C}{b\sigma^2} y^{\frac{\frac{2r}{\sigma^2 k^b}+1}{b}-1} \exp\left(-\frac{2r}{b\sigma^2} y\right),$$

which is the density function of the gamma distribution with

$$\nu = \frac{1}{b}\left(\frac{2r}{\sigma^2 k^b} + 1\right), \quad \alpha = \frac{b\sigma^2}{2r}.$$

If $a \neq 1$ and $a+b \neq 1$, $p(x)$ becomes

$$p(x) = \frac{C}{\sigma^2} x^{-2} \exp\left(\frac{2r}{(a-1)\sigma^2} x^{a-1} - \frac{2r}{(a+b-1)\sigma^2 k^b} x^{a+b-1}\right),$$

which cannot be the probability density function of any power transformed gamma distribution. □

It follows from Theorem 2 that the equilibrium distribution of the growth model (2.2), which is adopted in our analysis and corresponds to $a = 1$ and $b = 1$ in the general model (2.6), is the gamma distribution with the probability density function

$$p(x) = \frac{1}{\alpha^\nu \Gamma(\nu)} x^{\nu-1} \exp\left(-\frac{x}{\alpha}\right)$$

for $\nu = 2r/\sigma^2 - 1$ and $\alpha = \sigma^2 k/2r$, if $2r > \sigma^2$.

It is worth noting that the solution does not remain the same if the definition of the stochastic integral is not the Ito type integral for the stochastic differential equation (2.3). The equilibrium distribution is not necessarily the gamma distribution if other integral type such as the Stratonovich integral is employed (Feldman and Roughgarden, 1975).

Existence of the equilibrium distribution

For the existence of the equilibrium distribution $p(x) = \lim_{t\to\infty} p(t,x)$ in (2.4), one of the answers is given by Gihman and Skorohod (1979) in the framework of ergodic theory.

Theorem 3 (Gihman and Skorohod (1979), Theorem 3 in §18, Chapter 4)**.** *The equilibrium distribution of X_t exists for*

$$dX_t = \sigma(X_t)dB_t$$

and is written as

$$p(x) = \frac{\int_{-\infty}^{x} \frac{1}{\sigma^2(y)}dy}{\int_{-\infty}^{\infty} \frac{1}{\sigma^2(y)}dy},$$

provided that $\sigma(x)$ satisfies a first order Lipschitz condition and $\int_{-\infty}^{\infty} \frac{1}{\sigma^2(y)}dy < \infty$.

Theorem 3 can be applied for the case of (2.3). In fact, the function

$$f(x) = \int_0^x \exp\left(-\int_0^v \frac{2a(u)}{b^2(u)}du\right)dv$$

satisfies the equation

$$a(x)f'(x) + \frac{1}{2}b^2(x)f''(x) = 0,$$

so that $Y_t = f(X_t)$ satisfies the stochastic differential equation

$$dY_t = f'(X_t)b(X_t)dB_t$$

from the Ito formula. This is nothing more than the equation in Theorem 3 with $\sigma(x) = f'(x)b(x)$. Therefore, a sufficient condition for the existence of the equilibrium distribution for X_t of (2.3) is that both $a(x)$ and $b(x)$ satisfy a first order Lipschitz condition and

$$\int_{-\infty}^{\infty} \frac{1}{b^2(x)} \exp\left(2\int_0^x \frac{2a(u)}{b^2(u)} du\right) dx < \infty.$$

However, these conditions are very strong in practice. The Lipschitz condition for $a(x)$ is not satisfied for the case of (2.6) unless $a+b \leq 1$.

Another approach to showing the existence of the equilibrium distribution is to show the existence of the limit of the solution $p(t,x)$ of the Kolmogorov forward equation. We first state the following lemma for a general function $\psi(z)$, which is used in the proof of Theorem 4.

Lemma 1 (Levitan and Sargsjan (1991), Lemma 3.1.1 and its remark)**.** *For the spectrum of the problem*

$$\frac{d^2\psi(z)}{dz^2} + \{\lambda - U(z)\}\, \psi(z) = 0$$

$$\psi(0)\cos\alpha + \psi'(0)\sin\alpha = 0$$

to be discrete, it suffices that $U(z)$ tends to infinity as z tends to infinity.

We have the following theorem for

$$U(z) = \frac{d\tilde{a}(z)}{dz} + \tilde{a}^2(z),$$

where

$$\tilde{a}(z) = \frac{a(x)}{b(x)} - \frac{1}{2}\frac{db(x)}{dx}$$

and $z(x) = \int^x b^{-1}(u)du$. A sketch of the proof is given in Goel and Richter-Dyn (1974), however, necessary conditions are clarified in the following theorem.

Theorem 4. *Assume that $b(x) > 0$ and $U(z)$ is continuous and tends to infinity as z tends to infinity. Then,*

$$\lim_{t\to\infty} p(t,x) = \frac{q(x)}{b^2(x)},$$

where

$$q(x) = \frac{C_1 + C_2 \int^x P(u)du}{P(x)}$$

and

$$P(x) = \exp\left(-\int^x \frac{2a(u)}{b^2(u)}du\right).$$

Proof. From the definition of $z(x)$, it follows that $z(x)$ is a strictly increasing function of x since $b(x) > 0$. Let $g(t,z) = b(x)p(t,x)$. Then (2.4) can be rewritten as

$$\frac{\partial}{\partial t}\left\{\frac{g(t,z)}{b(x)}\right\} = -\frac{\partial}{\partial x}\left\{a(x)\frac{g(t,z)}{b(x)}\right\} + \frac{1}{2}\frac{\partial^2}{\partial x^2}\{g(t,z)b(x)\}$$

and we have

$$\frac{\partial g(t,z)}{\partial t} = -\frac{\partial}{\partial z}\{\tilde{a}(z)g(t,z)\} + \frac{1}{2}\frac{\partial^2}{\partial z^2}g(t,z). \tag{2.7}$$

Suppose that the solution is of the type $g(t,z) = Q(z)R(t)$. Then (2.7) becomes

$$\frac{dR(t)}{dt}\cdot\frac{1}{R(t)} = \frac{-2\frac{d}{dz}\{\tilde{a}(z)Q(z)\} + \frac{d^2}{dz^2}Q(z)}{2Q(z)}.$$

Therefore, we have

$$\frac{dR(t)}{dt}\cdot\frac{1}{R(t)} = \lambda$$

and

$$\frac{-2\frac{d}{dz}\{\tilde{a}(z)Q(z)\} + \frac{d^2}{dz^2}Q(z)}{2Q(z)} = \lambda$$

for a constant λ. The solutions of this simultaneous differential equation are written as $R(t) = e^{\lambda t+C}$ and $Q(z) = \pi^{-\frac{1}{2}}(z)\psi(z)$ by using $\pi(z) = \exp\left(-2\int^z \tilde{a}(u)du\right)$ and $\psi(z)$, which is the solution of

$$\frac{d^2\psi(z)}{dz^2} + \left\{\lambda - \frac{d\tilde{a}(z)}{dz} - \tilde{a}(z)^2\right\}\psi(z) = 0.$$

Here note that there exists an orthogonal basis $\{\psi(z,\lambda_n), n = 0,1,2,\ldots\}$ of $L^2([0,\infty),dz)$ from Lemma 1. Then $\{Q_n(z) = \psi(z,\lambda_n)\pi(z)^{-\frac{1}{2}}, n = 0,1,2,\ldots\}$ forms an orthogonal basis of $L^2([0,\infty),\pi dz)$. This implies that the solution of (2.7) is written as

$$g(t,z) = \sum_{n=0}^{\infty} \alpha_n(t)Q_n(z).$$

However, from the former discussion it is clear that $\alpha_n(t)$ can be written $\alpha_n(t) = a_n e^{\lambda_n t}$. We now have the solution of (2.7) is given by

$$g(t,z) = \sum_{n=0}^{\infty} a_n e^{\lambda_n t} Q_n(z).$$

For the existence of the limit of $g(t,z)$ in terms of t, all coefficients a_n have to be zero except a_k for the $\lambda_k = 0$. This implies that

$$\lim_{t\to\infty} g(t,z) = a_k Q_k(z).$$

It is enough to note that $g(t,z) = b(x)p(t,x)$ and $Q_k(z)$ is a base function for λ_k=0. □

We can now verify the existence of the equilibrium distribution for our growth model, $a(x) = rx(1-x/k)$ and $b(x) = \sigma x$, by checking the conditions in Theorem 4. It is clear that $b(x) > 0$ for any $x > 0$. The function

$$U(z) = \frac{d\tilde{a}(z)}{dz} + \tilde{a}(z)^2 = -\frac{rx}{k} + \left\{\frac{r}{\sigma}\left(1-\frac{x}{k}\right) - \frac{\sigma}{2}\right\}^2$$

is clearly continuous and $U(z)$ tends to infinity as $z(x)$ tends to infinity since

$$z(x) = \int^x b^{-1}(u)du = \frac{\log x}{\sigma}$$

tends to infinity as x tends to infinity and

$$\tilde{a}(z) = \frac{a(x)}{b(x)} - \frac{1}{2}\frac{db(x)}{dx} = \frac{r}{\sigma}\left(1-\frac{x}{k}\right) - \frac{\sigma}{2}.$$

The von Bertalanffy model

At the end of this subsection, we note the result when we tried to use the von Bertalanffy model for describing the weight distribution of animals on seabed instead of the growth model (2.2). A frequently used growth model for a scale, such as a length Y_t of animals, is so called the von Bertalanffy model (von Bertalanffy, 1960),

$$Y_t = Y_\infty - (Y_\infty - Y_0)\exp(-\beta t), \tag{2.8}$$

which is the solution of the differential equation

$$\frac{dY_t}{dt} = \beta(Y_\infty - Y_t),$$

where Y_∞ is the asymptotic length, Y_0 is the mean length at time 0, and β is the growth rate. Since this model is for a scale, the differential equation for the weight or volume becomes

$$dX_t = \rho\kappa X_t^{\frac{2}{3}}dt - \rho\kappa^{\frac{2}{3}}X_t dt \tag{2.9}$$

from (2.8) by putting $X_t = Y_t{}^3$. Here $\rho = 3\beta/Y_\infty{}^2$ and $\kappa = Y_\infty{}^3$. There are many articles which support the deterministic model (2.9) to use for describing the growth. For example, von Bertalanffy (1960) used the model for describing the difference between surface-proportional anabolism and weight-proportional catabolism. It then seems worthy of trying to fit a stochastic modification of (2.9) to our data,

$$dX_t = \rho\kappa X_t^{\frac{2}{3}}dt - \rho\kappa^{\frac{2}{3}}X_t dt + \sigma X_t dB_t. \tag{2.10}$$

However, as a result, it did not work well for our data. One of the reasons that the goodness-of-fit test of the model is rejected for many cases would be that the von Bertalanffy model is mainly for tracing the individual growth in size, for example, tracing the growth of plant or any other increasing size, which approaches to the growth limit. Therefore, this model is not good enough for describing an equilibrium in a population. In fact, the equilibrium distribution of the stochastic model (2.10) is a power transformed gamma distribution with power $-1/3$ as seen from Theorem 2. Such a negative power transformed gamma distribution does not seem reasonable.

Also, the use of such a distribution for which no reproducibility property holds true causes a lot of problems in the estimation of parameters and the goodness-of-fit test. The distribution of the total weight of each case, which is the only available observation, becomes much more complicated. For these reasons we concentrate ourselves on the model (2.2) in this analysis.

2.3.2 Maximum likelihood estimator

As has been mentioned before, only total weights for each case in each dredge were recorded in this survey since measuring individual weights takes time and money. Therefore, the jth observation Y_j is considered to be the sum of unobserved individual weights $\{X_{jk}, k = 1, 2, \ldots, m_j\}$, such as

$$Y_j = X_{j1} + X_{j2} + \cdots + X_{jm_j}, \quad j = 1, 2, \ldots, n,$$

where m_j is the number of individuals caught in the jth dredge. The variables $Y_1, Y_2, \ldots, Y_n$ for the observations are now independent but not identically distributed random variables. Fortunately, the reproducibility of the gamma distribution provides us a simple treatment of such non-identically distributed random variables. That is, the Y_j is still distributed as the gamma distribution $\mathrm{G_A}(m_j \nu, \alpha)$, provided that X_{jk}, $j = 1, \ldots, n$, $k = 1, 2, \ldots, m_j$, are independent and identically distributed as $\mathrm{G_A}(\nu, \alpha)$, i.e. all individuals share the same scale α and shape ν parameters. The maximum likelihood estimator of α is a function of ν,

$$\hat{\alpha} = \frac{y}{m\nu},$$

where $y = \sum_{j=1}^{n} y_j$ is the sum of observed total weights $y_1, y_2, \ldots, y_n$ and $m = \sum_{j=1}^{n} m_j$ is the sum of the number of individuals observed. Although no closed form is known for the maximum likelihood estimator of ν and α, we could obtain the numerical value of the estimate of ν by a numerical algorithm to maximize the profile likelihood,

$$L(\nu,\hat{\alpha}) \;=\; \sum_{j=1}^{n}\left\{-\log\left(\Gamma(m_j\nu)\right)+(m_j\nu-1)\log y_j\right\}-m\nu\log\left(\frac{y}{m\nu}\right)-m\nu.$$

The function “nlminb”, which is an implementation of the nonlinear minimization program on R, is used for the estimation.

The standard errors of the estimates are calculated from the inverse of the Fisher information matrix. The consistency and the asymptotic normality of the maximum likelihood estimator when the observations are independent but not identically distributed are already proved in Hoadley (1971) under suitable regularity conditions. Since such regularity conditions are satisfied in our case, the asymptotic variance covariance matrix of the estimators is given by the inverse of the Fisher information matrix. If we assume that the number of individuals caught $m_j, j=1,2,\ldots,n$, are reproduced even after the nth observation or the numbers are distributed with the same probability, the Fisher information matrix $I(\boldsymbol{\theta})$ is given by

$$I(\boldsymbol{\theta})=\begin{pmatrix} M_1 & M_2 \\ M_2 & \frac{\nu}{\alpha}M_2 \end{pmatrix},$$

where

$$M_1=\frac{1}{n}\sum_{j=1}^{n}m_j^2\psi'(m_j\nu), \quad M_2=\frac{1}{n\alpha}\sum_{j=1}^{n}m_j$$

with trigamma function $\psi'(\nu)=\frac{d^2}{d\nu^2}\log\Gamma(\nu)$. Then the standard errors of the estimates are obtained by

$$\frac{1}{\sqrt{n}}\left(\frac{\nu}{\alpha}M_1M_2-M_2^2\right)^{-1}\frac{\nu}{\alpha}M_2$$

for the shape parameter and

$$\frac{1}{\sqrt{n}}\left(\frac{\nu}{\alpha}M_1M_2-M_2^2\right)^{-1}M_1$$

for the scale parameter, respectively.

2.3.3 Goodness-of-fit

Checking goodness-of-fit is important when a probability distribution model is fitted to data. We adopt a P-P plot for a graphical method for checking goodness-of-fit since the observations are not identically distributed in this case. The P-P plot for n independent observations $y_1, y_2, \ldots, y_n$ is a plot of the n points,

$$\left(\frac{j-0.5}{n}, z_{(j)} \right),$$

where $z_j = F_j\left(y_j, \boldsymbol{\theta}\right)$ and $z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(n)}$ are order statistics of z_j, $j = 1, 2, \ldots, n$. Here the distribution function $F_j\left(y, \boldsymbol{\theta}\right)$ is that of the gamma distribution with the parameter $(m_j \nu, \alpha)$ for the common parameter $\boldsymbol{\theta} = (\nu, \alpha)$.

A goodness-of-fit test statistic for independent but not identically distributed random variables $Y_1, Y_2, \ldots, Y_n$ parallel to the P-P plot would be

$$\tilde{W}_n^2\left(\boldsymbol{\theta}\right) = \sum_{j=1}^{n} \left\{ Z_{(j)} - \frac{j-0.5}{n} \right\}^2 + \frac{1}{12n}, \qquad (2.11)$$

where $Z_j = F_j(Y_j, \boldsymbol{\theta})$, $j = 1, 2, \ldots, n$. This statistic is an extension of the Cramér-von Mises statistic $W_n^2(\boldsymbol{\theta})$ in (1.2) because $W_n^2(\boldsymbol{\theta})$ can be reduced from $\tilde{W}_n^2(\boldsymbol{\theta})$ when the observations are independent and identically distributed as $F(x, \boldsymbol{\theta})$. In addition, the distribution of $\tilde{W}_n^2(\boldsymbol{\theta})$ is equal to that of $W_n^2(\boldsymbol{\theta})$ when the parameters are known since Z_j is the transformation of Y_j, $j = 1, 2, \ldots, n$, by its distribution function so that Z_j follows the standard uniform distribution, which is the same when the observations are independent and identically distributed.

When the parameters are unknown and are necessary to be estimated from a sample, the asymptotic distribution of the statistic $\tilde{W}_n^2(\hat{\boldsymbol{\theta}})$, where the estimator $\hat{\boldsymbol{\theta}}$ is plugged in (2.11) instead of $\boldsymbol{\theta}$, cannot be obtained by a simple extension of the case when the observations are independent and identically distributed. There are some articles on the behavior of the empirical process $\sqrt{n}(\tilde{F}_n(z) - z)$, where $\tilde{F}_n(z)$ is the empirical distribution function of $Z_1, Z_2, \ldots, Z_n$, by Pierce and Kopecky (1979) and Loynes (1980). However, it is not directly useful to obtain the p-values of the test statistic $\tilde{W}_n^2(\hat{\boldsymbol{\theta}})$.

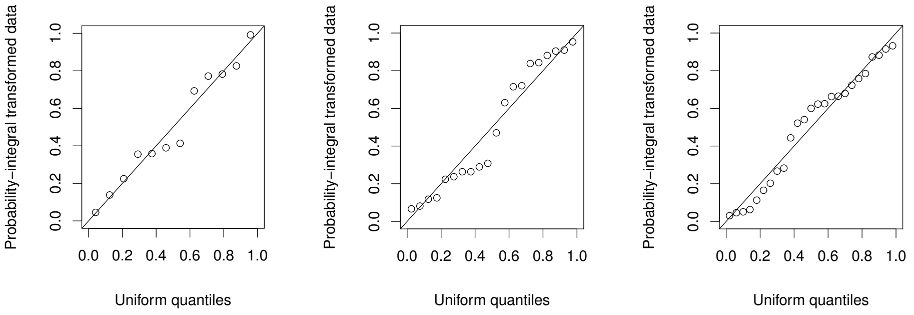
In the following, we first evaluate the p-value of $\tilde{W}_n^2(\hat{\boldsymbol{\theta}})$ because the parameters are necessary to be estimated from the observations. The p-values are obtained from 500 sets of generated random numbers from $\mathrm{G_A}(m_j\nu,\alpha)$, $j=1,2,\ldots,n$. Since it is not clear how the distribution of $\tilde{W}_n^2(\hat{\boldsymbol{\theta}})$ depends on the value of the estimate $\hat{\boldsymbol{\theta}}$, we evaluate the goodness-of-fit at several lattice points in the neighborhood, $(m_j\nu,\alpha)$ for $\nu=0.5\hat{\nu},0.75\hat{\nu},1.25\hat{\nu}$, and $1.5\hat{\nu}$ and $\alpha=0.5\hat{\alpha},0.75\hat{\alpha},1.25\hat{\alpha}$, and $1.5\hat{\alpha}$, not only at the point estimate $(m_j\hat{\nu},\hat{\alpha})$. As an example, Table 2.3 shows the p-values in the neighborhood for Case 2 in Table A.1 in the Appendix A with $\hat{\nu}=1.140$ and $\hat{\alpha}=0.967$. As seen from the Table 2.3, the p-value does not fluctuate so much, ranging from 0.091 to 0.133, so that we use the minimum in the neighborhood as a p-value through this analysis, which is favorable to the rejection of the fit. For all cases before trawling, we evaluate the p-value for the goodness-of-fit of the gamma distribution to the observations by the method described here.

Table 2.3: The p-values of the goodness-of-fit test for Case 2.

	$0.5\hat{\alpha}$	$0.75\hat{\alpha}$	$\hat{\alpha}$	$1.25\hat{\alpha}$	$1.5\hat{\alpha}$
$0.5\hat{\nu}$	0.119	0.106	0.108	0.121	0.096
$0.75\hat{\nu}$	0.125	0.127	0.126	0.114	0.114
$\hat{\nu}$	0.129	0.128	0.125	0.108	0.133
$1.25\hat{\nu}$	0.115	0.106	0.114	0.118	0.122
$1.5\hat{\nu}$	0.115	0.091	0.099	0.103	0.092

2.4 Distributions before trawling

The results for all species are shown in Table A.1 in the Appendix A, where the gamma distribution $\mathrm{G_A}(\nu,\alpha)$ is fitted to individual weights before trawling for the cases numbered from 1 to 80. Each case can be identified by a combination of its scientific name and the region name of the experiment. The class and family names are also listed as a reference. Species identified by scientific name are grouped into a family and several families are further grouped into a class. We can see what kind of animal was caught as a by-catch of the prawns from the seabed. The column labeled n indicates the number of nonzero observations out of 27 observations in each case. The maximum likelihood estimates of the parameters $\hat{\nu}$ and $\hat{\alpha}$ are also listed. The last two columns give the values of goodness-of-fit test statistic, $\tilde{W}_n^2\left(\hat{\boldsymbol{\theta}}\right)$, and the corresponding p-values obtained from the distribution of $\tilde{W}_n^2\left(\hat{\boldsymbol{\theta}}\right)$ with the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$. In the table p-values less than 0.1 are marked by $*$ as a reference. It seems reasonable to exclude these 23 cases for which the goodness-of-fit test is rejected at significance level $\alpha = 0.1$. For later analysis, we concentrate our attention on 57 cases out of 80 to investigate the effect of trawling because we are going to verify the effect through changes of the equilibrium distribution of the stochastic growth model (2.2). For the visual understanding of the goodness-of-fit of the 57 unmarked cases, P-P plots are given in Figure 2.2 for Cases 1, 2, and 3 as examples.



Case 1. Case 2. Case 3.

Figure 2.2: P-P plots for Cases 1, 2, and 3.

To understand the meaning of the estimated parameters for the 57 cases, a reasonable transform of the parameters would be $k = \alpha(\nu+1)$ and $\xi = \sqrt{2/(\nu+1)}$ because it is equivalent to rewrite the model (2.2) as

$$dX_s = X_s\left(1 - \frac{X_s}{k}\right)ds + \xi X_s dB_s,$$

where time is changed from t to $s = rt$. The parameter k is now the growth limit and ξ is the degree of randomness around the growth limit k. Figure 2.3 is a scatter plot of $\hat{\xi} = \sqrt{2/(\hat{\nu}+1)}$ and $\hat{k} = \hat{\alpha}(\hat{\nu}+1)$ where Case 12 is excluded because $\hat{k} = 735.870$ is very large as the growth limit with $\hat{\xi} = 1.219$. The points on the scatter plot are identified by initial letters of class names. For example, H is for the class Hydrozoa as described in the legend. We observe that the value of k is very large for 4 species, but it only implies that these species have heavy dry weights. However, it is interesting to note that species in the same class share similar $\hat{\xi}$ values for several classes. The value $\hat{\xi}$ is less than 0.8 for Bivalvia, greater than 1.0 for Hydrozoa and between 0.7 and 1.1 for Gymnolaemata.

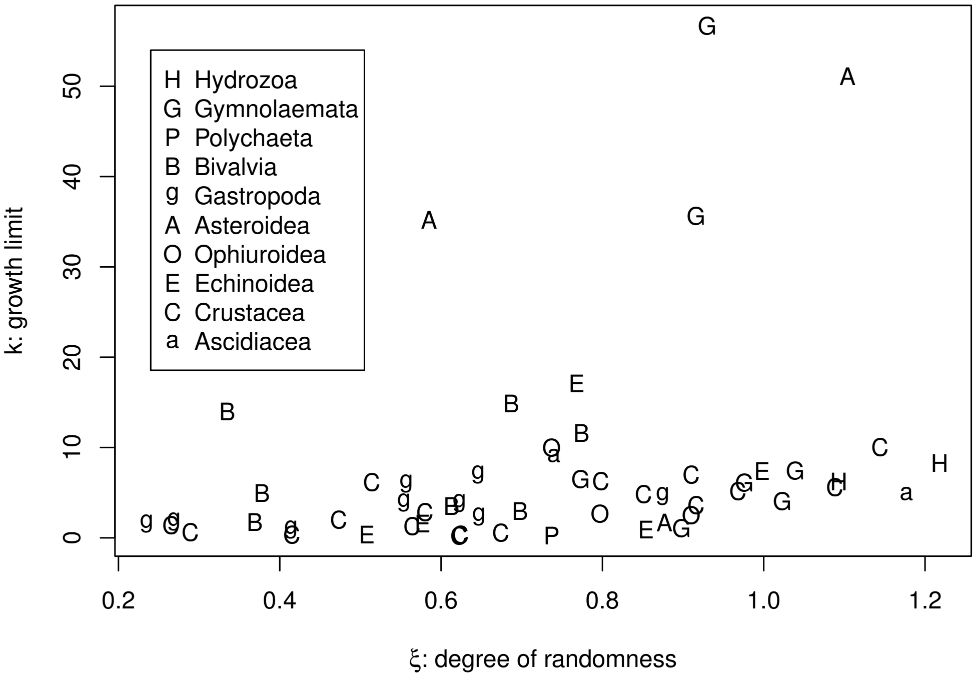


Figure 2.3: Scatter plot of the degree of randomness $\hat{\xi} = \sqrt{2/(\hat{\nu}+1)}$ and the growth limit $\hat{k} = \hat{\alpha}(\hat{\nu}+1)$.

There are some possible reasons that the gamma distribution does not fit well for the remaining 23 cases. For some species, the stochastic growth model (2.2) or its equilibrium distribution may not be a good model for describing their weights. On the other hand, there are some species for which the goodness-of-fit test of the gamma distribution is rejected for the observations on one region while it is not for the observations on the other region. In fact, only one species *Dardanus imbricatus* shows the gamma distribution can be used for the data observed both in the West and East out of four species in the list for which enough observations are available both in the West and East (Cases 6, 7, 41, 42, 50, 51, 57, and 58). Possibly, in these cases, environmental factors may have delayed species maturity. Besides, it is worth noting that there is no consistent rejection of the goodness-of-fit test over different species. This result also suggests that it is necessary to consider the species by species or case by case analysis.

2.5 The effect of trawling

In this section we verify the effect of trawling through discrepancies between the gamma distribution with the parameters estimated from the observations before trawling and the distribution of the weight observed after trawling. The data are obtained under a careful design of experiments (Haywood et al., 2005) so that it is natural to assume that there are no effects other than trawling. For this reason we verify the effect of trawling if the discrepancy is significant. Here we note that the number of target cases is now 47 since not enough observations are available after trawling for the remaining cases.

Although the existence of the discrepancy is examined by the goodness-of-fit test, the statistical framework for the calculation of the p-value is different from the method we have used for the case of before trawling. The goodness-of-fit test we apply here is to test the goodness-of-fit of the gamma distribution with the parameters estimated from the observations before trawling to the observations after trawling, which corresponds to the goodness-of-fit test when parameters are known. We hereafter denote the parameters estimated from the observations before trawling as $\boldsymbol{\theta}_0$ to make sure that the goodness-of-fit of the distribution when the parameters are known is examined. The distribution of $\tilde{W}_n^2(\boldsymbol{\theta}_0)$ is equal to that of $W_n^2(\boldsymbol{\theta}_0)$ as we explained in Section 2.3.3, and the asymptotic distribution of $W_n^2(\boldsymbol{\theta}_0)$ is given as a distribution of a weighted sum of chi-squared random variables with 1 degree of freedom, as described in Section 1.2. However, we obtain the p-value by generating 500 sets of random numbers distributed as the gamma distribution with the parameters estimated from the observations before trawling to obtain the distribution of the test statistic because the sample sizes may not be enough large to use the asymptotic distribution of the statistic.

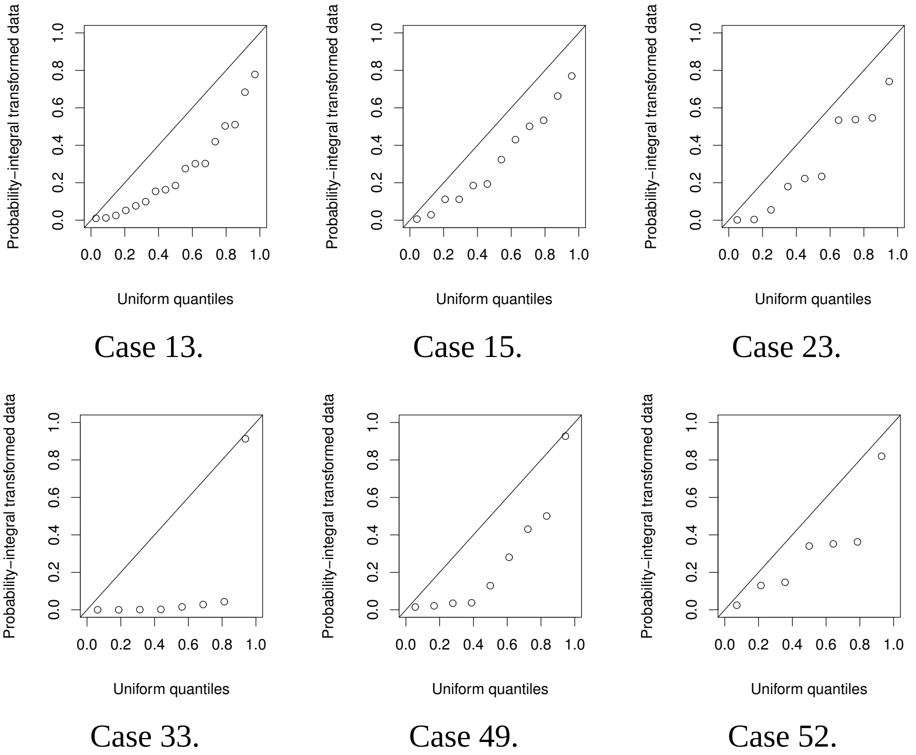
A summary of the results for the 47 cases is given in Table 2.4. The unaffected cases, for which the goodness-of-fit test is not rejected, are denoted by U in the column labeled “Effect.” For other cases, for which the goodness-of-fit test is rejected, the directions of the change of the weight distribution from the gamma distribution with the parameters estimated from the observations before trawling are denoted in the column of “Effect,” where L is for lighter cases and C or C(L) is for the cases that the weight distribution is changed but not consistently lighter or heavier. There are no heavier cases in our study. The direction of the change, which represents the type of the effect, is determined as follows. It is L if all points are below the line $y = x$ on the P-P plot, which indicates that $z_{(j)} < (j - 0.5)/n$ for all $j = 1, 2, \ldots, n$. Type C(L) is for the case when the type of the effect is almost same as the case for L, with a few exceptional points on the P-P plot.

We can see more details about the changes of the distribution through P-P plots. Figure 2.4 shows P-P plots for 6 cases of type L. A possible reason that the weight distribution is changed to the lighter direction after trawling would be that those species have difficulty to avoid the trawl net and only individuals smaller than the net size remain, so that the distribution is skewed in the lighter direction.

Figure 2.5 shows P-P plots for 5 cases of type C and C(L). It is observed that the distribution is skewed to lighter direction for Case 21, Case 53, and Case 75, if a single point on the P-P plot is ignored. However, there is no clear direction

Table 2.4: The effect of trawling (U:Unaffected, L:Lighter, C:Changed).

Case	n	$\tilde{W}_n^2(\boldsymbol{\theta}_0)$	p-value	Effect	Case	n	$\tilde{W}_n^2(\boldsymbol{\theta}_0)$	p-value	Effect
1	6	0.092	0.636	U	34	9	0.115	0.550	U
3	15	0.106	0.584	U	36	14	0.091	0.650	U
5	9	0.076	0.734	U	40	7	0.319	0.106	U
7	13	0.059	0.838	U	41	6	0.319	0.112	U
9	8	0.077	0.748	U	45	7	0.096	0.650	U
11	8	0.144	0.424	U	46	18	0.264	0.184	U
12	6	0.097	0.604	U	47	11	0.143	0.404	U
13	17	1.075	0.000	L	49	9	0.658	0.022	L
14	8	0.223	0.212	U	51	13	0.166	0.348	U
15	12	0.441	0.058	L	52	7	0.367	0.088	L
19	9	0.150	0.422	U	53	12	1.556	0.000	C(L)
20	10	0.044	0.918	U	57	11	0.211	0.248	U
21	13	0.549	0.030	C(L)	58	8	0.191	0.276	U
22	12	0.183	0.334	U	59	14	0.242	0.210	U
23	10	0.446	0.054	L	60	7	0.097	0.640	U
24	12	0.125	0.530	U	61	12	0.390	0.094	C
25	9	0.212	0.272	U	63	10	0.200	0.250	U
26	15	0.458	0.036	C	70	7	0.066	0.808	U
27	14	0.184	0.322	U	71	15	0.188	0.300	U
28	8	0.225	0.208	U	73	6	0.277	0.136	U
30	9	0.122	0.516	U	75	10	0.697	0.008	C(L)
31	11	0.197	0.276	U	76	6	0.327	0.100	U
32	18	0.253	0.210	U	78	12	0.355	0.118	U
33	8	1.663	0.000	L					

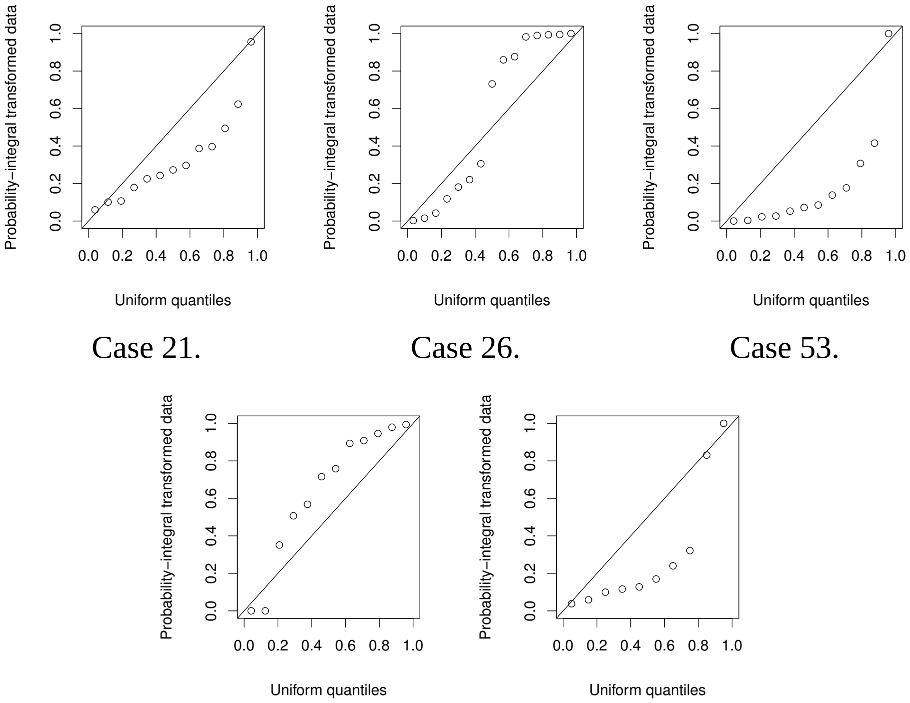


Case 13. Case 15. Case 23.

Case 33. Case 49. Case 52.

Figure 2.4: P-P plots for 6 cases where the weight distribution became lighter after trawling. The null distributions are $\mathrm{G_A}(1.313, 24.512)$, $\mathrm{G_A}(2.687, 0.062)$, $\mathrm{G_A}(4.320, 0.660)$, $\mathrm{G_A}(35.112, 0.043)$, $\mathrm{G_A}(2.388, 5.049)$, and $\mathrm{G_A}(5.018, 0.259)$, respectively.

of change for the other cases, Case 26 and Case 61. The reason why the weight distribution is changed in no clear direction might be that such species are more sensitive to other factors like the local unevenness of the environment rather than the trawling effect.



Case 21.

Case 26.

Case 53.

Case 61.

Case 75.

Figure 2.5: P-P plots for 5 cases where the weight distribution changed without direction after trawling. The null distributions are $\mathrm{G_A}(2.342,3.467)$, $\mathrm{G_A}(26.664,0.0463)$, $\mathrm{G_A}(1.007,3.693)$, $\mathrm{G_A}(27.462,0.049)$, and $\mathrm{G_A}(0.690,3.327)$, respectively.

2.6 Comparisons with simple mean tests

We have seen how the trawling effect is verified by using the extended version of the Cramér-von Mises goodness-of-fit test of the gamma distribution, which is the model for the weight distribution of animals on seabed. It would be worthwhile to compare this result with that obtained by a simple mean difference test statistic like Welch's t-test statistic as

$$T = \frac{\overline{W_a} - \overline{W_b}}{\sqrt{\frac{{s_a}^2}{n_a} + \frac{{s_b}^2}{n_b}}}.$$

Here $\overline{W_a}$ is the sample mean of the weights normalized by the number of individuals observed after trawling, ${s_a}^2$ is the sample variance, and n_a is the sample size. $\overline{W_b}$, ${s_b}^2$, and n_b are those for the normalized weights before trawling. Table 2.5 shows the p-values for $\tilde{W}_n^2(\boldsymbol{\theta}_0)$ given in Table 2.4 and for Welch's t-test statistic for the two sided alternative hypothesis in the case of type U. Also the p-values for Student's t-test statistic,

$$T = \frac{\overline{W_a} - \overline{W_b}}{\sqrt{\frac{{s_a}^2}{n_a}}},$$

where $\overline{W_b}$ is assumed to be known, are shown in the table as a reference. This is because the p-values for $\tilde{W}_n^2(\theta_0)$ are obtained for the case when the parameters are known. The sign of Welch's t-test statistic, which is the same sign of Student's t-test statistic, is also given in Table 2.5.

It seems reasonable that the p-values for Welch's t-tests are all large for the cases of type U. However, the values themselves are not consistent with those for $\tilde{W}_n^2(\theta_0)$, particularly for the three cases marked † in Table 2.5. The reason is that the discrepancy from the weight distribution is symmetric so that the mean difference fails in describing such a discrepancy as seen in the P-P plots given in Figure 2.6. We also note that the p-values for Student's t-tests are not consistent with those for Welch's t-tests, particularly for the three cases marked $*$ in this table. It can be seen from the P-P plots for those three cases given in Figure 2.7 that Student's t-test is sensitive to a small shift of the distribution.

Table 2.6 shows the result for type L, corresponding to the result for type U in Table 2.5. It is clear that Welch's t-test fails in detecting changes in the 3 cases

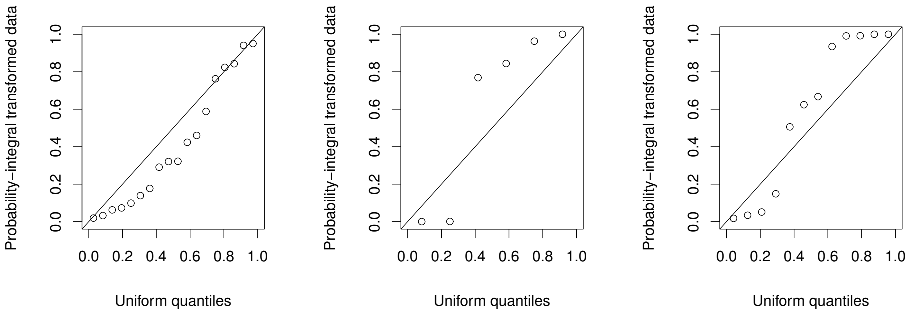
Table 2.5: The p-values for type U.

Case	1	3	5	7	9	11	12
p-value ($\tilde{W}_n^2(\boldsymbol{\theta}_0)$)	0.636	0.584	0.734	0.838	0.748	0.424	0.604
p-value (Welch's t-test)	0.524	0.541	0.735	0.642	0.376	0.756	0.804
p-value (Student's t-test)	0.315	0.428	0.690	0.507	0.354	0.734	0.679
sign(T)	–	–	+	–	+	+	–

Case	14	19	20	22	24	25	27	28	30	31^*
p-value ($\tilde{W}_n^2(\boldsymbol{\theta}_0)$)	0.212	0.422	0.918	0.334	0.530	0.272	0.322	0.208	0.516	0.276
p-value (Welch's t-test)	0.318	0.777	0.975	0.757	0.810	0.465	0.423	0.628	0.674	0.154
p-value (Student's t-test)	0.130	0.756	0.960	0.592	0.771	0.135	0.298	0.269	0.526	0.097
sign(T)	–	–	–	+	–	+	–	+	+	+

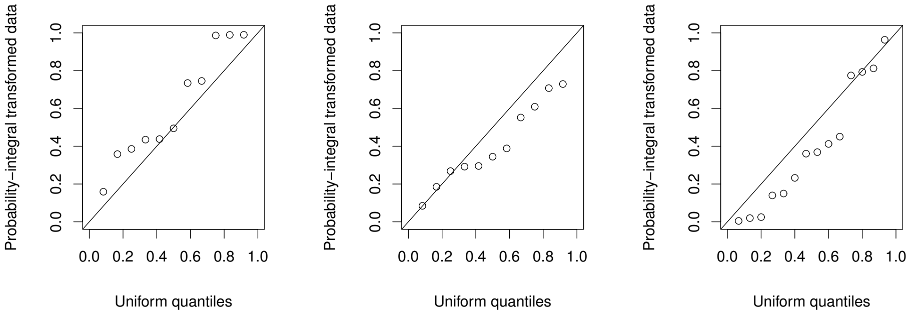
Case	32	34	36	40	41	45	$46^\dagger$	47	51	57^*
p-value ($\tilde{W}_n^2(\boldsymbol{\theta}_0)$)	0.210	0.550	0.650	0.106	0.112	0.650	0.184	0.404	0.348	0.248
p-value (Welch's t-test)	0.851	0.967	0.721	0.177	0.260	0.981	0.992	0.954	0.393	0.157
p-value (Student's t-test)	0.808	0.955	0.609	0.105	0.186	0.976	0.990	0.947	0.279	0.016
sign(T)	+	+	–	–	+	+	–	–	–	–

Case	58	59^*	60	63	70	71	73	$76^\dagger$	$78^\dagger$
p-value ($\tilde{W}_n^2(\boldsymbol{\theta}_0)$)	0.276	0.210	0.640	0.250	0.808	0.300	0.136	0.100	0.118
p-value (Welch's t-test)	0.372	0.269	0.963	0.766	0.730	0.425	0.338	0.606	0.833
p-value (Student's t-test)	0.330	0.055	0.945	0.682	0.670	0.354	0.301	0.595	0.797
sign(T)	–	–	+	+	–	+	–	+	–



Case 46. Case 76. Case 78.

Figure 2.6: P-P plots for Cases 46, 76, and 78 after trawling.



Case 31. Case 57. Case 59.

Figure 2.7: P-P plots for Cases 31, 57, and 59 after trawling.

marked by †, although the values of Welch's t-test statistics are all negative. The sensitivity of Student's t-test makes a difference for Case 23 as well as for type U.

Table 2.6: The p-values for type L.

Case	13	$15^\dagger$	$23^\dagger$	33	49	$52^\dagger$
p-value ($\tilde{W}_n^2(\boldsymbol{\theta}_0)$)	0.000	0.058	0.054	0.000	0.022	0.088
p-value (Welch's t-test)	0.005	0.291	0.194	0.010	0.020	0.486
p-value (Student's t-test)	0.000	0.174	0.014	0.009	0.013	0.384
sign(T)	−	−	−	−	−	−

Table 2.7 is for type C and C(L). In this case, Welch's t-test fails in detecting changes at a level of 0.1 in 3 cases out of the 5 cases. A significant difference is shown for Case 75, where the sign of Welch's t-test is positive although it belongs to type C(L).

In summary, Welch's t-test tends to fail in the detection of distributional changes when the weight distribution after trawling differs from the weight distribution before trawling in a symmetric manner. On the other hand, Student's t-test seems very sensitive for slight differences from the weight distribution before trawling. Such mean difference tests are simple and easy to use, but not strong enough for investigating distributional changes since the distributions are only identified by the mean in their tests.

Table 2.7: The p-values for type C or C(L).

Case	21	$26^{\dagger}$	53	$61^{\dagger}$	$75^{\dagger}$
p-value ($\tilde{W}_n^2(\boldsymbol{\theta}_0)$)	0.030	0.036	0.000	0.094	0.008
p-value (Welch's t-test)	0.084	0.168	0.056	0.937	0.720
p-value (Student's t-test)	0.020	0.073	0.008	0.928	0.696
sign(T)	–	+	–	–	+

2.7 Concluding remarks

We have shown that the gamma distribution, the equilibrium distribution of the stochastic growth model, can describe the distribution of weight of animals on seabed. Goodness-of-fit of the distribution is examined by using the extended version of the Cramér-von Mises statistic with a P-P plot. One of the reasons why we need such a test is that only the total weights of catches for each species are recorded in the survey. As a result the integrated use of numerical and graphical methods for checking goodness-of-fit shows the trawling effect on the weight distribution of animals on seabed through the change of the distribution.

Another approach to examining the difference of the weight distributions between before and after trawling would be the likelihood ratio test, which compares the parameters of the gamma distribution. However, this approach does not give an answer for the case when the gamma distribution does not fit to the data observed after trawling, which is happened in some cases. Also our approach, checking the direction of the change of the weight distribution from the gamma distribution with the parameters estimated from the observations before trawling, would give a simpler understanding how the weight distribution changed than comparing the changes of the parameters. For these reasons we have examined the distributional change rather than the change of the parameters of the gamma distribution in this analysis.

Chapter 3

The effect of freshwater flows on the growth of banana prawns

Another case study, which is the modeling the length-frequency data of banana prawns, is presented in this chapter. By using the derived probability distribution model, the effect of freshwater flows on the growth of banana prawns is investigated.

3.1 Introduction

It is important to understand the role of freshwater flows into estuaries, the downstream sections of rivers and streams, and the requirement for a sustainable environment, especially in Australia, because the water resources are limited but a demand for human use is increasing. A wide review of the need of freshwater flows for estuarine fisheries in tropical areas can be found in Robins et al. (2005). For that reason the project "Environmental flows for sub-tropical estuaries: understanding the freshwater needs for sustainable fisheries production and assessing the impacts of water regulation." was initiated in Australia, whose data are analyzed in this chapter. The project aimed at an investigation of the effects of freshwater flows on estuarine fisheries production. Although a preliminary analysis is published in Halliday and Robins (2007), there still remain problems unsolved.

In this chapter, we focus on banana prawn (*Penaeus merguiensis*), which is known to be one of the significant target species in the trawl fisheries of northern Australia and has been investigated by many researchers. Lucas et al. (1979) assessed the state of the banana prawn stocks in the Gulf of Carpentaria, Australia, by yield per recruit analysis based on the studies of migration, growth, and mortality. The effects of temperature and salinity on growth and survival were examined by Staples and Heales (1991) from laboratory experiments. Haywood and Staples (1993) investigated growth and mortality of juvenile banana prawns from data sampled from 1986 to 1989 in the north-eastern Gulf of Carpentaria. The size-dependent mortality of juvenile banana prawns was suggested by Wang and Haywood (1999). For the behavior of postlarval penaeid prawns, including banana prawns, the effect of tide and day/night on the vertical migration was explored by Vance and Pendrey (2008).

The effects of freshwater flows on the growth rate of banana prawns have been investigated in Halliday and Robins (2007) for the data we analyze in this chapter. They decomposed length-frequency distributions of banana prawns into components of the normal distributions to identify means and found the links of the means to identify the cohort. For each links, the first and last dates the cohort was sampled were set to be t_1 and t_2 and the mean carapace lengths on the dates were L_{t_1} and L_{t_2}, respectively. Then they investigated the effects of environmental factors by modeling the growth rate K in the von Bertalanffy model

$$L_{t_2} = L_{t_1} + (L_\infty - L_{t_1}) \left\{1 - e^{-K(t_2 - t_1)}\right\},$$

which is already introduced in (2.8), as a function of freshwater inflow and other environmental factors. For example, the final model for the growth rate K for the Calliope River is in the form of

$$K = \beta_0 + \beta_1 T + \beta_2 T^2 + \beta_3 W_0 + \beta_4 W_4,$$

where T is temperature, W_0 is the total freshwater inflow for the period between t_1 and t_2, and W_4 is the total freshwater inflow four weeks before t_1.

Although the effects of freshwater flows and other environmental factors are included in their model, the explanation led by this model is not convincing enough. For example, the distribution of the length has different shape time by time so that using only the mean would not be enough to investigate the effect of environmental factors on the growth. Also salinity of water is observed but it was not used in their model. To overcome such weakness of their analysis, we introduce a new probability distribution model for the length of banana prawns.

Descriptions of the data we analyze are given in Section 3.2. We build the probability distribution model for length-frequency data of banana prawns in Section 3.3. In Section 3.4, the methods for fitting the model to the data and testing goodness-of-fit of the distribution are described. Results of the fit to the data are given in Section 3.5.

3.2 Data

The data we analyze here are obtained in the Fisheries Research and Development Corporation (FRDC) funded Project 2001/022, "Environmental flows for sub-tropical estuaries: understanding the freshwater needs for sustainable fisheries production and assessing the impacts of water regulation." in Australia. Although the surveys were done for some species in three rivers, the Fitzroy River, the Calliope River, and the Boyne River, we focus on the banana prawns catch data observed in the Calliope River from December 4th, 2002 to April 19th, 2004 in this analysis. The data are observed fortnightly from the beginning of the survey to July 12th, 2003 and in 4 weeks after then.

The target data are the carapace length-frequency data of banana prawns in the estuary. Table 3.1 is a part of the length-frequency data. Banana prawns caught were measured to a truncated 1 mm Carapace Length (CL) size-class, that is, 1.00 to 1.99 mm are counted for 1 mm CL. The size of the carapace length is ranged from 1 mm truncated CL size-class to 33 mm truncated CL size-class. The total catches of banana prawns for each size-class within 8 sites are observed.

Table 3.1: Part of the length-frequency data.

Date	Length class (mmCL)	Number of catches
2003/12/4	7	1
2003/12/4	8	2
2003/12/4	9	3
2003/12/4	10	9

There are also data of environmental factors: temperature, salinity, pH, and turbidity. Table 3.2 is a part of the data of the environmental factors. Temperature is of water and salinity is recorded in ‰. Turbidity gives the depth in meter no longer visible the Secchi disc and the large value indicates that water is transparent. Although the number of catches is recorded by summing up among 8 sites, these environmental factor data are observed in each site. For this reason we use the environmental factor data by averaging over sites.

Table 3.2: Part of the data of the environmental factors.

Unique site number	Temperature	Salinity	pH	Turbidity
5-1	30.11	38.64	7.94	0.50
5-2	30.82	38.87	7.72	0.30
5-3	33.43	38.61	7.96	0.20
5-4	31.17	39.26	7.90	0.20
5-5	33.21	38.54	7.99	0.35
5-5	33.21	38.54	7.99	0.35
5-6	32.97	38.65	7.96	0.40
5-7	31.81	38.72	7.89	0.30
5-8	33.60	38.47	7.98	0.40

3.3 Probability distribution model of the carapace length of banana prawns

To construct a probability distribution model of the length of banana prawns, we assume that the observations are constituted of two kinds of cohorts because of the life cycle of banana prawns and the interval of the samplings. Banana prawns are spawned in offshore waters and larvae and post-larvae migrate into estuaries. After several months in the estuary they migrate to coastal marine waters (Halliday and Robins, 2007). They have around one year life cycle, and on the other hand, the samplings were done fortnightly or 4 weeks in the estuary in this survey. Therefore, it is natural to consider that the observations are a mixture of two kinds of cohorts, a cohort which has been stayed in the estuary from the previous sampling and a cohort which migrates from offshore waters to the estuary after the previous sampling. We also note that it is reasonable to assume that banana prawns migrate from offshore waters in a cohort because it is known that peaks of spawning of banana prawns are on new and full moon.

3.3.1 Cohort stayed in the estuary

We first consider constructing a probability distribution model of the carapace length of banana prawns for a cohort stayed in the estuary by using the data observed in the previous sampling. One of the natural models would be given by a transformation of the previous carapace length distribution with reflecting growth and survival.

Let $f_{t_0}(x)$ be a probability density function for the distribution of the carapace length at time t_0. Also we define $g(x,t_0,t)$ and $q(x,t_0,t)$ to be a carapace increment and a survival rate during the period (t_0,t), where x is the length at time t_0. If we assume that a proportion of prawns migrating to coastal marine waters is constant for each length, then the distribution function of the carapace length at time $t > t_0$

becomes

$$G(x,t_0,t) = c_1 \int_0^x f_{t_0}(y - g(y,t_0,t))q(y,t_0,t)dy \tag{3.1}$$

with a normalized constant c_1. In this analysis, $f_{t_0}(x)$ is given by a polygon approximation of the observed length-frequency data. To construct models for the carapace increment $g(x,t_0,t)$ and the survival rate $q(x,t_0,t)$, we use models suggested by Staples and Heales (1991) and Wang and Haywood (1999) as follows.

Carapace increment $g(x,t_0,t)$

For the carapace increment $g(x,t_0,t)$, we use two models, for the intermoult period and for the moult increment, obtained by Staples and Heales (1991). From laboratory experiments, they derived the models such that

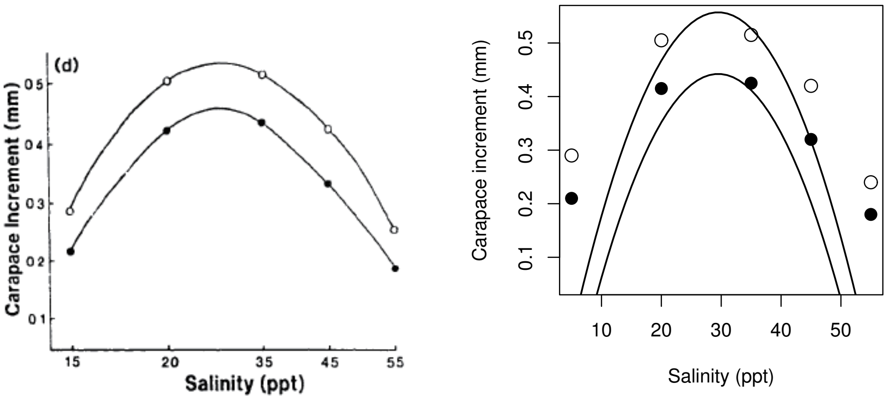
$$\begin{aligned} t_m - t_{m-1} = 13.919 - 0.411T + 0.027(T-25)^2 \\ - 0.014S + 0.001(S-30)^2 + 0.201x_{m-1} \end{aligned}$$

for the intermoult period and

$$\begin{aligned} x_m - x_{m-1} = 0.039 + 0.012T - 0.002(T-25)^2 \\ - 0.001S - 0.001(S-30)^2 + 0.023x_{m-1} \end{aligned} \tag{3.2}$$

for the moult increment. Here t_m denotes the day of the mth moult, x_m is the carapace length in mm after the mth moult, T is temperature, and S is salinity (‰), where temperature and salinity were held constant in their experiments. These models show that the intermoult period and the moult increment depend on temperature and salinity.

Before applying these models to the data, we give a modification to the model (3.2) because there seems to be some rounding errors in the coefficients of the model (3.2). Figure 3.1 (a) is the figure given in Staples and Heales (1991), describing the moult increment against salinity at temperature 28°C. The black dots are for 5 mm CL and the white dots are for 10 mm CL. On the other hand, in Figure 3.1 (b), the curves based on the model (3.2) at temperature 28°C are drawn and the points are plotted as imitating the points in Figure 3.1 (a) to make it easy to compare. For this inconsistency, we have estimated the coefficients for S and $(S-30)^2$ to fit the curves in Figure 3.1 (a) given by Staples and Heales (1991) and use the modified model



(a) Figure from Staples and Heales (1991). (b) Curve based on the model (3.2).

Figure 3.1: Plots of the carapace increment against salinity (‰).

$$x_m - x_{m-1} = 0.039 + 0.012T - 0.002(T-25)^2$$
$$- 0.00126S - 0.0004(S-30)^2 + 0.023x_{m-1} \qquad (3.3)$$

instead of (3.2). As a reference, the curves based on the model (3.3) at temperature 28°C is drawn in Figure 3.2 with the points plotted as imitating the points in Figure 3.1 (a).

Moreover, we note that these models suggest that a growth rate, which is the carapace increment per day, is approximately constant in length. If we rewrite the

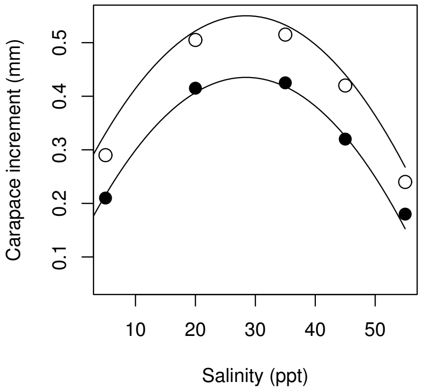


Figure 3.2: Plot of the carapace increment against salinity (‰). The curve based on the model (3.3) are drawn.

models as

$$t_m - t_{m-1} = f_1(T,S) + 0.201 x_{m-1} \tag{3.4}$$

$$x_m - x_{m-1} = f_2(T,S) + 0.023 x_{m-1} \tag{3.5}$$

with the functions $f_1(T,S)$ and $f_2(T,S)$ of T and S, then the sums of the intermoult period and the moult increment from the m_0th moult to the mth moult are given as

$$\begin{aligned} t_m - t_{m_0} &= \frac{0.201}{0.023}\left\{x_{m_0} + \frac{f_2(T,S)}{0.023}\right\}\left\{(1+0.023)^{m-m_0} - 1\right\} \\ &\quad + (m-m_0)\left\{f_1(T,S) - \frac{0.201}{0.023} f_2(T,S)\right\} \\ &\approx (m-m_0)\left\{f_1(T,S) + 0.201 x_{m_0}\right\} \end{aligned}$$

and

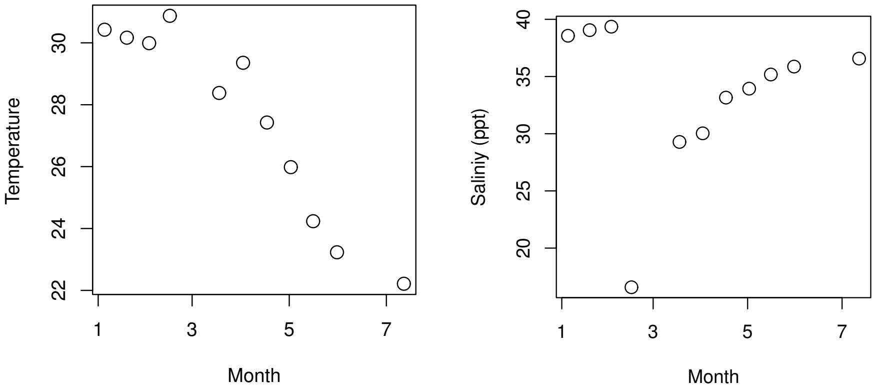
$$\begin{aligned} x_m - x_{m_0} &= \left\{(1+0.023)^{m-m_0} - 1\right\} x_{m_0} + \frac{(1+0.023)^{m-m_0} - 1}{0.023} f_2(T,S) \\ &\approx (m-m_0)\left\{f_2(T,S) + 0.023 x_{m_0}\right\} \end{aligned}$$

using an approximation $(1+0.023)^{m-m_0} \approx 1+0.023(m-m_0)$. Therefore, we have

$$\frac{x_m - x_{m_0}}{t_m - t_{m_0}} \approx \frac{f_2(T,S)+0.023x_{m_0}}{f_1(T,S)+0.201x_{m_0}} = \frac{x_{m_0+1} - x_{m_0}}{t_{m_0+1} - t_{m_0}} \tag{3.6}$$

and this approximation suggests that the growth rate is approximately constant in length if temperature and salinity hold constant.

By using these models for the intermoult period and for the moult increment, we employ the following steps to obtain the carapace increment $g(x,t_0,t)$ in (3.1). First we use the locally weighted scatter plot smoothing (loess) to the data of temperature and salinity because they were only observed on the day of the sampling and were changing throughout the survey. Figure 3.3 gives plots of temperature and salinity from January to July in 2003 as an example. As shown in Figure 3.3, temperature changes along with the season and salinity changes suddenly, which is because of freshwater flows.



(a) Temperature. (b) Salinity.

Figure 3.3: Plots of temperature and salinity for each sampling date from January to July in 2003.

Using temperature and salinity obtained by the loess, the number of moults during the period (t_0,t) is given by m which is the maximum number satisfying

$t > t_m$, where t_m is given by

$$t_m - t_0 = 0.201 \left\{ \frac{(1+0.023)^m - 1}{0.023} x_0 + \sum_{l=0}^{m-1} \sum_{l'=1}^{l} (1+0.023)^{l'-1} f_2(T_{t_{l-l'}}, S_{t_{l-l'}}) \right\}$$
$$+ \sum_{l=0}^{m-1} f_1(T_{t_l}, S_{t_l}),$$

which is obtained from (3.4) and (3.5) by reflecting the changes of temperature and salinity. Here we assume that the t_0 is the time immediately after the moult and the length at t_0 is x_0. T_t and S_t are temperature and salinity at time t obtained by the loess. By using the number of moults m given above, the total carapace increment during the period (t_0, t) is then given as

$$x_m - x_0 = \{(1+0.023)^m - 1\} x_0 + \sum_{l=0}^{m-1} (1+0.023)^l f_2(T_{t_{m-l+1}}, S_{t_{m-l+1}}).$$

We denote $g^*(x_0, t_0, t) = x_m - x_0$ as the total carapace increment during the period (t_0, t) for a banana prawn whose carapace length is x_0 at time t_0 and use

$$g(x, t_0, t) = \frac{1}{30} \sum_{y=1}^{30} g^*(y, t_0, t) \tag{3.7}$$

for the carapace increment $g(x, t_0, t)$ in our analysis. Here we note that $g(x, t_0, t)$ does not depend on the carapace length x. The averaging over the carapace length $y = 1, 2, \ldots, 30$ is to obtain a good approximation of the carapace increment. This is because the growth rate is approximately constant in length and the averaging would help to take an account of errors of the moult day. We also note that the growth rate is approximately constant in time, which is also shown by (3.6).

Survival rate $q(y, t_0, t)$

For the survival rate, we use the size-dependent mortality rate model proposed by Wang and Haywood (1999) from data observed in the Gulf of Carpentaria. Assume that the instantaneous mortality rate at carapace length x has a form of the exponential function $\alpha e^{\beta x}$, then the size-dependent mortality of banana prawns after δ weeks is given as

$$q^*(x_{t_0},\delta,\gamma) = \exp\left[-\frac{\alpha}{\beta\gamma}\left\{e^{\beta(x_{t_0}+\gamma\delta)} - e^{\beta x_{t_0}}\right\}\right], \tag{3.8}$$

where the growth rate γ mm per week is constant and x_{t_0} is the length at time t_0. For the parameters α and β, we use the values $\hat{\alpha} = 1.594$ and $\hat{\beta} = -0.2919$, which were obtained in Wang and Haywood (1999).

Using the size-dependent mortality rate model, the survival rate $q(x,t_0,t)$ in the model (3.1), which is for a cohort stayed in the estuary, is given as

$$q(x,t_0,t) = q^*\left(x, t-t_0, \frac{g(x,t_0,t)}{t-t_0}\right),$$

where the growth rate γ in (3.8) is given by averaging the carapace total increment $g(x,t_0,t)$ of (3.7) during the period (t_0,t).

3.3.2 Cohort migrated from offshore waters to the estuary

To construct a model for a cohort migrated from offshore waters to the estuary, we use the size-dependent mortality rate model (3.8) again. We assume that the carapace length immediately after hatching follows the normal distribution with mean μ_0 and variance σ^2. Also we assume that the growth rate in offshore waters is $\gamma_0 = 1$ (Haywood and Staples, 1993, Wang and Haywood, 1999). Then the distribution of the carapace length δ weeks after hatching is given as

$$H(x,\delta',\sigma) = c_2 \int_0^x \phi\left(\frac{y-\delta'}{\sigma}\right) q^*(y,\delta',1)dy, \tag{3.9}$$

where c_2 is a constant and $\delta' = \delta + \mu_0$. Here we have used an approximation of $e^{-\beta\mu_0}$ by 1 because we assume that μ_0 is small enough.

3.4 Fitting the model to the data

We fit the model to the data for 19 cases shown in Table 3.3 in this analysis. This is because the model for a cohort stayed in the estuary $G(x,t_0,t)$ given in (3.1) is constructed from the data observed in the previous sampling. For example, Case 1 in Table 3.3 refers to modeling the data observed in December 16th, 2002 by the mixture of cohorts stayed in the estuary from December 4th, 2002 and migrated from offshore waters to the estuary. We exclude the cases for December 4th, 2002 and October 26th, 2003 because there is no sampling prior to these dates.

Table 3.3: Target cases.

Case	Date	Number of observations
-	2002/12/04	23
1	2002/12/16	41
2	2003/01/19	43
3	2003/02/02	54
4	2003/02/15	75
5	2003/03/18	1715
6	2003/04/02	382
7	2003/04/17	465
8	2003/05/02	307
9	2003/05/16	341
10	2003/05/31	64
11	2003/07/12	26
-	2003/10/26	23
12	2003/11/25	170
13	2003/12/23	88
14	2004/01/18	633
15	2004/01/22	761
16	2004/02/16	204
17	2004/02/20	251
18	2004/03/21	174
19	2004/04/19	35

3.4.1 Model for a mixture of different cohorts

In general case, a mixture model of cohorts stayed in the estuary and cohorts migrated from offshore waters to the estuary is given as

$$F(x, \boldsymbol{\theta}) = r_0 G(x, t_0, t) + \sum_{j \geq 1} r_j H(x, \delta_j, \sigma_j),$$

where $0 \leq r_j \leq 1,\ j = 0, 1, \ldots$, and $\sum_{j \geq 0} r_j = 1$, however, it turns out that only the following two models are useful in this analysis because the results show that these models can be used for 15 cases out of 19.

- Model 1 :

  $$F(x, \boldsymbol{\theta}) = r_0 G(x, t_0, t) + r_1 H(x, \delta_1, \sigma_1),$$

  where $\boldsymbol{\theta} = (r_0, \delta_1, \sigma_1)$ and $r_1 = 1 - r_0$.

- Model 2 :

  $$F(x, \boldsymbol{\theta}) = r_0 G(x, t_0, t) + r_1 H(x, \delta_1, \sigma_1) + r_2 H(x, \delta_2, \sigma_2)$$

  where $\boldsymbol{\theta} = (r_0, r_1, \delta_1, \delta_2, \sigma_1, \sigma_2)$ and $r_2 = 1 - r_0 - r_1$.

Although we have examined the mixture model with more than 3 components for the other 4 cases, the goodness-of-fit tests were rejected for any number of components.

Since the data for the carapace length are binned data, we consider a grouped distribution of $F(x, \boldsymbol{\theta})$, which is defined as

$$p_j(\boldsymbol{\theta}) = F(j+1, \boldsymbol{\theta}) - F(j, \boldsymbol{\theta}), \quad j = 1, 2, \ldots, 33,$$

for the probability distribution model of the carapace length of banana prawns.

3.4.2 Parameter estimation

The parameter $\boldsymbol{\theta}$ in the probability distribution model $p_j(\boldsymbol{\theta}),\ j = 1,2,\ldots,33$, is estimated by minimizing the Crámer-von Mises statistic for discrete distributions

$$W_n^{(d)2}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{k=1}^{K}\left\{\sum_{j=1}^{k} o_j - n\sum_{j=1}^{k} p_j(\boldsymbol{\theta})\right\}^2 p_k(\boldsymbol{\theta}),$$

which is already introduced in (1.4), with two constrains in the estimation of the parameters. One of the constrains is for the rate parameters, r_0 in Model 1 and r_0, r_1, and r_2 in Model 2, to take on values between 0 and 1. The other constrain is for δ_2 in Model 2. To avoid to become a too flexible model, we assume δ_2 as $\delta_2 = \delta_1 + d$ and adopt the value d which gives the highest p-value among $d = 2,4,6,\ldots$. Since it is known that peaks of spawning of banana prawns are on new and full moon, we here assume that the interval of the migration to the estuary is $2,4,6,\ldots$ weeks. For this reason the constrain for δ_2 to be $\delta_2 = \delta_1 + d$ would be reasonable.

The standard errors of the estimates of the parameters are calculated using a parametric bootstrap, with 500 bootstrap samples, because of the constrains.

3.4.3 Goodness-of-fit test

To check goodness-of-fit of the derived probability distribution model to the data, we use the Crámer-von Mises statistic for discrete distributions when the parameters are estimated by the minimum distance method with two constrains described in Section 3.4.2. We calculate the p-values using the parametric bootstrap, as same as for the standard errors of the estimates of the parameters.

Although the p-values are calculated using parametric bootstrap in this analysis, as a reference, we give the following theorem, which shows the asymptotic distribution of the Crámer-von Mises statistic for discrete distributions when parameters are estimated by the minimum distance method, which is the estimation method of finding the value $\hat{\boldsymbol{\theta}}$ which makes the Crámer-von Mises statistic for discrete distributions a minimum. We note that the asymptotic distribution of the statistic when parameters are estimated by the maximum likelihood method is given by Lockhart et al. (2007).

Theorem 5. *Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables following a discrete distribution with K cells labeled $1, 2, \ldots, K$ and probability $p_j(\boldsymbol{\theta})$ of falling into cell j, $j = 1, 2, \ldots, K$. Then the asymptotic distribution of $W_n^{(d)2}(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the minimum distance estimator, is given as a distribution of a weighted sum of chi-squared random variables with 1 degree of freedom, such that*

$$\sum_{j=1}^{K-1} \lambda_j V_j^2,$$

where V_j follows the standard normal distribution and λ_j is an eigenvalue of a $K \times K$ matrix

$$\Sigma_y P(\boldsymbol{\theta}) \left[I - AZ(\boldsymbol{\theta}) \left\{ Z(\boldsymbol{\theta})^\top A^\top P(\boldsymbol{\theta}) AZ(\boldsymbol{\theta}) \right\}^{-1} Z(\boldsymbol{\theta})^\top A^\top P(\boldsymbol{\theta}) \right], \tag{3.10}$$

for $j = 1, 2, \ldots, K-1$. Here $\Sigma_y, P(\boldsymbol{\theta})$, A, and $Z(\boldsymbol{\theta})$ are defined in Section 1.3.

Proof. We first show that the estimation error $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$ is approximated by

$$-\left\{ Z(\boldsymbol{\theta})^\top A^\top P(\boldsymbol{\theta}) AZ(\boldsymbol{\theta}) \right\}^{-1} Z(\boldsymbol{\theta})^\top A^\top P(\boldsymbol{\theta}) \boldsymbol{y},$$

where $\boldsymbol{y}$ is defined in Section 1.3. Since the minimum distance estimator $\hat{\boldsymbol{\theta}}$ is a solution of

$$\frac{\partial}{\partial \boldsymbol{\theta}} \left[\frac{1}{n} \sum_{k=1}^{K} \left\{ \sum_{j=1}^{k} o_j - n \sum_{j=1}^{k} p_j(\boldsymbol{\theta}) \right\}^2 p_k(\boldsymbol{\theta}) \right] \Bigg|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} = \mathbf{0},$$

where $\mathbf{0} = (0, 0, \ldots, 0)^\top$, and

$$\frac{1}{n} \sum_{k=1}^{K} \left\{ \sum_{j=1}^{k} o_j - n \sum_{j=1}^{k} p_j\left(\hat{\boldsymbol{\theta}}\right) \right\} \left\{ -n \sum_{j=1}^{k} \frac{\partial}{\partial \boldsymbol{\theta}} p_j(\boldsymbol{\theta}) \Bigg|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \right\} \left\{ p_k\left(\hat{\boldsymbol{\theta}}\right) - p_k(\boldsymbol{\theta}) \right\}$$

converges to 0 as n tends to infinity, it is shown that

$$\frac{1}{n} \sum_{k=1}^{K} \left\{ \sum_{j=1}^{k} o_j - n \sum_{j=1}^{k} p_j\left(\hat{\boldsymbol{\theta}}\right) \right\} \left\{ -n \sum_{j=1}^{k} \frac{\partial}{\partial \boldsymbol{\theta}} p_j(\boldsymbol{\theta}) \Bigg|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \right\} p_k(\boldsymbol{\theta}) \tag{3.11}$$

converges to 0 as n tends to infinity. Applying a Taylor expansion to (3.11) around $\boldsymbol{\theta}$ gives an approximation of $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ as

$$\left[\sum_{k=1}^{K}\left\{\sum_{j=1}^{k}\frac{\partial}{\partial\boldsymbol{\theta}}p_j(\boldsymbol{\theta})\right\}\left\{\sum_{j=1}^{k}\frac{\partial}{\partial\boldsymbol{\theta}^\top}p_j(\boldsymbol{\theta})\right\}p_k(\boldsymbol{\theta})\right.$$
$$\left.+\frac{1}{n}\sum_{k=1}^{K}\left\{\sum_{j=1}^{k}o_j-n\sum_{j=1}^{k}p_j(\boldsymbol{\theta})\right\}\left\{\sum_{j=1}^{k}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}p_j(\boldsymbol{\theta})\right\}p_k(\boldsymbol{\theta})\right]^{-1}$$
$$\times\sqrt{n}\left[\frac{1}{n}\sum_{k=1}^{K}\left\{\sum_{j=1}^{k}o_j-n\sum_{j=1}^{k}p_j(\boldsymbol{\theta})\right\}\left\{\sum_{j=1}^{k}\frac{\partial}{\partial\boldsymbol{\theta}}p_j(\boldsymbol{\theta})\right\}p_k(\boldsymbol{\theta})\right].$$

Note that

$$\frac{1}{n}\left\{\sum_{j=1}^{k}o_j-n\sum_{j=1}^{k}p_j(\boldsymbol{\theta})\right\}$$

converges to 0 as n tends to infinity for any $k=1,2,\ldots K$, thus $\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)$ can be approximated as

$$\left[\sum_{k=1}^{K}\left\{\sum_{j=1}^{k}\frac{\partial}{\partial\boldsymbol{\theta}}p_j(\boldsymbol{\theta})\right\}\left\{\sum_{j=1}^{k}\frac{\partial}{\partial\boldsymbol{\theta}^\top}p_j(\boldsymbol{\theta})\right\}p_k(\boldsymbol{\theta})\right]^{-1}$$
$$\times\sqrt{n}\left[\frac{1}{n}\sum_{k=1}^{K}\left\{\sum_{j=1}^{k}o_j-n\sum_{j=1}^{k}p_j(\boldsymbol{\theta})\right\}\left\{\sum_{j=1}^{k}\frac{\partial}{\partial\boldsymbol{\theta}}p_j(\boldsymbol{\theta})\right\}p_k(\boldsymbol{\theta})\right]$$
$$=\left\{Z(\boldsymbol{\theta})^\top A^\top P(\boldsymbol{\theta})AZ(\boldsymbol{\theta})\right\}^{-1}Z(\boldsymbol{\theta})^\top A^\top P(\boldsymbol{\theta})\boldsymbol{y}.$$

On the other hand, applying a Taylor expansion to $W_n^{(d)2}(\boldsymbol{\theta})$ around the minimum distance estimator $\hat{\boldsymbol{\theta}}$ and using the approximation shown above, we have

$$W_n^{(d)2}(\hat{\boldsymbol{\theta}})+n\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^\top Z(\boldsymbol{\theta})^\top A^\top P(\boldsymbol{\theta})AZ(\boldsymbol{\theta})\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right).$$

From these approximations it follows that $W_n^{(d)2}(\hat{\boldsymbol{\theta}})$ can be approximated by

$$\boldsymbol{y}^\top P(\boldsymbol{\theta})\left[I - AZ(\boldsymbol{\theta})\left\{Z(\boldsymbol{\theta})^\top A^\top P(\boldsymbol{\theta})AZ(\boldsymbol{\theta})\right\}^{-1} Z(\boldsymbol{\theta})^\top A^\top P(\boldsymbol{\theta})\right]\boldsymbol{y}$$

$$= \left(\Sigma_y^{-\frac{1}{2}}\boldsymbol{y}\right)^\top D\left(\Sigma_y^{-\frac{1}{2}}\boldsymbol{y}\right),$$

where D is a $K \times K$ matrix defined by

$$D = \Sigma_y^{\frac{1}{2}} P(\boldsymbol{\theta})\left[I - AZ(\boldsymbol{\theta})\left\{Z(\boldsymbol{\theta})^\top A^\top P(\boldsymbol{\theta})AZ(\boldsymbol{\theta})\right\}^{-1} Z(\boldsymbol{\theta})^\top A^\top P(\boldsymbol{\theta})\right]\Sigma_y^{\frac{1}{2}}.$$

As used by Choulakian et al. (1994) in their proof, the distribution of $\Sigma_y^{-\frac{1}{2}}\boldsymbol{y}$ converges to the multivariate standard normal distribution, therefore, the asymptotic distribution of $W_n^{(d)2}(\hat{\boldsymbol{\theta}})$ is given as a distribution of a weighted sum of chi-squared random variables with 1 degree of freedom, where the weights are the eigenvalues of the matrix D. The equivalence of the eigenvalues of D and (3.10) is easily checked. □

3.5 Results

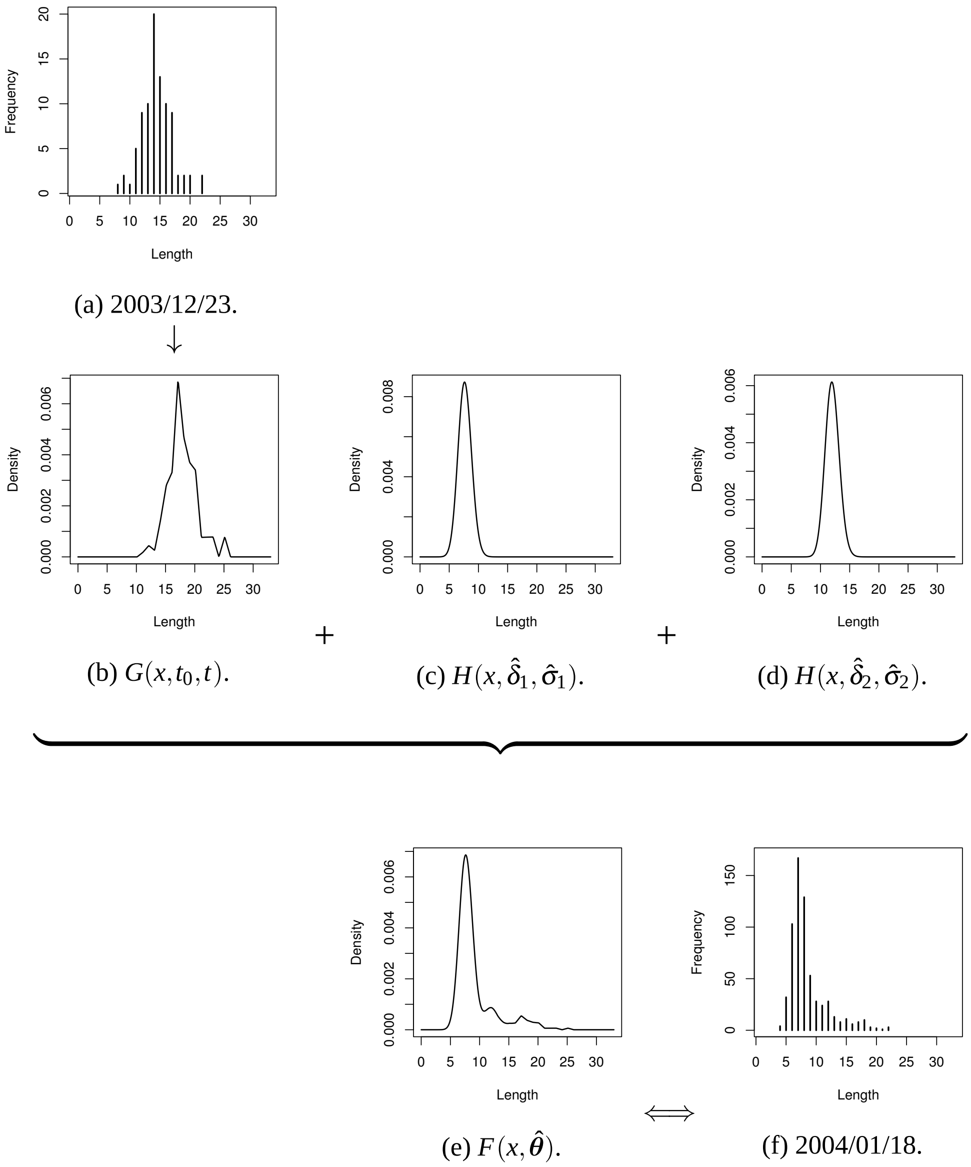
Figure 3.4 shows the result for Case 14 fitting Model 2 as an example. The data observed on December 23rd, 2003 are shown in Figure 3.4 (a). By using this distribution, a model for a cohort stayed in the estuary $G(x,t_0,t)$ is determined as described in Section 3.3.1, where t_0 is December 23rd , 2003 and t is January 18th, 2004. The probability density function of $G(x,t_0,t)$ is drawn in Figure 3.4 (b). On the other hand, probability density functions of models for two cohorts migrated from offshore waters to the estuary $H(x,\hat{\delta}_1,\hat{\sigma}_1)$ and $H(x,\hat{\delta}_2,\hat{\sigma}_2)$ are drawn in Figure 3.4 (c) and (d) with the estimated parameters $\hat{\delta}_1 = 6.014$, $\hat{\sigma}_1 = 1.396$, $\hat{\delta}_2 = 10.014$, and $\hat{\sigma}_2 = 1.498$. Combining the distributions $G(x,t_0,t)$, $H(x,\hat{\delta}_1,\hat{\sigma}_1)$, and $H(x,\hat{\delta}_2,\hat{\sigma}_2)$, the mixture model is given as

$$F(x,\hat{\boldsymbol{\theta}}) = \hat{r}_0 G(x,t_0,t) + \hat{r}_1 H(x,\hat{\delta}_1,\hat{\sigma}_1) + \hat{r}_2 H(x,\hat{\delta}_2,\hat{\sigma}_2)$$

with the estimated rate parameters $\hat{r}_0 = 0.079$, $\hat{r}_1 = 0.786$, and $\hat{r}_2$=0.135. The probability density function of $F(x, \hat{\boldsymbol{\theta}})$ is drawn in Figure 3.4 (e). The target data for Case 14 are the data observed on January 18th, 2004, which is shown in Figure 3.4 (f). It is observed from Figure 3.4 (e) and (f) that the data can be modeled by the mixture model for Case 14. For this case, the p-value obtained by the parametric bootstrap is 0.984.

If we assume that the model can be used when the p-value is higher than 0.1, which implies that the goodness-of-fit test is not rejected with the significance level $\alpha = 0.1$, it is shown that Model 1 can be used for 8 cases, as shown in Table 3.4, and Model 2 can be used for other 7 cases, as shown in Table B.1 in Appendix B. The values with * in both tables denote that the values are not estimated because the estimates were close to 0 or 1 so that the values are fixed to 0 or 1 to make the model simple. For these 15 cases, the model can explain the effects of the changes in temperature and salinity of water caused by freshwater flows on the growth of banana prawns.

For Cases 2, 5, 6, and 16, on the other hand, the goodness-of-fit test is rejected for both models. From Figure 3.5, it might be because of the small number of observations and some outliers for Case 2. For Cases 5 and 6, there might be a large cohort constructed for some reason because much more prawns were caught on March 18th, 2003 compared to the data on February 15th and April 2nd, 2003, as shown in Figure 3.6 and Figure 3.7, respectively. For Case 16, there are some large prawns observed on February 16th compared to January 22nd, 2004 as shown in Figure 3.8, so there might be some reason to make the growth of banana prawns faster than the model we have applied to.



(a) 2003/12/23.

(b) $G(x,t_0,t)$.

(c) $H(x,\hat{\delta}_1,\hat{\sigma}_1)$.

(d) $H(x,\hat{\delta}_2,\hat{\sigma}_2)$.

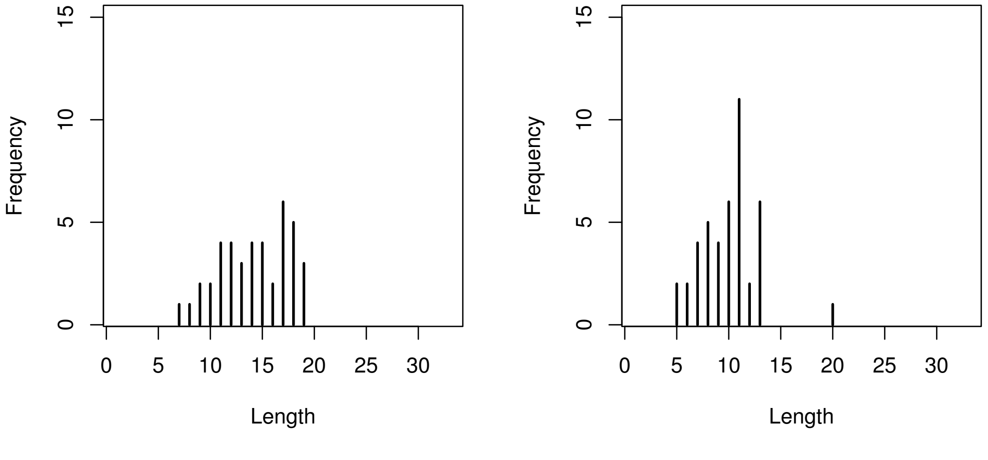
(e) $F(x,\hat{\boldsymbol{\theta}})$.

(f) 2004/01/18.

Figure 3.4: Result for Case 14. (a) and (f) are length-frequency data observed on 2003/12/23 and 2004/1/18. Others are probability density functions of distribution functions $G(x,t_0,t)$, $H(x,\hat{\delta}_1,\hat{\sigma}_1)$, $H(x,\hat{\delta}_2,\hat{\sigma}_2)$, and $F(x,\hat{\boldsymbol{\theta}})$, respectively.

Table 3.4: Parameters and results of goodness-of-fit test for Model 1.

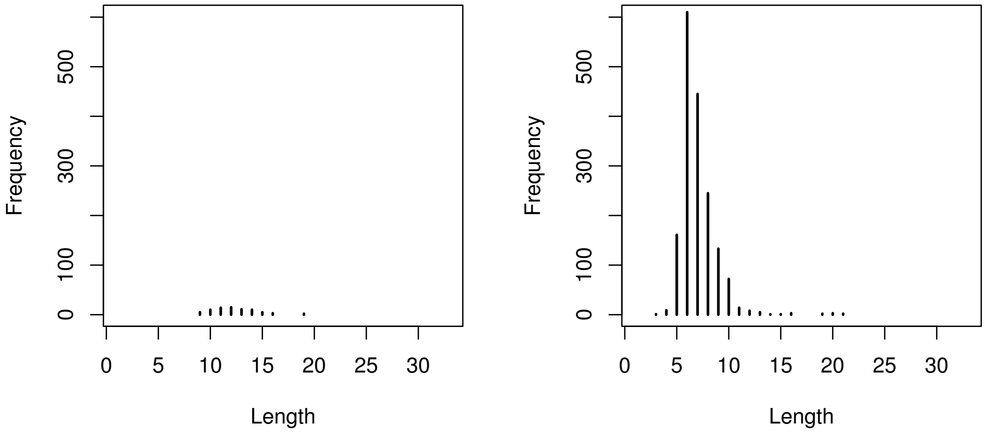
Case	$\hat{r}_0$ (SE)	$\hat{r}_1 = 1 - \hat{r}_0$	$\hat{\delta}_1$ (SE)	$\hat{\sigma}_1$ (SE)	$W_n^{(d)2}\left(\hat{\boldsymbol{\theta}}\right)$	p-value
1	0.314 (0.112)	0.686	11.045 (1.607)	3.828 (0.995)	0.040	0.186
4	0.165 (0.181)	0.835	8.888 (1.202)	2.765 (0.603)	0.005	0.944
8	0.000^*	1.000^*	6.788 (0.293)	4.169 (0.261)	0.043	0.360
9	0.155 (0.198)	0.845	6.023 (0.387)	3.779 (0.575)	0.015	0.774
10	0.836 (0.199)	0.164	5.444 (1.005)	1.112 (1.345)	0.066	0.254
11	0.410 (0.135)	0.590	7.092 (0.995)	1.523 (0.791)	0.008	0.880
13	0.073 (0.071)	0.927	10.346 (0.849)	3.246 (0.527)	0.025	0.522
15	0.000^*	1.000^*	4.351 (0.146)	3.454 (0.138)	0.022	0.734



(a) 2002/12/16.

(b) 2003/1/19.

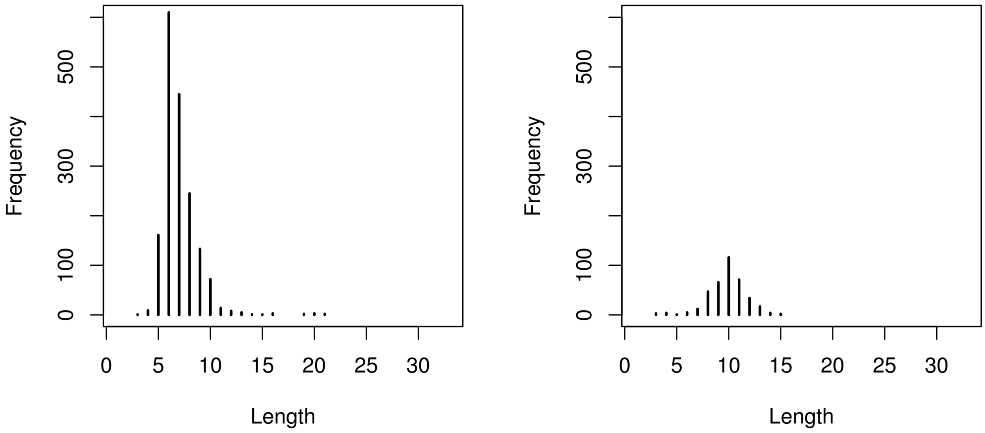
Figure 3.5: Length-frequency data for Case 2.



(a) 2003/2/15.

(b) 2003/3/18.

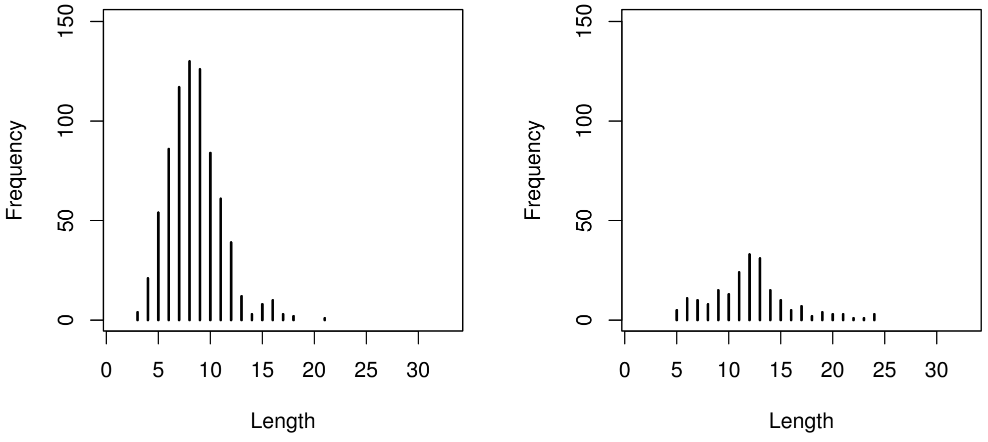
Figure 3.6: Length-frequency data for Case 5.



(a) 2003/3/18.

(b) 2003/4/2.

Figure 3.7: Length-frequency data for Case 6.



(a) 2004/1/22.

(b) 2004/2/16.

Figure 3.8: Length-frequency data for Case 16.

3.6 Concluding remarks

We have shown that a mixture of two probability distribution models, for a cohort stayed in the estuary and for a cohort migrated from offshore waters to the estuary, can be used for describing the distribution of the carapace length of banana prawns in the estuary. Since our model is an elaborated model to reflect the changes of the environmental factors, it makes easy to detect outlying cases where some unknown cause exists. On the contrary, the model can explain the effects of the changes in temperature and salinity of water caused by freshwater flows on the growth of banana prawns for the cases where the model can be used to describe the distribution of the carapace length. We hope that our results will be useful for further understanding of the length-frequency data.

From a statistical point of view, how the choice of parameter estimation affects the goodness-of-fit test would be an interesting problem for the case of discrete distributions as well as for the case of continuous distributions. In this chapter, the parameters are estimated by minimizing the Crámer-von Mises statistic for discrete distributions with some constrains. The reason we used the minimum distance estimator is that such an estimator chosen to minimize some distance is known to be robust to contamination, which we will explain in Section 4.5.1, and such property would be favorable to the situation where one wishes to give an approximation model of the data. We will investigate how the combination of the parameter estimation and the goodness-of-fit test works in the next chapter.

Chapter 4

Asymptotic behavior of the Cramér-von Mises statistic when contamination exists

In the two case studies, we have used the goodness-of-fit test to check whether the derived probability distribution model can be used or not. In this chapter, we investigate the asymptotic behavior of the Cramér-von Mises statistic when contamination exists because it often happens in practice that the data are contaminated.

4.1 Introduction

In this chapter, we assume that $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables from a distribution function $F_\varepsilon(x, \boldsymbol{\theta})$ and $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ are their order statistics. Here the distribution $F_\varepsilon(x, \boldsymbol{\theta})$ is contaminated as

$$F_\varepsilon(x, \boldsymbol{\theta}) = \left(1 - \frac{\varepsilon}{\sqrt{n}}\right) F(x, \boldsymbol{\theta}) + \frac{\varepsilon}{\sqrt{n}} G(x),$$

where $F(x, \boldsymbol{\theta})$ is a continuous distribution with a parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^\top \in \Theta \subset \mathbb{R}^m$, $G(x)$ is the distribution of the contamination, and $\varepsilon \geq 0$. We hereafter assume that both distributions $F(x, \boldsymbol{\theta})$ and $G(x)$ have bounded and smooth probability density functions $f(x, \boldsymbol{\theta})$ and $g(x)$, respectively.

In this chapter, we use

$$W_n^2(\boldsymbol{\theta}) = \sum_{j=1}^{n} \left\{ F\left(X_{(j)}, \boldsymbol{\theta}\right) - \frac{j}{n+1} \right\}^2$$

as the Cramér-von Mises statistic for simplicity. This definition is slightly different from the definition in (1.2), however, the asymptotic behaviors of the statistics for each definition are identical. Although the asymptotic behavior of $W_n^2(\boldsymbol{\theta})$ when no contamination exists has been thoroughly investigated as introduced in Section 1.2, only a few works have been performed for the case that contamination exists.

We first derive the asymptotic distribution of $W_n^2(\boldsymbol{\theta})$ via an elementary matrix calculation in Section 4.2.1. It follows from the result that the asymptotic distribution of $W_n^2(\boldsymbol{\theta})$ is given as a distribution of a weighted infinite sum of non-central chi-squared random variables with 1 degree of freedom and the effect of contamination appears only in the non-centralities. In Section 4.2.2, the result given in Section 4.2.1 is extended to the case where the parameters are estimated by the minimum distance method, which is the estimation method of finding the value $\hat{\boldsymbol{\theta}}$ which makes the Cramér-von Mises statistic a minimum. An approximation of the distribution of the statistic based on the result given in Section 4.2.2 is described in Section 4.3. Some remarks on the weights in the asymptotic distribution of the Cramér-von Mises statistic are given in Section 4.4. The robustness of the Cramér-von Mises goodness-of-fit test when the minimum distance estimator is used is investigated by extending the robustness of the estimator and demonstrated by numerical experiments in Section 4.5.

4.2 Asymptotic distribution of the Cramér-von Mises statistic

4.2.1 When the parameters are known

We rewrite $W_n^2(\boldsymbol{\theta})$ as $W_n^2(\boldsymbol{\theta}) = \|(n+1)S_n\boldsymbol{U}_n\|^2$ by introducing an $n \times (n+1)$ matrix

$$S_n = \left(\frac{1}{n+1}\left(1_{j\geq k} - \frac{j}{n+1}\right); 1 \leq j \leq n, 1 \leq k \leq n+1\right)$$

and an $n+1$-dimensional vector

$$\boldsymbol{U}_n = \left(F\left(X_{(j)}, \boldsymbol{\theta}\right) - F\left(X_{(j-1)}, \boldsymbol{\theta}\right); 1 \leq j \leq n+1\right)^\top.$$

Here we define $F\left(X_{(0)}, \boldsymbol{\theta}\right) = 0$ and $F\left(X_{(n+1)}, \boldsymbol{\theta}\right) = 1$ for convenience. We also define a diagonal matrix B with diagonal elements, $b_1 = b_2$ and

$$b_{j+1} = \frac{f(F_\varepsilon^{-1}(\frac{j}{n+1}, \boldsymbol{\theta}), \boldsymbol{\theta})}{f_\varepsilon(F_\varepsilon^{-1}(\frac{j}{n+1}, \boldsymbol{\theta}), \boldsymbol{\theta})}, \quad j = 1, 2, \ldots, n,$$

where $f_\varepsilon(x, \boldsymbol{\theta})$ and $F_\varepsilon^{-1}(u, \boldsymbol{\theta}) = x$ are the probability density function and the inverse function of $F_\varepsilon(x, \boldsymbol{\theta})$, respectively.

A Taylor expansion of

$$\begin{aligned} &F\left(X_{(j)}, \boldsymbol{\theta}\right) - F\left(X_{(j-1)}, \boldsymbol{\theta}\right) \\ &\qquad = F\left(F_\varepsilon^{-1}\left(F_\varepsilon\left(X_{(j)}, \boldsymbol{\theta}\right), \boldsymbol{\theta}\right), \boldsymbol{\theta}\right) - F\left(F_\varepsilon^{-1}\left(F_\varepsilon\left(X_{(j-1)}, \boldsymbol{\theta}\right), \boldsymbol{\theta}\right), \boldsymbol{\theta}\right) \end{aligned}$$

around $F_\varepsilon\left(X_{(j-1)}, \boldsymbol{\theta}\right)$, $j = 1, 2, \ldots, n+1$, yields an approximation of $\boldsymbol{U}_n$ as

$$B\left\{\boldsymbol{U}_n^* - \frac{1}{n+1}(\mathbf{1} - \boldsymbol{c}_n)\right\},$$

where

$$\boldsymbol{U}_n^* = \left(F_\varepsilon\left(X_{(j)}, \boldsymbol{\theta}\right) - F_\varepsilon\left(X_{(j-1)}, \boldsymbol{\theta}\right); 1 \leq j \leq n+1\right)^\top,$$

$\boldsymbol{c}_n$ is an $n+1$-dimensional vector where the first element is $(n+1)F(F_\varepsilon^{-1}(1/(n+1),\boldsymbol{\theta}),\boldsymbol{\theta})/b_1$ and all others are equal to 1, and $\mathbf{1}=(1,1,\dots,1)^\top$. Again we define $F_\varepsilon\left(X_{(0)},\boldsymbol{\theta}\right)=0$ and $F_\varepsilon\left(X_{(n+1)},\boldsymbol{\theta}\right)=1$ for convenience.

We now see that it is enough to know the distribution of

$$(\boldsymbol{V}_n+\boldsymbol{\mu}_n)^\top \Lambda_n (\boldsymbol{V}_n+\boldsymbol{\mu}_n)=\sum_{j=1}^{n}\lambda_{nj}(V_{nj}+\mu_{nj})^2 \tag{4.1}$$

instead of $W_n^2(\boldsymbol{\theta})$, where

$$\boldsymbol{V}_n=(V_{n1},V_{n2},\dots,V_{nn})^\top=(n+1)\Lambda_n{}^{-\frac{1}{2}}P_n^\top S_n B\left(\boldsymbol{U}_n^*-\frac{1}{n+1}\mathbf{1}\right)$$

and

$$\boldsymbol{\mu}_n=(\mu_{n1},\mu_{n2},\dots,\mu_{nn})^\top=\Lambda_n{}^{-\frac{1}{2}}P_n^\top S_n B\boldsymbol{c}_n.$$

Here Λ_n is a diagonal matrix of eigenvalues $\lambda_{n1}\geq\lambda_{n2}\geq\cdots\geq\lambda_{nn}$, and P_n is an orthogonal matrix of eigenvectors $\boldsymbol{p}_j^{(n)}$, $j=1,2,\dots$, of $S_nB^2S_n^\top$. The following proposition gives the limits of these eigenvalues and eigenvectors. It follows from this proposition that the eigenvalues and the eigenvectors become independent of the contamination in the limit.

Proposition 1. *For any fixed $j>0$, as n tends to infinity λ_{nj} converges to $\lambda_j=1/(j\pi)^2$ and $\sqrt{n}p_{\lceil nu\rceil j}^{(n)}$ converges to $f_j(u)=\sqrt{2}\sin(\pi ju)$ for $0<u<1$, which are the eigenvalues and the eigenfunctions of the integral equation*

$$\lambda f(u)=\int_0^1 \rho_0(u,v)f(v)dv,$$

where the kernel function $\rho_0(u,v)=\min(u,v)-uv$, $p_{kj}^{(n)}$ is the kth element of $\boldsymbol{p}_j^{(n)}$, and $\lceil x\rceil$ is the minimum integer which is greater than or equal to x.

Before giving the proof of Proposition 1, we will make sure of the convergence of the eigenvalues and the eigenvectors. The following lemma can be derived from the theorem on page 372 of Riesz and Sz.-Nagy (1990), which states that the eigenvalues and eigenfunctions of an integral equation are continuous with respect to the kernel function of the integral equation as far as the kernel function belongs to the space

$$L^2(\nu \times \nu) = \left\{ k(x,y); \int\int k^2(x,y) d\nu(x) d\nu(y) < \infty \right\}$$

with the norm $\|\cdot\|$, where ν is a sigma finite measure. We will use the following lemma in some of the proofs through this chapter.

Lemma 2. *Let λ_{nj} and $f_n^{(j)}(u)$ be the jth eigenvalue and eigenfunction of the integral equation*

$$\lambda f(u) = \int k_n(u,v) f(v) dv$$

and λ_j and $f^{(j)}(u)$ be the jth eigenvalue and eigenfunction of the integral equation

$$\lambda f(u) = \int k(u,v) f(v) dv.$$

If $k_n(x,y)$ is a compact operator and $\|k_n - k\|$ converges to 0 as n tends to infinity, then λ_{nj} converges to λ_j and $\left\| f_n^{(j)} - f^{(j)} \right\|$ converges to 0 as n tends to infinity for $j = 1, 2, \ldots$ when $f_n^{(j)}$ and $f^{(j)}$ are properly normalized, including their signs.

Proof of Proposition 1. We first rewrite the equation $\lambda_{nj} \boldsymbol{p}_j^{(n)} = S_n B^2 S_n{}^\top \boldsymbol{p}_j^{(n)}$ as the integral equation

$$\lambda_{nj} f_n^{(j)}(u) = \int k_n(u,v) f_n^{(j)}(v) dv$$

with the kernel function

$$k_n(u,v) = \frac{n}{(n+1)^2} \sum_{l=1}^{n+1} \left(1_{\lceil nu \rceil \geq l} - \frac{\lceil nu \rceil}{n+1} \right) b_l^2 \left(1_{\lceil nv \rceil \geq l} - \frac{\lceil nv \rceil}{n+1} \right)$$

and the eigenfunction $f_n^{(j)}(u) = \sqrt{n} p_{\lceil nu \rceil j}^{(n)}$. Noting that $k_n(u,v)$ can be approximated as

$$k^*(u,v) = \int_0^1 (1_{u \geq s} - u)(1_{v \geq s} - v) b(s)^2 ds,$$

where $b(s) = f(F_\varepsilon(s,\boldsymbol{\theta}),\boldsymbol{\theta})/f_\varepsilon(F_\varepsilon(s,\boldsymbol{\theta}),\boldsymbol{\theta})$, then the convergence of $k^*(u,v)$ to

$$\rho_0(u,v) = \min(u,v) - uv = \int_0^1 (1_{u\geq s} - u)(1_{v\geq s} - v)\,ds$$

is clear from Lebesgue's dominated convergence theorem because $0 \leq b(s) \leq (1-\varepsilon/\sqrt{n})^{-1}$ for any $0 \leq s \leq 1$ and $\varepsilon^2 < n$. The desired result follows from Lemma 2 because the eigenvalues and the eigenfunctions of the integral equation for the kernel $\rho_0(u,v)$ are $1/(j\pi)^2$ and $\sqrt{2}\sin(\pi ju)$, $j = 1,2,\ldots$. □

We also give the following proposition for the convergence of $\boldsymbol{V}_n$.

Proposition 2. *Any finite-dimensional random vector* $\left(V_{nj_1}, V_{nj_2}, \ldots, V_{nj_p}\right)^\top$ *converges in distribution to a normally distributed random vector* $\boldsymbol{V} = \left(V_{j_1}, V_{j_2}, \ldots, V_{j_p}\right)^\top$ *with mean* $\mathbf{0}$ *and variance* I_p *as n tends to infinity.*

Proof. We first note that the kth element of $\boldsymbol{U}_n^*$ can be replaced by $E_k/\sum_{j=1}^{n+1} E_j$, where E_j, $j = 1,2,\ldots,n+1$, are independent and identically distributed exponential random variables with mean 1. This is because of a property of order statistics of a sample from the standard uniform distribution, for example, LePage et al. (1981) used this property to prove the convergence of the normalized partial sums to a stable distribution. Let c_{j_lk} be the (j_l,k) element of $\Lambda_n^{-\frac{1}{2}} P_n^\top S_n B$. Then, it is enough to show that for any $\boldsymbol{t} = (t_1,t_2,\ldots,t_p) \in \mathbb{R}^p$,

$$\frac{n+1}{\sum_{j=1}^{n+1} E_j} \sum_{l=1}^{p} t_l \sum_{k=1}^{n+1} c_{j_lk}(E_k - 1) = \frac{n+1}{\sum_{j=1}^{n+1} E_j} \sum_{k=1}^{n+1} \left(\sum_{l=1}^{p} t_l c_{j_lk}\right)(E_k - 1)$$

converges to a normally distributed random variable $V = \boldsymbol{t}^\top \boldsymbol{V}$ with mean 0 and variance $\boldsymbol{t}^\top\boldsymbol{t}$. It is easily seen that the Lindeberg condition for the central limit theorem is satisfied for

$$V_n = \sum_{k=1}^{n+1} \left(\sum_{l=1}^{p} t_l c_{j_lk}\right)(E_k - 1).$$

In fact, the inequality

$$\frac{1}{\boldsymbol{t}^\top\boldsymbol{t}} \sum_{k=1}^{n+1} \left(\sum_{l=1}^{p} t_l c_{j_lk}\right)^2 \mathrm{E}\left\{(E_k - 1)^2 1_{\left|\left(\sum_{l=1}^{p} t_l c_{j_lk}\right)(E_k - 1)\right| > \varepsilon\sqrt{\boldsymbol{t}^\top\boldsymbol{t}}}\right\} \leq H\left(\frac{\varepsilon}{\sqrt{p}c_n}\right)$$

implies the desired result, where the function

$$H(x) = \mathrm{E}\left\{(E_k - 1)^2 1_{|E_k-1|>x}\right\}$$

is a monotone decreasing to 0 as x increases and

$$c_n = \max_{1\le l\le p, 1\le k\le n+1} \left|c_{j_l k}\right| \le \max_{1\le l\le p, 1\le k\le n+1} \frac{b_k \lambda_{nj_l}^{-\frac{1}{2}}}{n+1} \left\{\sum_{l'=1}^{n} \left(1_{l'\ge k} - \frac{l'}{n+1}\right)^2\right\}^{\frac{1}{2}}.$$

□

Combining the results in Proposition 1 and Proposition 2, we see that $\sum_{j=1}^{p} \lambda_{nj}(V_{nj} + \mu_{nj})^2$ converges to $\sum_{j=1}^{p} \lambda_j (V_j + \mu_j)^2$, where $V_1, V_2, \ldots, V_p$ follow the standard normal distribution and

$$\mu_j = \varepsilon \lambda_j^{-\frac{1}{2}} \int_0^1 \int_0^1 f_j(u)(1_{u\ge v} - u)\left\{1 - \frac{g(F^{-1}(v,\boldsymbol{\theta}))}{f(F^{-1}(v,\boldsymbol{\theta}),\boldsymbol{\theta})}\right\} du dv, \quad j = 1, 2, \ldots.$$

On the other hand, $\sum_{j=1}^{n} \lambda_{nj}\mu_{nj}^2$ converges to $\sum_{j=1}^{\infty} \lambda_j \mu_j^2$ because of the boundedness of $\sum_{j=1}^{n} \lambda_{nj}\mu_{nj}^2$. We therefore have the following theorem. It follows from this theorem that the contamination affects only the non-centralities μ_j, $j = 1, 2, \ldots$, with the proportion ε.

Theorem 6. *$W_n^2(\boldsymbol{\theta})$ converges to $\sum_{j=1}^{\infty} \lambda_j (V_j + \mu_j)^2$ in distribution as n tends to infinity, where $V_1, V_2, \ldots$ are independent and identically distributed random variables with the standard normal distribution.*

The asymptotic distribution of $W_n^2(\boldsymbol{\theta})$ for the case that no contamination exists, which we have introduced in Section 1.2 as a known result, can be reduced from Theorem 6. We note that our derivation of the asymptotic distribution is different from others, for example Darling (1955) and Shorack and Wellner (1986), since many of the results are derived as an application of the theory of the empirical processes $\sqrt{n}\{F_n(x) - F(x,\boldsymbol{\theta})\}$.

A closely related result to Theorem 6 is given by Guttorp and Lockhart (1988). They developed a general theory for an asymptotic distribution of quadratic forms of order statistics from a uniform distribution under contiguous alternatives, where densities are of the form $1+\delta\eta(u)/n^{\frac{1}{2}}$ under the condition $\int_0^1 \eta(u)^2 du = 1$. If we consider the case $\eta(F(x,\boldsymbol{\theta})) = g(x)/f(x,\boldsymbol{\theta}) - 1$, the same result as in Theorem 6 can be derived from their theory. However, a major difference is that Theorem 6 is free from the constraint $\int_0^1 \eta(u)^2 du < \infty$. We have used only the fact that $f(x,\boldsymbol{\theta})$ and $g(x)$ are probability density functions in the proof.

4.2.2 When the parameters are estimated by the minimum distance method

We hereafter assume the followings in order to derive the asymptotic distribution of $W_n^2\left(\hat{\boldsymbol{\theta}}\right)$, the Cramér-von Mises statistic when the parameters are estimated by the minimum distance method, which is the estimation method of finding the value $\hat{\boldsymbol{\theta}}$ which makes the Cramér-von Mises statistic a minimum.

Assumption 1 (Identifiability)**.**

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} \{F(x,\boldsymbol{\theta}_n) - F(x,\boldsymbol{\theta})\}^2 \, dF(x,\boldsymbol{\theta}) = 0$$

implies the convergence of $\boldsymbol{\theta}_n$ to $\boldsymbol{\theta}$ in Θ.

Assumption 2 (Regularity)**.**

- *$F(x,\boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$.*
- *$g_k(u,\boldsymbol{\theta}) = \frac{\partial}{\partial\theta_k}F(x,\boldsymbol{\theta})$, $k = 1,2,\ldots,m$, are all continuous and square-integrable with respect to $u \in (0,1)$, where $u = F(x,\boldsymbol{\theta})$.*
- *The matrix $A = \left(\int_0^1 g_j(u,\boldsymbol{\theta})g_k(u,\boldsymbol{\theta})du; 1 \le j,k \le m\right)$ is of full rank.*
- *$\frac{\partial}{\partial\theta_k}g_j(u,\boldsymbol{\theta})$, $j,k = 1,2,\ldots,m$ exist and are continuous for $u \in (0,1)$.*
- *$\sup_{0<u<1}\left|\frac{\partial}{\partial\theta_l}g_k(u,\boldsymbol{\theta})\right| < \infty$ for any $k,l = 1,2,\ldots,m$.*
- *$\sup_{0<u<1}\left|\frac{\partial}{\partial u}g_k(u,\boldsymbol{\theta})\right| < \infty$ for any $k = 1,2,\ldots,m$.*

Let $\bar{\lambda}_{n1}(\boldsymbol{\theta}) \geq \bar{\lambda}_{n2}(\boldsymbol{\theta}) \geq \cdots \geq \bar{\lambda}_{n\,n-m}(\boldsymbol{\theta})$ and $\boldsymbol{q}_1^{(n)}, \boldsymbol{q}_2^{(n)}, \ldots, \boldsymbol{q}_{n-m}^{(n)}$ be the eigenvalues and the eigenvectors of the matrix

$$D_n(\boldsymbol{\theta}) = \Lambda_n^{\frac{1}{2}} \left[I - P_n^\top Z_n(\boldsymbol{\theta}) \left\{ Z_n(\boldsymbol{\theta})^\top Z_n(\boldsymbol{\theta}) \right\}^{-1} Z_n(\boldsymbol{\theta})^\top P_n \right] \Lambda_n^{\frac{1}{2}},$$

where

$$Z_n(\boldsymbol{\theta}) = \left(\frac{1}{\sqrt{n}} g_k \left(\frac{j}{n+1}, \boldsymbol{\theta} \right); 1 \leq j \leq n,\ 1 \leq k \leq m \right).$$

Without loss of generality, we may assume that $Z_n(\boldsymbol{\theta})$ is of full rank for each n, in view of the third assumption in Assumption 2.

Since the minimum distance estimator $\hat{\boldsymbol{\theta}}$ is the solution of

$$\left. \sum_{j=1}^{n} \left\{ F(X_{(j)}, \boldsymbol{\theta}) - \frac{j}{n+1} \right\} \boldsymbol{g}\left(F\left(X_{(j)}, \boldsymbol{\theta} \right), \boldsymbol{\theta} \right) \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} = 0$$

for $\boldsymbol{g}(u, \boldsymbol{\theta}) = (g_1(u, \boldsymbol{\theta}), g_2(u, \boldsymbol{\theta}), \ldots, g_m(u, \boldsymbol{\theta}))^\top$, a Taylor expansion of the left hand side around $\boldsymbol{\theta}$ gives an approximation of $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$ as

$$\begin{aligned} - \Bigg[\frac{1}{n} \sum_{j=1}^{n} \boldsymbol{g}\left(F\left(X_{(j)}, \boldsymbol{\theta} \right), \boldsymbol{\theta} \right) \boldsymbol{g}\left(F\left(X_{(j)}, \boldsymbol{\theta} \right), \boldsymbol{\theta} \right)^\top & \\ + \frac{1}{n} \sum_{j=1}^{n} \left\{ F(X_{(j)}, \boldsymbol{\theta}) - \frac{j}{n+1} \right\} \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{g}\left(F\left(X_{(j)}, \boldsymbol{\theta} \right), \boldsymbol{\theta} \right)^\top \Bigg]^{-1} & \\ \times \left[\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left\{ F(X_{(j)}, \boldsymbol{\theta}) - \frac{j}{n+1} \right\} \boldsymbol{g}\left(F\left(X_{(j)}, \boldsymbol{\theta} \right), \boldsymbol{\theta} \right) \right]. & \end{aligned}$$

Here we have

$$\lim_{n \to \infty} \mathrm{E} \left[\sum_{j=1}^{n} \left\{ F(X_{(j)}, \boldsymbol{\theta}) - \frac{j}{n+1} \right\}^2 \right] = \sum_{j=1}^{\infty} \lambda_j \mu_j^2 < \infty,$$

thus $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$ can be approximated as

$$\begin{aligned} -\sqrt{n} \left\{ \sum_{j=1}^{n} \boldsymbol{g}\left(F\left(X_{(j)}, \boldsymbol{\theta} \right), \boldsymbol{\theta} \right) \boldsymbol{g}\left(F\left(X_{(j)}, \boldsymbol{\theta} \right), \boldsymbol{\theta} \right)^\top \right\}^{-1} & \\ \times \left[\sum_{j=1}^{n} \left\{ F(X_{(j)}, \boldsymbol{\theta}) - \frac{j}{n+1} \right\} \boldsymbol{g}\left(F\left(X_{(j)}, \boldsymbol{\theta} \right), \boldsymbol{\theta} \right) \right]. & \qquad (4.2) \end{aligned}$$

Moreover, it is further approximated as

$$-(n+1)\left\{Z_n(\boldsymbol{\theta})^\top Z_n(\boldsymbol{\theta})\right\}^{-1} Z_n(\boldsymbol{\theta})^\top S_n \boldsymbol{U}_n \tag{4.3}$$

because it can be shown that $\boldsymbol{g}\left(F\left(X_{(j)},\boldsymbol{\theta}\right),\boldsymbol{\theta}\right)$ in (4.2) can be replaced by $\boldsymbol{g}\left(j/(n+1),\boldsymbol{\theta}\right)$ for $j = 1,2,\ldots,n$ in the approximation after the tedious calculation. Also $W_n^2\left(\hat{\boldsymbol{\theta}}\right)$ is asymptotically equivalent to

$$W_n^2(\boldsymbol{\theta}) - n\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^\top \left\{Z_n(\boldsymbol{\theta})^\top Z_n(\boldsymbol{\theta})\right\}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right),$$

therefore, we see from (4.1) and (4.3) that it is enough to derive the asymptotic distribution of

$$\left(\boldsymbol{V}_n+\boldsymbol{\mu}_n\right)^\top D_n(\boldsymbol{\theta})\left(\boldsymbol{V}_n+\boldsymbol{\mu}_n\right) = \sum_{j=1}^{n-m} \bar{\lambda}_{nj}(\boldsymbol{\theta})\left\{Y_{nj}+\bar{\mu}_{nj}(\boldsymbol{\theta})\right\}^2$$

instead of $W_n^2\left(\hat{\boldsymbol{\theta}}\right)$, where $Y_{nj} = \boldsymbol{q}_j^{(n)\top}\boldsymbol{V}_n$ and $\bar{\mu}_{nj}(\boldsymbol{\theta}) = \boldsymbol{q}_j^{(n)\top}\boldsymbol{\mu}_n$, $j = 1,2,\ldots,n-m$. The following propositions show the limits of the eigenvalues and the eigenvectors of the matrix $D_n(\boldsymbol{\theta})$ and the convergence of Y_{nj}, $j = 1,2,\ldots,n-m$.

Proposition 3. *For any fixed $j>0$, $\bar{\lambda}_{nj}(\boldsymbol{\theta})$ converges to $\bar{\lambda}_j$ and $\boldsymbol{q}_j^{(n)}$ converges to $\boldsymbol{q}_j$, where $\bar{\lambda}_j$ and $\boldsymbol{q}_j = (q_{1j},q_{2j},\ldots)^\top$ are the jth eigenvalue and eigenvector of the infinite-dimensional matrix*

$$D_\infty(\boldsymbol{\theta}) = \left(\lambda_j^{\frac{1}{2}}\lambda_k^{\frac{1}{2}}\int_0^1\int_0^1 f_j(u)h(u,v)f_k(v)dudv; 1 \le j,k < \infty\right), \tag{4.4}$$

where $h(u,v) = \delta(u-v) - \boldsymbol{g}(u,\boldsymbol{\theta})^\top A^{-1}\boldsymbol{g}(v,\boldsymbol{\theta})$ with the Dirac delta function $\delta(u)$.

Proof. By taking ν as a counting measure in Lemma 2, $D_n(\boldsymbol{\theta})$ and $D_\infty(\boldsymbol{\theta})$ can be considered as compact operators on $L^2(\nu)$. To evaluate $\|D_n(\boldsymbol{\theta})-D_\infty(\boldsymbol{\theta})\|$, we first note that

$$\begin{aligned}&\|D_n(\boldsymbol{\theta})-D_\infty(\boldsymbol{\theta})\|\\ &\le \sum_{j=1}^{p}\sum_{k=1}^{p}\left\{d_{jk}^{(n)}-d_{jk}\right\}^2 + 2\sum_{j=1}^{n}\sum_{k=p+1}^{n}\left\{d_{jk}^{(n)}-d_{jk}\right\}^2 + 2\sum_{j=1}^{\infty}\sum_{k=n+1}^{\infty} d_{jk}^2,\end{aligned} \tag{4.5}$$

where $d_{jk}^{(n)}$ and d_{jk} are the (j,k) element of $D_n(\boldsymbol{\theta})$ and of $D_\infty(\boldsymbol{\theta})$, respectively. The first term of (4.5) converges to 0 for any fixed p, since $d_{jk}^{(n)}$ converges to d_{jk} for fixed j and k. The convergence of the second term of (4.5) to 0 can be shown as follows. Since the matrix

$$I - P_n^\top Z_n(\boldsymbol{\theta})\left\{Z_n(\boldsymbol{\theta})^\top Z_n(\boldsymbol{\theta})\right\}^{-1} Z_n(\boldsymbol{\theta})^\top P_n$$

is a projection, we have $\left|d_{jk}^{(n)}\right| \leq \lambda_{nj}^{\frac{1}{2}}\lambda_{kj}^{\frac{1}{2}}$, so that

$$\sum_{k=p+1}^{n}\left\{\sum_{j=1}^{n} {d_{jk}^{(n)}}^2 + \sum_{j=1}^{n} d_{jk}^2\right\} \leq \left(\sum_{j=1}^{n}\lambda_{nj}\right)\sum_{k=p+1}^{n}\lambda_{nk} + \alpha\left(\sum_{j=1}^{n}\lambda_j\right)\sum_{k=p+1}^{n}\lambda_k, \tag{4.6}$$

where $\alpha = \max_{j,k}\lambda_j^{-\frac{1}{2}}\lambda_k^{-\frac{1}{2}}\left|d_{jk}\right|$. Here $\sum_{j=1}^{\infty}\lambda_j = \pi^{-2}\sum_{j=1}^{\infty} j^{-2} = 1/6$ and

$$\sum_{j=1}^{n}\lambda_{nj} = \text{trace}\left(S_n B^2 S_n^\top\right) = \sum_{j=1}^{n}\sum_{k=1}^{n+1}\left(1_{j\geq k} - \frac{j}{n+1}\right)^2 b_k^2 \leq \frac{1}{6}\left(1 - \frac{\varepsilon}{\sqrt{n}}\right)^{-2},$$

because $0 \leq f(x,\boldsymbol{\theta})/f_\varepsilon(x,\boldsymbol{\theta}) \leq (1-\varepsilon/\sqrt{n})^{-1}$ for any x and $\varepsilon^2 < n$. Therefore, by taking a large enough value of p, the right hand side of (4.6) will be sufficiently small. The convergence of the last term of (4.5) is clear from the inequality

$$\sum_{j=1}^{\infty}\sum_{k=n+1}^{\infty} d_{jk}^2 \leq \frac{1}{6}\alpha^2\left(\sum_{k=n+1}^{\infty}\lambda_k\right).$$

Since $D_n(\boldsymbol{\theta})$ is a compact operator, the proof is complete from Lemma 2.

□

Proposition 4. *Any finite-dimensional random vector* $\left(Y_{nj_1}, Y_{nj_2}, \ldots, Y_{nj_p}\right)$ *converges in distribution to a normally distributed random vector* $\left(Y_{j_1}, Y_{j_2}, \ldots, Y_{j_p}\right)$ *with mean* $\mathbf{0}$ *and variance* I_p *as n tends to infinity.*

Proof. Using a similar argument to that given in the proof of Proposition 2, it is enough to show that

$$\max_{1\leq m\leq p, 1\leq k\leq n}\left|\sum_{l=1}^{n} q_{lm}^{(n)} c_{j_l k}\right|$$

converges to 0 as n tends to infinity. We have

$$\left|\sum_{l=1}^{n} q_{lm}^{(n)} c_{j_l k}\right| \leq \left|\sum_{l=1}^{p'} q_{lm}^{(n)} c_{j_l k}\right| + \left\{\sum_{l=p'+1}^{n} q_{lm}^{(n)2}\right\}^{\frac{1}{2}} \left(\sum_{l=p'+1}^{n} c_{j_l k}^{2}\right)^{\frac{1}{2}} \tag{4.7}$$

and see that the first term on the right hand side of (4.7) converges to 0 as n tends to infinity for any $p' < n$ from the fact in the proof of Proposition 2 that $c_n = \max_{1 \leq l \leq p, 1 \leq k \leq n} \left|c_{j_l k}\right|$ converges to 0 as n tends to infinity. Noting that $\sum_{l=1}^{n} q_{lm}^{(n)2} = 1$ for any m, $\sum_{l=1}^{p} c_{j_l k}^{2} \leq \sum_{l=1}^{n} c_{lk}^{2} \leq 1$ for any k and Lemma 2, the proof is complete. □

From the fact that $\sum_{j=1}^{n-m} \bar{\lambda}_{nj}(\boldsymbol{\theta}) \bar{\mu}_{nj}^{2}(\boldsymbol{\theta}) \leq \sum_{j=1}^{n} \lambda_{nj} \mu_{nj}^{2}$, we have the following theorem by using the similar argument to that for Theorem 6.

Theorem 7. *$W_n^2\left(\hat{\boldsymbol{\theta}}\right)$ converges to $\sum_{j=1}^{\infty} \bar{\lambda}_j \left(Y_j + \bar{\mu}_j\right)^2$ in distribution as n tends to infinity, where $Y_1, Y_2, \ldots$ are independent and identically distributed random variables with the standard normal distribution and*

$$\bar{\mu}_j = \sum_{l=1}^{\infty} q_{lj} \mu_l, \quad j = 1, 2, \ldots.$$

Here μ_l, $l = 1, 2, \ldots$, are defined in Theorem 6.

4.3 An approximation of the distribution of the Cramér-von Mises statistic

The derivations of Theorem 6 and Theorem 7 suggest a good way of the approximations of the distributions of $W_n^2(\boldsymbol{\theta})$ and $W_n^2\left(\hat{\boldsymbol{\theta}}\right)$. For the distribution of $W_n^2(\boldsymbol{\theta})$, it follows from Theorem 6 that a distribution of a weighted finite sum of non-central chi-squared random variables with 1 degree of freedom $\sum_{j=1}^{p} \lambda_{nj} \left(V_j + \mu_{nj}\right)^2$ would give a good approximation of the distribution of $W_n^2(\boldsymbol{\theta})$ for an appropriate choice of $p \leq n$, where λ_{nj} and μ_{nj} are obtained from the eigenvalues and the eigenvectors of $S_n B^2 S_n^{\top}$.

Similarly, for the distribution of $W_n^2(\hat{\boldsymbol{\theta}})$, it follows from Theorem 7 that a distribution of a weighted finite sum of non-central chi-squared random variables with 1 degree of freedom $\sum_{j=1}^{p}\bar{\lambda}_{nj}(\hat{\boldsymbol{\theta}})\left\{Y_j+\bar{\mu}_{nj}\left(\hat{\boldsymbol{\theta}}\right)\right\}^2$ would give a good approximation of the distribution of $W_n^2(\hat{\boldsymbol{\theta}})$, where $\bar{\lambda}_{nj}\left(\hat{\boldsymbol{\theta}}\right)$ and $\bar{\mu}_{nj}\left(\hat{\boldsymbol{\theta}}\right)$ are obtained from the eigenvalues and the eigenvectors of $D_n\left(\hat{\boldsymbol{\theta}}\right)$. The replacement of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$ in the calculation is justified by the consistency of $\hat{\boldsymbol{\theta}}$ irrespective of the existence of contamination. In fact,

$$\sum_{j=1}^{n-m}\bar{\lambda}_{nj}(\hat{\boldsymbol{\theta}})\left\{Y_j+\bar{\mu}_j\left(\hat{\boldsymbol{\theta}}\right)\right\}^2-\sum_{j=1}^{n-m}\bar{\lambda}_{nj}(\boldsymbol{\theta})\left\{Y_j+\bar{\mu}_j\left(\boldsymbol{\theta}\right)\right\}^2$$

converges to 0 in probability as n tends to infinity. It follows from the strong consistency of $\hat{\boldsymbol{\theta}}$ given by Woodward et al. (1984) and Lemma 2.

Example 1

An example for the approximation of the distribution function of $W_n^2\left(\hat{\boldsymbol{\theta}}\right)$ by that of $\sum_{j=1}^{n-m}\bar{\lambda}_{nj}\left(\hat{\boldsymbol{\theta}}\right)\left\{Y_j+\bar{\mu}_{nj}\left(\hat{\boldsymbol{\theta}}\right)\right\}^2$ is shown in Figure 4.1. The distribution $F(x,\boldsymbol{\theta})$ is the exponential distribution with mean 0 and the distribution of the contamination $G(x)$ is the normal distribution with mean 7 and variance 1 in this experiment. The sample size is $n=100$. In Figure 4.1 the broken lines stand for the distribution of $\sum_{j=1}^{n-m}\bar{\lambda}_{nj}\left(\hat{\boldsymbol{\theta}}\right)\left\{Y_j+\bar{\mu}_{nj}\left(\hat{\boldsymbol{\theta}}\right)\right\}^2$. The R function “imhof ” developed by Duchesne and De Micheaux (2010) is used for calculating the probability of the distribution of the weighted sum of non-central chi-squared random variables with 1 degree of freedom. The solid lines stand for the distributions of of $W_n^2\left(\hat{\boldsymbol{\theta}}\right)$ obtained from 30,000 times random number experiments. Three gray scales, black, dark, and light, are used for indicating different rates of contamination, $\varepsilon=0,0.25,0.5$, respectively. Figure 4.1 shows that the approximation works fine even when $n=100$. The figure also shows that the distribution slightly shifts toward the right as ε increases. Such an insensitivity of $W_n^2\left(\hat{\boldsymbol{\theta}}\right)$ will lead us the robustness of the test shown in Section 4.5.

Practical procedure to obtain the weights for the goodness-of-fit test

In the Cramér-von Mises goodness-of-fit test, we only need to obtain the weights $\bar{\lambda}_{nj}\left(\hat{\boldsymbol{\theta}}\right)$, $j=1,2,\ldots$, because the distribution of $W_n\left(\hat{\boldsymbol{\theta}}\right)$ when no contamination

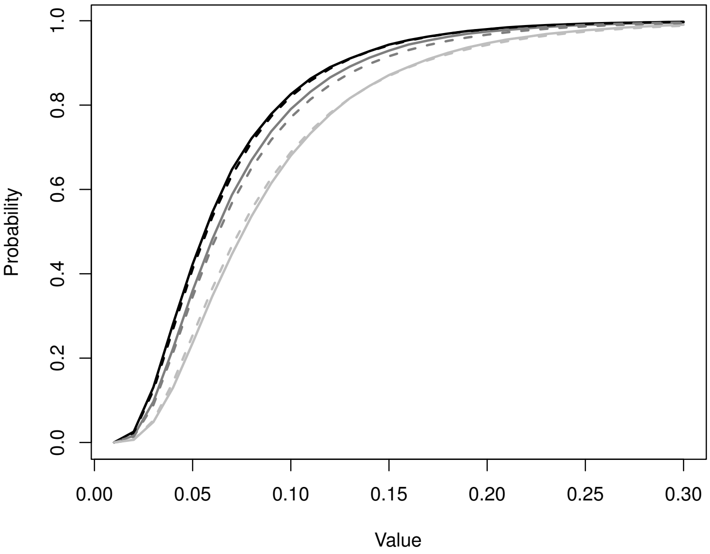


Figure 4.1: Distribution functions of $W_n^2\left(\hat{\boldsymbol{\theta}}\right)$ when the observations are contaminated.

exists can be approximated as $\sum_{j=1}^{n-m} \bar{\lambda}_{nj}\left(\hat{\boldsymbol{\theta}}\right) Y_j^2$. Noting that $B = I$ when no contamination exists, a practical procedure for obtaining $\left\{\bar{\lambda}_{nj}\left(\hat{\boldsymbol{\theta}}\right)\right\}$ would be

1. Obtain the eigenvalues and the eigenvectors of $S_n S_n^\top$ to make Λ_n and P_n by a singular value decomposition of S_n.

2. Find the eigenvalues of $\left\{I - U\left(\hat{\boldsymbol{\theta}}\right) U\left(\hat{\boldsymbol{\theta}}\right)^\top\right\}\Lambda_n$, where $U(\hat{\boldsymbol{\theta}})$ is an orthogonal matrix obtained by a singular value decomposition of $P_n Z_n(\hat{\boldsymbol{\theta}})$ as $P_n Z_n(\hat{\boldsymbol{\theta}}) = U\left(\hat{\boldsymbol{\theta}}\right) D\left(\hat{\boldsymbol{\theta}}\right) V\left(\hat{\boldsymbol{\theta}}\right)^\top$.

3. The squared values of the eigenvalues obtained in 2. are $\left\{\bar{\lambda}_{nj}\left(\hat{\boldsymbol{\theta}}\right)\right\}$.

Note that it is enough to obtain Λ_n and P_n only once, since those matrices are solely determined from the constant matrix S_n.

Example 2

Here we demonstrate the validity of the approximation through the critical values for the Cramér-von Mises goodness-of-fit test when the parameters are estimated by the minimum distance method. As a simple example, consider the gamma distribution with shape $\nu > 0$ and scale $a > 0$. The probability density function is

$$f(x,\boldsymbol{\theta}) = \frac{1}{a\Gamma(\nu)}\left(\frac{x}{a}\right)^{\nu-1}\exp\left(-\frac{x}{a}\right)$$

with $\boldsymbol{\theta} = (\nu, a)$. The elements of $Z_n(\boldsymbol{\theta})$ is calculated by using the formulas

$$g_1(u,\boldsymbol{\theta}) = -u\psi(\nu) - u\log a + \int_0^{F^{-1}(u,\boldsymbol{\theta})} f(x,\boldsymbol{\theta})\log x\,dx$$

and

$$g_2(u,\boldsymbol{\theta}) = -\frac{F^{-1}(u,\boldsymbol{\theta})}{a} f(F^{-1}(u,\boldsymbol{\theta}),\boldsymbol{\theta})$$

with the digamma function $\psi(\nu) = \frac{d}{d\nu}\log\Gamma(\nu)$. We have performed 30,000 times random number simulations for the case $\nu = 2$ and $a = 1$. For each sample, the parameters are estimated by the minimum distance method and the critical value for the significance level α is obtained from the distribution of $\sum_{j=1}^{n-m}\overline{\lambda}_{nj}\left(\hat{\boldsymbol{\theta}}\right)Y_j^2$. The R function "qchiapprox" developed by Tong et al. (2010) is used for obtaining the critical value of the distribution of the weighted sum of chi-squared random variables with 1 degree of freedom. Table 4.1 shows the proportion of acceptance of the null hypothesis in the 30,000 times random number simulations for each α and $n = 50, 100, 150, 200$. It shows that the proportion is close to the α as far as $\alpha \geq 0.8$ even if $n = 50$.

Table 4.1: Validation of the approximation in the case of the gamma distribution.

$n\backslash\alpha$	0.8	0.85	0.9	0.95	0.99
50	0.828	0.872	0.918	0.960	0.992
100	0.814	0.862	0.909	0.956	0.991
150	0.810	0.859	0.906	0.955	0.991
200	0.808	0.857	0.905	0.954	0.990

4.4 Equivalence of the weights in the asymptotic distribution of $W_n\left(\hat{\boldsymbol{\theta}}\right)$

In this section, we show that the weights $\{\bar{\lambda}_j\}$ in the asymptotic distribution of $W_n\left(\hat{\boldsymbol{\theta}}\right)$ given in Theorem 7 can be obtained as the eigenvalues of two different integral equations. This equivalence implies that the weights can be obtained by solving the infinite-dimensional matrix or the integral equations. In our result, the infinite-dimensional matrix is derived as the limit of the finite-dimensional matrix. Therefore, it is a natural way to use the eigenvalues of the finite-dimensional matrix as an approximation of the weights, which we have demonstrated in the previous section.

As we have introduced in Section 1.2, the asymptotic distribution of the Cramér-von Mises statistic when the parameters are estimated by a general estimation method and no contamination exists is well known. Along with the known result, the asymptotic distribution of the statistic when the parameters are estimated by the minimum distance method is given as a distribution of a infinite weighted sum of chi-squared random variables with 1 degree of freedom, where the weights are the eigenvalues of the integral equation

$$\lambda f(u) = \int_0^1 \rho(u,v) f(v) dv \tag{4.8}$$

with the kernel function

$$\rho(u,v) = \rho_0(u,v) - \boldsymbol{g}(u,\boldsymbol{\theta})^\top \boldsymbol{h}(v) - \boldsymbol{h}(u)^\top \boldsymbol{g}(v,\boldsymbol{\theta}) + \boldsymbol{g}(u,\boldsymbol{\theta})^\top \Sigma\, \boldsymbol{g}(v,\boldsymbol{\theta}). \tag{4.9}$$

Here $\rho_0(u,v)$ is given in Proposition 1,

$$\boldsymbol{h}(u) = \lim_{n\to\infty} \mathrm{E}\left[\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n} 1_{F(X_i,\boldsymbol{\theta})<u} - u\right\}\left\{\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\right\}\right],$$

and

$$\Sigma = \lim_{n\to\infty} n\mathrm{E}\left\{\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^\top\right\}.$$

For example, Beran (1984) gave an explicit representation of $\rho(u,v)$ in the case of the minimum distance estimator.

We first give the following lemma to prove that the eigenvalues $\{\bar{\lambda}_j\}$ of the matrix $D_\infty(\boldsymbol{\theta})$ in Proposition 3 are also those of the integral equation (4.8).

Lemma 3.

$$\rho(s,t) = \int_0^1 \int_0^1 h(s,u)h(t,v)\rho_0(u,v)dudv,$$

where $h(u,v)$ is given in Proposition 3.

Proof. We first note that

$$\boldsymbol{h}(u) = \mathrm{E}\left\{w(u)A^{-1}\int_0^1 w(v)\boldsymbol{g}(v,\boldsymbol{\theta})dv\right\} = A^{-1}\int_0^1 \rho_0(u,v)\boldsymbol{g}(v,\boldsymbol{\theta})dv,$$

where $u = F(x,\boldsymbol{\theta})$ and $w(u)$ is a Brownian bridge and the limit of the empirical process $\sqrt{n}\{F_n(x) - F(x,\boldsymbol{\theta})\}$. Next we see that

$$\begin{aligned}\Sigma &= A^{-1}\left[\int_0^1\int_0^1 \mathrm{E}\{w(u)w(v)\}\boldsymbol{g}(u,\boldsymbol{\theta})\boldsymbol{g}(v,\boldsymbol{\theta})^\top dudv\right]A^{-1}\\ &= A^{-1}\left\{\int_0^1\int_0^1 \rho_0(u,v)\boldsymbol{g}(u,\boldsymbol{\theta})\boldsymbol{g}(v,\boldsymbol{\theta})^\top dudv\right\}A^{-1}\end{aligned}$$

because $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$ converges to $A^{-1}\int_0^1 w(u)\boldsymbol{g}(u,\boldsymbol{\theta})du$. The desired result follows from (4.9) and the representations of $\boldsymbol{h}(u)$ and Σ given above. □

Proposition 5. $\{\bar{\lambda}_j\}$ *are also the eigenvalues of the integral equation (4.8).*

Proof. It is easily verified that the function

$$f(u) = \int_0^1 h(u,v)\left\{\sum_{k=1}^{\infty}\lambda_k^{\frac{1}{2}} f_k(v)q_k^{(j)}\right\}dv$$

is the solution of (4.8) for $\lambda = \bar{\lambda}_j$ from Lemma 3 and the fact that $\rho_0(u,v)$ can be written by using $f_j(u), j = 1,2,\ldots,$ in Proposition 1 as

$$\rho_0(u,v) = \sum_{j=1}^{\infty}\lambda_j f_j(u)f_j(v). \tag{4.10}$$

For the proof of the converse, we first note that the kernel function $\rho(u,v)$ is a compact operator because it is bounded and continuous. The integral equation (4.8) thus has only discrete bounded spectra, $\check{\lambda}_1 \geq \check{\lambda}_2 \geq \ldots$. Denote the corresponding eigenfunctions as $\check{f}_1(u), \check{f}_2(u), \ldots$, then it becomes clear that the infinite-dimensional vector $\check{\boldsymbol{q}}_j$ with the elements

$$\check{q}_k^{(j)} = \lambda_k^{\frac{1}{2}} \int_0^1 \int_0^1 f_k(u)h(u,v)\check{f}_j(v)dudv, \quad k = 1,2,\ldots,$$

is also the solution of (4.4) for $\lambda = \check{\lambda}_j$ from Lemma 3 together with (4.10). □

It is also interesting to note that a simpler kernel function of the integral equation instead of (4.9) is available in the case of the minimum distance estimator. The reason is that $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$ can be approximated by a simple function of the empirical process.

Proposition 6. *The* $\left\{\bar{\lambda}_j\right\}$ *are also the eigenvalues of the integral equation*

$$\lambda f(u) = \int_0^1 \zeta(u,v) f(v)dv \tag{4.11}$$

with the kernel function

$$\zeta(u,v) = \rho_0(u,v) - \boldsymbol{h}(u)^\top \boldsymbol{g}(v,\boldsymbol{\theta}),$$

where

$$\boldsymbol{h}(u) = A^{-1} \int_0^1 \rho_0(u,v)\boldsymbol{g}(v,\boldsymbol{\theta})dv.$$

Proof. Similarly as in the proof of Proposition 5, the function

$$f(u) = \sum_{k=1}^{\infty} \lambda_k^{\frac{1}{2}} f_k(u) q_k^{(j)}$$

is the solution of (4.11) for $\lambda = \bar{\lambda}_j$, and

$$\tilde{q}_k^{(j)} = \lambda_j^{\frac{1}{2}} \int_0^1 \int_0^1 f_j(u)h(u,v)\tilde{f}_k(v)dudv$$

is the solution of (4.4) for $\lambda = \tilde{\lambda}_j$, where $\tilde{f}_k(v), k = 1,2,\ldots$, are the eigenfunctions for (4.11) with $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots$. Therefore, $\left\{\bar{\lambda}_j\right\}$ are also the eigenvalues of the integral equation (4.11). □

4.5 Robustness

In this section, we investigate the robustness of the Cramér-von Mises statistic when the parameters are estimated by the minimum distance method. Several theoretical results are developed from the robustness of the minimum distance estimator to that of the test statistic. To compare with the case that the parameters are estimated by the maximum likelihood method, the numerical experiments are presented.

4.5.1 Minimum distance estimators and their robustness

In general, a minimum distance estimator is referred as “an estimator chosen to minimize a certain distance of two functions.” General review and the bibliography of the minimum distance estimator can be found in Parr (1981). Because of its general name, there are various kinds of estimators called “minimum distance estimator.” A fundamental difference comes from functions to be measured for the distance. In this context, we focus on distribution functions, that is, the distance between an empirical distribution function and a distribution function is focused. For the distance based on probability density functions, see Basu et al. (2011) for example.

Various characteristics of a minimum distance estimator which is chosen to minimize a distance based on the empirical distribution function $F_n(x)$ and the distribution function $F(x,\boldsymbol{\theta})$ are investigated. Sahler (1970) gave conditions under which minimum distance estimators exist and are consistent. Bolthausen (1977) showed the weak convergence of minimum distance estimators for general parameters than location parameters and other norm than integral-type ones.

One of the advantages of using minimum distance estimators is their robustness. Robustness is a word widely used in many senses and there are many results showing the robustness of minimum distance estimators. In Parr and Schucany (1980), Monte Carlo results show that minimum distance estimators are competitive with other estimators in the sense of the variance of the location parameter of a symmetric distribution. A mathematical framework for describing the robustness of minimum distance estimators is constructed by Millar (1981). Donoho and Liu (1988) showed that minimum distance estimators are “automatically” robust, in the sense of the stability of the quantity estimated.

In addition to the results for general minimum distance estimators described above, the robustness of the minimum distance estimator which is chosen to minimize the Cramér-von Mises statistic is also studied by many researchers. Woodward et al. (1984) demonstrated by numerical experiments that the minimum distance estimator is better than the maximum likelihood estimator under symmetric departures from normality of each component in normal mixture models. Since then, the minimum distance estimator is often used in practice for mixture models to avoid instability of the identification of the distribution due to small number of outlying observations (Beutner and Bordes, 2011, García-Dorado and Marin, 1998). The robustness based on the influence function for complete and grouped data is considered in Duchesne et al. (1997). Moreover, the minimum distance estimator which is chosen to minimize the Cramér-von Mises statistic shares the same loss function with the goodness-of-fit test if we adopt the Cramér-von Mises statistic as a goodness-of-fit test statistic. It seems natural to employ the same loss function for both parameter estimation and a goodness-of-fit test.

4.5.2 Millar’s robustness and the minimum distance estimator

We first introduce the result on the robustness given by Millar (1981) because it is suitable for considering the relationship between a minimum distance estimator and a test statistic. Let H be a finite measure on $\mathbb{R}^1$ and define $|\cdot|_H$ and $\langle\cdot,\cdot\rangle_H$ to be norm and inner product of $L^2(H)$. Millar (1981) considered a risk of parameter estimation when observations are from a contaminated distribution $G_{nq}(x) = F(x,\boldsymbol{\theta}) + \frac{1}{\sqrt{n}}q(x)$, where q is in $N(c) = \left\{q \in L^2(H); \int q(x)dH(x) < c\right\}$ and chosen so that $G_{nq}(x)$ is a distribution function, and proved that under suitable regularity conditions,

$$\lim_n \inf_{\tilde{\theta}} \sup_{q\in N(c_n)} n \int \left|F(\cdot,\theta^*) - F(\cdot,\tilde{\theta})\right|_H^2 dG_{nq}^n \geq \mathrm{E}\left\{|\pi w(F(\cdot,\theta))|_H^2\right\} \tag{4.12}$$

for any increasing sequence c_n. Here an operator π is an orthogonal projection in $L^2(H)$ to the subspace $\Sigma = \{\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \xi\rangle; \boldsymbol{\theta}_n \in \mathbb{R}^m\}$, where $|\cdot|$ and $\langle , \rangle$ are the Euclidean norm and inner product and ξ is a function such that $\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \xi\rangle \in L^2(H)$ for all $\boldsymbol{\theta}_n$ and

$$|F(\cdot,\boldsymbol{\theta}_n) - F(\cdot,\boldsymbol{\theta}) - \langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \xi\rangle|_H = o(|\boldsymbol{\theta}_n - \boldsymbol{\theta}|)$$

for any $\boldsymbol{\theta}_n$ which goes to $\boldsymbol{\theta}$. The $\boldsymbol{\theta}^*$ is the pseudo true value which attains

$$\inf_{\boldsymbol{\theta}} \left|G_{nq}(\cdot) - F(\cdot,\boldsymbol{\theta})\right|_H = \left|G_{nq}(\cdot) - F(\cdot,\boldsymbol{\theta}^*)\right|_H$$

and $w(u)$ is a Brownian bridge. Millar (1981) defined any sequence of estimators $\boldsymbol{\theta}_n$ for which the limiting minimax risk,

$$\lim_n \sup_{q\in N(c_n)} n \int |F(\cdot,\theta^*) - F(\cdot,\boldsymbol{\theta}_n)|_H^2 dG_{nq}^n$$

in this case, is equal to the lower bound of (4.12) as “H-robust” and showed that an estimator $\boldsymbol{\theta}'$ that attains

$$\inf_{\boldsymbol{\theta}} |F_n(\cdot) - F(\cdot,\boldsymbol{\theta})|_H = \left|F_n(\cdot) - F(\cdot,\boldsymbol{\theta}')\right|_H$$

is “H-robust.” The following lemma shows that the lower bound of (4.12) can be written in other form.

Lemma 4.

$$\mathrm{E}\left\{|\pi w(F(\cdot,\theta))|_H^2\right\} = \lim_{n\to\infty} n\mathrm{E}\left\{\left|\left(\boldsymbol{\theta}_0' - \boldsymbol{\theta}\right)^\top \boldsymbol{g}(F(x,\boldsymbol{\theta}),\boldsymbol{\theta})\right|_H^2\right\}, \tag{4.13}$$

where $\boldsymbol{\theta}_0'$ is the estimator which satisfies

$$\inf_{\boldsymbol{\theta}} \left|F_n^0(\cdot) - F(\cdot,\boldsymbol{\theta})\right|_H = \left|F_n^0(\cdot) - F(\cdot,\boldsymbol{\theta}_0')\right|_H$$

and $F_n^0(x)$ is the empirical distribution function for observations from $F(x,\boldsymbol{\theta})$.

Proof. Since $F_n^0(x)$ is the empirical distribution function for observations from $F(x,\boldsymbol{\theta})$, the empirical process $\sqrt{n}\left\{F_n^0(x)-F(x,\theta)\right\}$ converges to a Brownian bridge $w(F(x,\theta))$ so that the left hand side of (4.13) is equal to

$$\lim_{n\to\infty} n\mathrm{E}\left\{\left|\pi\left(F_n^0(\cdot)-F(\cdot,\boldsymbol{\theta})\right)\right|_H^2\right\}.$$

We have

$$\pi\left(F_n^0(\cdot)-F(\cdot,\theta)\right)=\pi\left(F_n^0(\cdot)-F(\cdot,\boldsymbol{\theta}_0')\right)+\pi\left(F(\cdot,\boldsymbol{\theta}_0')-F(\cdot,\theta)\right) \tag{4.14}$$

and it is shown as follows that the first term on the right hand side of (4.14) converges to 0 as n tends to infinity. Noting that $\boldsymbol{\theta}_0'$ satisfies

$$\left\langle F_n^0(\cdot)-F(\cdot,\boldsymbol{\theta}_0'),\ \boldsymbol{g}(F(\cdot,\theta),\boldsymbol{\theta}_0')\right\rangle=\mathbf{0},$$

it follows that

$$\begin{aligned}&\left\langle F_n^0(\cdot)-F(\cdot,\boldsymbol{\theta}_0'),\ F(\cdot,\boldsymbol{\theta}_0')-F(\cdot,\theta)\right\rangle\\ &=\left\langle F_n^0(\cdot)-F(\cdot,\boldsymbol{\theta}_0'),\ \left(\boldsymbol{\theta}_0'-\boldsymbol{\theta}\right)^\top\boldsymbol{g}(F(\cdot,\boldsymbol{\theta}),\boldsymbol{\theta}_0')+o(|\boldsymbol{\theta}_0'-\boldsymbol{\theta}|)\right\rangle\end{aligned}$$

converges to 0 as n tends to infinity. Then the first term on the right hand side of (4.14) converges to 0 as n tends to infinity because π is the orthogonal projection to Σ.

The second term on the right hand side of (4.14) is asymptotically equal to

$$\pi\left(\left(\boldsymbol{\theta}_0'-\boldsymbol{\theta}\right)^\top\frac{\partial}{\partial\boldsymbol{\theta}}F(x,\boldsymbol{\theta})\right)=\left(\boldsymbol{\theta}_0'-\boldsymbol{\theta}\right)^\top\boldsymbol{g}(F(x,\boldsymbol{\theta}),\boldsymbol{\theta})$$

from a Taylor expansion

$$F(x,\theta)-F(x,\boldsymbol{\theta}_0')=\left(\boldsymbol{\theta}_0'-\boldsymbol{\theta}\right)^\top\boldsymbol{g}(F(x,\boldsymbol{\theta}),\boldsymbol{\theta})+o(|\theta|)$$

and the proof is complete. □

In the case of the minimum distance estimator

The result of Lemma 4 is easily interpreted in the case that observations are from the contaminated distribution $F_\varepsilon(x,\boldsymbol{\theta})$ and the parameters are estimated by the minimum distance method, which is the estimation method of finding the value $\hat{\boldsymbol{\theta}}$ which makes the Cramér-von Mises statistic a minimum. Consider the case $H(x) = F(x,\boldsymbol{\theta})$ and $G_{nq}(x) = F_\varepsilon(x,\boldsymbol{\theta})$ with $q(x) = \varepsilon\{G(x) - F(x,\boldsymbol{\theta})\}$, then the following theorem is derived from Millar's result and Lemma 4.

Theorem 8. *For any increasing sequence c_n,*

$$\lim_{n\to\infty} \inf_{\tilde{\theta}} \sup_{\varepsilon<c_n, G(x)} n\mathrm{E}_{F_\varepsilon}\left[\int_{-\infty}^{\infty}\left\{F(x,\boldsymbol{\theta}^*) - F(x,\tilde{\boldsymbol{\theta}})\right\}^2 dF(x,\boldsymbol{\theta})\right]$$

$$\geq \int_0^1\int_0^1 \rho_0(u,v)\boldsymbol{g}^\top(u,\boldsymbol{\theta})A^{-1}\boldsymbol{g}(v,\boldsymbol{\theta})dudv, \tag{4.15}$$

where $\boldsymbol{\theta}^*$ attains

$$\inf_\theta \int_{-\infty}^{\infty}\{F_\varepsilon(x,\boldsymbol{\theta}) - F(x,\boldsymbol{\theta})\}^2 dF(x,\boldsymbol{\theta}) = \int_{-\infty}^{\infty}\{F_\varepsilon(x,\boldsymbol{\theta}) - F(x,\boldsymbol{\theta}^*)\}^2 dF(x,\boldsymbol{\theta}).$$

Proof. It follows from Lemma 4 that we only need to calculate

$$\lim_{n\to\infty} n\mathrm{E}\left[\int_{-\infty}^{\infty}\left\{\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^\top \boldsymbol{g}(F(x,\boldsymbol{\theta}),\boldsymbol{\theta})\right\}^2 dF(x,\boldsymbol{\theta})\right],$$

where $\hat{\boldsymbol{\theta}}$ is the minimum distance estimator and estimated from observations following the distribution $F(x,\boldsymbol{\theta})$. The proof is complete from the fact that $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$ converges to $A^{-1}\int_0^1 w(u)\boldsymbol{g}(u,\boldsymbol{\theta})du$, which is also used in the proof of Lemma 3.

□

As shown by Millar (1981), the minimum distance estimator $\hat{\boldsymbol{\theta}}$ is robust in this framework of the robustness. The following result also shows that the limiting risk for $\hat{\boldsymbol{\theta}}$

$$\lim_{n\to\infty} n\mathrm{E}_{F_\varepsilon}\left[\int_{-\infty}^{\infty}\left\{F(x,\boldsymbol{\theta}^*) - F(x,\hat{\boldsymbol{\theta}})\right\}^2 dF(x,\boldsymbol{\theta})\right]$$

does not depend on $G(x)$ and ε.

Remark 1. *For the minimum distance estimator $\hat{\boldsymbol{\theta}}$,*

$$\lim_{n\to\infty} n\mathrm{E}_{F_\varepsilon}\left[\int_{-\infty}^{\infty}\left\{F(x,\boldsymbol{\theta}^*)-F(x,\hat{\boldsymbol{\theta}})\right\}^2 dF(x,\boldsymbol{\theta})\right]$$
$$=\int_0^1\int_0^1 \rho_0(u,v)\boldsymbol{g}^\top(u,\boldsymbol{\theta})A^{-1}\boldsymbol{g}(v,\boldsymbol{\theta})dudv. \qquad (4.16)$$

For the minimum distance estimator, it follows from Remark 1 that the limiting risk attains the lower bound (4.15), however, it does not always attain for other estimators. As an example, we consider an M-estimator $\bar{\boldsymbol{\theta}}$, which satisfies $\sum_{j=1}^{n}\psi\left(X_j,\bar{\theta}\right)=\mathbf{0}$, where $\psi(x,\boldsymbol{\theta})$ is a m-dimensional function and differentiable with respect to $\boldsymbol{\theta}$. M-estimators are proposed by Huber (1964) as a generalization of the maximum likelihood estimator.

Remark 2. *For the M-estimator $\bar{\boldsymbol{\theta}}$,*

$$\lim_{n\to\infty} n\mathrm{E}_{F_\varepsilon}\left[\int_{-\infty}^{\infty}\left\{F(x,\boldsymbol{\theta}^*)-F(x,\bar{\boldsymbol{\theta}})\right\}^2 dF(x,\boldsymbol{\theta})\right]$$
$$=\int_0^1 \boldsymbol{g}^\top(u,\boldsymbol{\theta})\Big\{A^{-1}\boldsymbol{q}\boldsymbol{q}^\top A^{-1}-A^{-1}\boldsymbol{q}\boldsymbol{r}^\top\Psi_1^{-1}$$
$$-\Psi_1^{-1}\boldsymbol{r}\boldsymbol{q}^\top A^{-1}+\Psi_1^{-1}\left(\Psi_2^{-1}+\boldsymbol{r}\boldsymbol{r}^\top\right)\Psi_1^{-1}\Big\}\boldsymbol{g}(u,\boldsymbol{\theta})du, \qquad (4.17)$$

where $\boldsymbol{q}$ and $\boldsymbol{r}$ are m-dimensional vectors such that

$$\boldsymbol{q}=\varepsilon\int_0^1\left\{G(F^{-1}(u,\boldsymbol{\theta}))-u\right\}\boldsymbol{g}(u,\boldsymbol{\theta})du$$
$$\boldsymbol{r}=\varepsilon\int_{-\infty}^{\infty}\left\{g(x)-f(x,\boldsymbol{\theta})\right\}\psi(x,\boldsymbol{\theta})dx$$

and Ψ_1 and Ψ_2 are $m\times m$ matrices such that

$$\Psi_1=\int_{-\infty}^{\infty}\frac{\partial}{\partial\boldsymbol{\theta}^\top}\psi(x,\boldsymbol{\theta})dF(x,\boldsymbol{\theta})$$
$$\Psi_2=\int_{-\infty}^{\infty}\psi(x,\boldsymbol{\theta})\psi(x,\boldsymbol{\theta})^\top dF(x,\boldsymbol{\theta}).$$

It can be seen from Remark 2 that the limiting risk for the M-estimator (4.17) depends on $G(x)$ and ε, while the limiting risk for the minimum distance estimator (4.16) does not. The difference is demonstrated in the following example.

Example 3

In this example, we consider the case where $F(x,\boldsymbol{\theta}) = \Phi(x-\theta_0)$ with $\boldsymbol{\theta} = \theta_0$ and $G(x) = \Phi(x-\theta_1)$, where $\Phi(x)$ is the distribution function of the standard normal distribution. Then, we have

$$F_\varepsilon(x,\boldsymbol{\theta}) = \left(1 - \frac{\varepsilon}{\sqrt{n}}\right)\Phi(x-\theta_0) + \frac{\varepsilon}{\sqrt{n}}\Phi(x-\theta_1),$$

which is the normal distribution contaminated with another normal distribution. We here note that the results on the calculations of the distribution function and the probability density function of the normal distribution given by Owen (1980) and Patel and Read (1996) are used in the following calculations. As shown in the previous, the lower bound (4.15) does not depend on $G(x)$ and ε and becomes

$$\sqrt{12}\pi\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\min\left(\Phi(x),\Phi(y)\right)\phi^2(x)\phi^2(y)dxdy - \frac{1}{16\pi}\right\}, \tag{4.18}$$

where $\phi(x)$ is the probability density function of the standard normal distribution.

Here we consider a simple example of M-estimators, such that $\psi(x,\boldsymbol{\theta}) = [x-\theta]_{-b}^{b}$, where $[y]_{-b}^{b} = y$ for $|y| < b$ and 0 otherwise. Then the limiting risk is given as

$$\frac{1}{\sqrt{12}\pi}\left(\frac{1}{\Phi(b)-\Phi(-b)} + \varepsilon^2\left[-\sqrt{\frac{3\pi}{2}}\left\{\Phi\left(\frac{\sqrt{2}\theta_2}{\sqrt{3}}\right) - \frac{1}{2}\right\}\right.\right.$$
$$\left.\left. + \frac{\phi(b+\theta_2) - \phi(-b+\theta_2) - \theta_2\left\{\Phi(b+\theta_2) - \Phi(-b+\theta_2)\right\}}{\Phi(b)-\Phi(-b)}\right]^2\right)$$

with $\theta_2 = \theta_0 - \theta_1$. In particular, the limiting risk for the maximum likelihood estimator, which is the case $b = \infty$ in the estimator given above, is given as

$$\frac{1}{\sqrt{12}\pi}\left(1 + \varepsilon^2\left[-\sqrt{\frac{3\pi}{2}}\left\{\Phi\left(\frac{\sqrt{2}\theta_2}{\sqrt{3}}\right) - \frac{1}{2}\right\} - \theta_2\right]^2\right).$$

Figure 4.2 illustrates the limiting risk curves for the minimum distance estimator (MDE), the maximum likelihood estimator (MLE), and the

M-estimators (M ($b = 2$), M ($b = 3$), M ($b = 4$)) against ε and $|\theta_0 - \theta_1|$, respectively. For both cases the limiting risk curves for the minimum distance estimator, which is given by (4.18), are obtained by a Monte Carlo simulation with 1,000,000 replications. These results show that the limiting risk of the minimum distance estimator is a little larger than those of other estimators for very small ε and for very small $|\theta_0 - \theta_1|$, however, the limiting risk curve stays when ε or $|\theta_0 - \theta_1|$ increases, while those of other estimators increase, especially when ε increases.

4.5.3 The Cramér-von Mises statistic when an estimator is plugged in

Here we investigate the property of the Cramér-von Mises statistic when contamination exists using the robustness of the parameters, which we have shown in the previous section. Noting that for any estimator $\tilde{\boldsymbol{\theta}}$, we have

$$
\begin{aligned}
& n\int_{-\infty}^{\infty} \{F_n(x) - F(x,\boldsymbol{\theta}^*)\}^2 dF(x,\boldsymbol{\theta}) \\
&\quad = n\int_{-\infty}^{\infty} \left\{F_n(x) - F(x,\tilde{\boldsymbol{\theta}})\right\}^2 dF(x,\boldsymbol{\theta}) + \xi_1\left(\tilde{\boldsymbol{\theta}}\right) - \xi_2\left(\tilde{\boldsymbol{\theta}}\right),
\end{aligned}
$$

where

$$
\xi_1\left(\tilde{\boldsymbol{\theta}}\right) = n\int_{-\infty}^{\infty} \left\{F(x,\boldsymbol{\theta}^*) - F(x,\tilde{\boldsymbol{\theta}})\right\}^2 dF(x,\boldsymbol{\theta})
$$

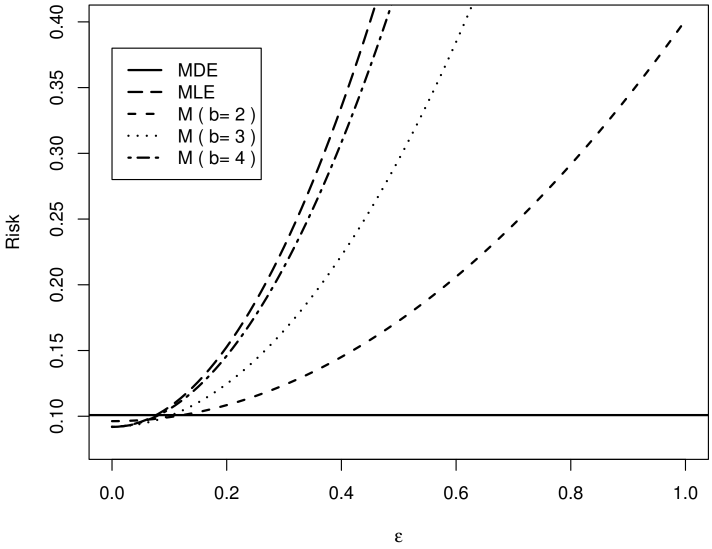
and

$$
\xi_2\left(\tilde{\boldsymbol{\theta}}\right) = 2n\int_{-\infty}^{\infty} \left\{F_n(x) - F(x,\tilde{\boldsymbol{\theta}})\right\}\left\{F(x,\boldsymbol{\theta}^*) - F\left(x,\tilde{\boldsymbol{\theta}}\right)\right\} dF(x,\boldsymbol{\theta}),
$$

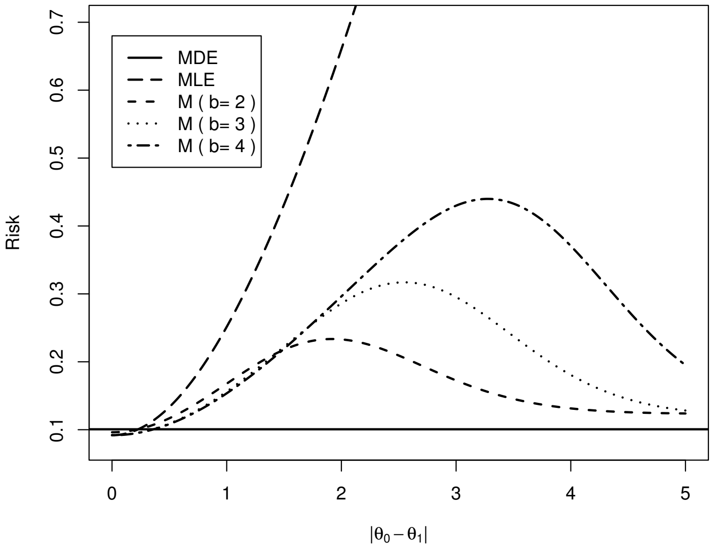
the Cramér-von Mises statistic $W_n^2\left(\tilde{\boldsymbol{\theta}}\right)$ can be divided into three parts as

$$
W_n^2\left(\tilde{\boldsymbol{\theta}}\right) = n\int_{-\infty}^{\infty} \{F_n(x) - F(x,\boldsymbol{\theta}^*)\}^2 dF(x,\boldsymbol{\theta}) - \xi_1(\tilde{\boldsymbol{\theta}}) + \xi_2(\tilde{\boldsymbol{\theta}}). \tag{4.19}
$$

We can see that the first term on the right hand side of (4.19) is independent of the estimator $\tilde{\boldsymbol{\theta}}$ and the limit of the expectation of $\xi_1\left(\tilde{\boldsymbol{\theta}}\right)$ is the limiting risk, which we have evaluated in the previous section.



(a) The limiting risk against ε.



(b) The limiting risk against $|\theta_0 - \theta_1|$.

Figure 4.2: Comparisons of the limiting risks for the minimum distance estimator, the maximum likelihood estimator, and the M-estimators.

If the minimum distance estimator $\hat{\boldsymbol{\theta}}$ is used, the limit distribution of $\xi_1\left(\hat{\boldsymbol{\theta}}\right)$ is independent of $G(x)$ and ε and the limit of the expectation of $\xi_1\left(\hat{\boldsymbol{\theta}}\right)$ is given as a constant, as shown in Remark 1. In addition, we have $\xi_2\left(\hat{\boldsymbol{\theta}}\right) = 0$ since the minimum distance estimator $\hat{\boldsymbol{\theta}}$ satisfies

$$\int_{-\infty}^{\infty} \left\{F_n(x) - F\left(x, \hat{\boldsymbol{\theta}}\right)\right\} \boldsymbol{g}(F(x,\boldsymbol{\theta}),\boldsymbol{\theta})dF(x,\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}.$$

Therefore, the Cramér-von Mises statistic when the parameters are estimated by the minimum distance method can be divided into two parts; the term independent of the estimator and the term independent of $G(x)$ and ε.

These characteristics are distinctive of the estimator because $\xi_1\left(\tilde{\boldsymbol{\theta}}\right)$ usually depends on $G(x)$ and ε and $\xi_2\left(\tilde{\boldsymbol{\theta}}\right)$ is not always 0. We can see them in the case of an M-estimator $\bar{\boldsymbol{\theta}}$ as an example. As we have shown, the limit of the expectation of $\xi_1\left(\tilde{\boldsymbol{\theta}}\right)$ is (4.17). Note that $\xi_2\left(\bar{\boldsymbol{\theta}}\right)$ is asymptotically equal to

$$2n\left[\left(\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}}\right)^{\top} \int_{-\infty}^{\infty} \{F_n(x) - F_\varepsilon(x,\boldsymbol{\theta})\} \boldsymbol{g}(F(x,\boldsymbol{\theta}),\boldsymbol{\theta})dF(x,\boldsymbol{\theta}) + \xi_1(\bar{\boldsymbol{\theta}})\right].$$

If $\psi(x)$ is chosen as $\bar{\boldsymbol{\theta}}$ to follow the normal distribution asymptotically, $\sqrt{n}\left(\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}}\right)$ converges to the normal distribution with mean $A^{-1}\boldsymbol{q} - \Psi_1^{-1}\boldsymbol{r}$ and variance $\Psi_1^{-1}\Psi_2\Psi_1^{-1}$ and $\xi_2\left(\bar{\boldsymbol{\theta}}\right)$ depends on $G(x)$ and ε.

4.5.4 Power and robustness

In this section, via Monte Carlo simulations we compare the Cramér-von Mises goodness-of-fit tests when the parameters are estimated by the minimum distance method and by the maximum likelihood method.

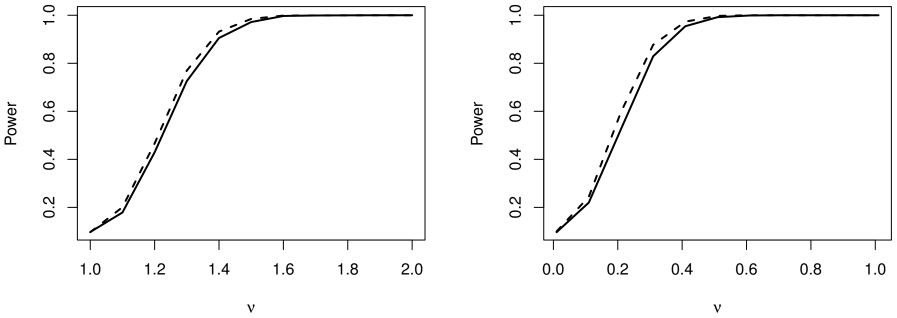
In the following comparisons, we set as follows. The sample size here is fixed at 200, the number of the replications in Monte Carlo simulations is 3,000, and the significance level is 0.1. Testing exponentiality indicates testing the goodness-of-fit of the exponential distribution when the mean parameter is estimated and testing normality indicates testing the goodness-of-fit of the normal distribution when only the mean parameter is estimated and the variance parameter is fixed as 1 as the known parameter. The solid line in each panel is for the case of the minimum distance estimator and the dotted line is for the case of the maximum likelihood estimator.

The power curves when no contamination exists are drawn in Figure 4.3. Figure 4.3 (a) shows the power curves for testing exponentiality when observations are from the gamma distribution with shape ν and scale 1 and Figure 4.3 (b) shows the power curves for testing normality when observations are from Student's t-distribution, which has the probability density function

$$f(x,d) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{d\pi}\,\Gamma\left(\frac{d}{2}\right)}\left(1+\frac{x^2}{d}\right)^{-\frac{d+1}{2}}$$

with the parameter $\nu = 1/d$, where d is the number of degrees of freedom. The power curves are drawn against the shape parameter ν in Figure 4.3 (a) and the parameter $\nu = 1/d$ in Figure 4.3 (b), respectively. These results suggest that there is no significant difference between the minimum distance and the maximum likelihood estimator when no contamination exists.

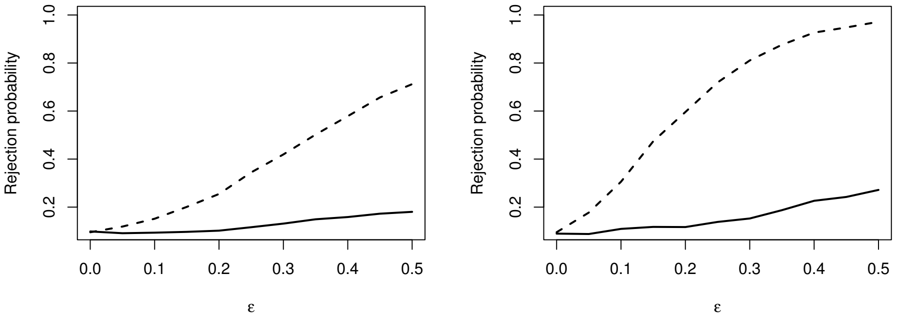
On the other hand, the rejection probability of the Cramér-von Mises goodness-of-fit test of a distribution $F(x,\boldsymbol{\theta})$ when the observations follow the distribution $F_\varepsilon(x,\boldsymbol{\theta})$ are shown in Figure 4.4. The distribution of the contamination $G(x)$ is the normal distribution with mean 7 and variance 1. Figure 4.4 (a) is for the case testing exponentiality when the distribution $F(x,\boldsymbol{\theta})$ is the exponential distribution with mean 1 and Figure 4.4 (b) is for testing normality when the distribution $F(x,\boldsymbol{\theta})$ is the standard normal distribution. It is observed from both figures that the rejection probability quickly increases for the maximum likelihood estimator, while it does not for the minimum distance estimator. These results indicate that the use of the minimum distance estimator makes the goodness-of-fit test robust to contamination.



(a) Testing exponentiality.

(b) Testing normality.

Figure 4.3: Power curves against the parameter ν.



(a) Testing exponentiality.

(b) Testing normality.

Figure 4.4: Rejection probabilities against the rate of contamination ε.

4.6 Concluding remarks

The asymptotic distributions of the Cramér-von Mises statistic are derived when the observations are contaminated for both cases where the parameters are known and where the parameters are estimated by the minimum distance method. These results are consistent to the known result when no contamination exists. For both cases, the asymptotic distribution is given as a distribution of a weighted infinite sum of non-central chi-squared random variables with 1 degree of freedom and the effect of contamination on the asymptotic distribution appears only in the non-centralities. Moreover, the derivations of the asymptotic distributions suggest a simple procedure to obtain the approximation of distribution of the Cramér-von Mises statistic.

We also show that the extension of the Millar's result on robustness of the minimum distance estimator associates with the robustness of the Cramér-von Mises goodness-of-fit test. Numerical experiments indicate that the use of the minimum distance estimator makes the test less sensitive to contamination, although the power of the test stays almost the same as that for the maximum likelihood estimator. Such insensitivity would be harmful when the aim of the test is to detect the existence of contamination. However, it becomes an advantage if the aim of the test is to check whether the underlying probability distribution model can be used or not. It often happens in practice that the hypothesis testing is not a goal but the beginning of an analysis. In such case the robust goodness-of-fit test, which is insensitive to small number of contaminations, would be preferred.

Chapter 5

Conclusion

We have investigated the role of the goodness-of-fit test of distributions from two case studies and the asymptotic behavior of the Cramér-von Mises statistic when contamination exists.

In the first case study, the trawling effect is verified by using the gamma distribution, which is derived as the equilibrium distribution of the stochastic growth model, as a model of the weight of animals on seabed. To examine the goodness-of-fit of the gamma distribution to the weight distribution, we have used the extended version of the Cramér-von Mises statistic because the observations are independent but not identically distributed. It is shown that the gamma distribution can be used for the model of the weight distribution before trawling for 57 cases out of 80. For 47 cases with large enough sample size of the data after trawling out of the 57 cases, we have classified the change of the weight distribution into three types: unaffected, lighter, and changed. The classification is based on the goodness-of-fit test of the gamma distribution with the parameters estimated from the observations before trawling and on the direction of the change of the weight distribution observed from the P-P plots. The results show the effect of trawling on the weight distribution of animals on seabed through the change of distribution.

In the second case study, a mixture distribution model is derived for the carapace length of banana prawns to investigate the effect of freshwater flows.

The model is a mixture of two kinds of probability distributions, for a cohort stayed in the estuary and for a cohort migrated from offshore waters to the estuary, derived by combining models for the carapace increment and for the survival rate. It is shown that the model can be used for 15 cases out of 19 by using the Cramér-von Mises statistic for discrete distributions. For these 15 cases, the model can explain the effects of the changes in temperature and salinity of water caused by freshwater flows on the growth of banana prawns.

For both case studies, the goodness-of-fit of the probability distribution model is examined by the Cramér-von Mises type statistics. As described in both case studies, the goodness-of-fit test is not a goal but the beginning of the analysis. For example, in the first case study, the goodness-of-fit test of the gamma distribution to the weight distribution of animals on seabed before trawling is just for validating whether the model can be used or not to give an approximation of the weight distribution. This is because the purpose of the analysis is to investigate the effect of trawling, not to judge whether the weights of animals on seabed follow the gamma distribution or not. From this point of view, the robustness of goodness-of-fit test to contamination would be attractable when one wishes to give an approximation model to analyze the data.

Theoretical studies of the asymptotic behavior of the Cramér-von Mises statistic when contamination exists provide a key to the robust method of the goodness-of-fit test. The asymptotic distribution of the Cramér-von Mises statistic for contaminated data is derived as a distribution of a weighted infinite sum of non-central chi-squared random variables with 1 degree of freedom for both cases when the parameters are known and when the parameters are estimated by the minimum Cramér-von Mises distance method. The effect of the contamination appears only in the non-centralities. We extended the mathematical framework of the robustness of the minimum distance estimator to that of the goodness-of-fit test statistic. The theoretical results and the numerical experiments show that using the minimum distance estimator makes the Cramér-von Mises goondess-of-fit test robust.

Goodness-of-fit test is often investigated from the view point of detecting the existence of contamination, so that robust property is not much of interest. However, the robustness would become a good property when the aim of the test is to check whether the probability distribution model is applicable or not, for example, in the two case studies we have explored. This is because it often happens in practice that there is small number of contaminations in the data and the goodness-of-fit test lies the beginning of the analysis. We hope that our results will help to connect between theoretical studies and practical demands.

Appendix A

Table A.1: Goodness-of-fit of the gamma distribution before trawling.

Case	Class	Family	Scientific name	Region	*n*
1	Hydrozoa		Hydroid OPNO 006	East	12
2	Hydrozoa		Hydroid OPNO 156	West	20
3	Hydrozoa		Hydroid OPNO 184	East	25
4	Hydrozoa		Hydroid OPNO 201	East	6
5	Gymnolaemata	Flustridae	*Retiflustra cornea*	East	18
6	Gymnolaemata		Cheilostomata sp OPNO 142	West	7
7	Gymnolaemata		Cheilostomata sp OPNO 142	East	17
8	Gymnolaemata		Scrupocellaria sp OPNO 215	East	7
9	Gymnolaemata		Bryozoan OPNO 142a	West	11
10	Gymnolaemata		Bryozoan OPNO 142b	West	21
11	Gymnolaemata		Bryozoan OPNO 171a	East	16
12	Gymnolaemata		Bryozoan OPNO 171b	East	16
13	Gymnolaemata		Bryozoan OPNO 203	East	21
14	Gymnolaemata		Bryozoan OPNO 216	East	9
15	Polychaeta		Tubeworm OPNO 006	East	21
16	Bivalvia	Nuculidae	*Leionucula superba*	East	6
17	Bivalvia	Glycymerididae	*Melaxinaea vitrea*	East	14
18	Bivalvia	Malleidae	*Malleus (Malleus) malleus*	East	25
19	Bivalvia	Pectinidae	*Amusium pleuronectes*	East	9
20	Bivalvia	Pectinidae	*Annachlamys flabellata*	East	13
21	Bivalvia	Spondylidae	Spondylidae OPNO 193	East	18
22	Bivalvia	Cardiidae	Cardiidae OPNO 151	West	9
23	Bivalvia	Veneridae	Lioconcha sp OPNO 004	East	7
24	Bivalvia	Veneridae	Placamen sp OPNO 156	West	10
25	Gastropoda	Neritidae	Neritidae OPNO 142	West	10
26	Gastropoda	Modulidae	Modulidae OPNO 151	West	9
27	Gastropoda	Strombidae	Strombus sp OPNO 142	West	13
28	Gastropoda	Strombidae	Strombus sp OPNO 150	West	9
29	Gastropoda	Muricidae	Chicoreus sp OPNO 184	East	24
30	Gastropoda	Muricidae	Murex sp OPNO 002	East	8
31	Gastropoda	Muricidae	Murex sp OPNO 172	East	9
32	Gastropoda	Cerithiidae	Cerithiidae OPNO 142	West	21
33	Gastropoda	Cancellariidae	Cancellariidae OPNO 151	West	10
34	Gastropoda	Architectonicidae	Architectonica sp OPNO 151	West	8
35	Gastropoda	Smaragdinellidae	Smaragdinellidae OPNO 151	West	15
36	Asteroidea	Luidiidae	Luidiidae OPNO 006	East	22
37	Asteroidea	Astropectinidae	Astropectinidae OPNO 006	East	18
38	Asteroidea	Astropectinidae	Astropectinidae OPNO 142	West	26
39	Asteroidea	Goniasteridae	Stellaster sp OPNO 006a	East	9
40	Asteroidea	Goniasteridae	Stellaster sp OPNO 006b	East	9

Case (again)	shape $\hat{\nu}$ (SE)	scale $\hat{\alpha}$ (SE)	$\tilde{W}_n^2\left(\hat{\theta}\right)$	p-value	
1	0.674(0.236)	3.712(1.774)	0.045	0.653	
2	1.140(0.321)	0.967(0.339)	0.111	0.091	$*$
3	0.349(0.088)	6.127(1.651)	0.075	0.409	
4	1.190(0.614)	1.366(0.871)	0.125	0.042	$*$
5	1.480(0.449)	0.439(0.158)	0.067	0.335	
6	1.024(0.483)	1.168(0.702)	0.120	0.059	$*$
7	2.344(0.754)	1.947(0.698)	0.102	0.107	
8	0.912(0.425)	2.124(1.300)	0.047	0.610	
9	0.852(0.315)	4.030(1.989)	0.035	0.793	
10	0.686(0.180)	13.502(5.021)	0.163	0.021	$*$
11	1.098(0.346)	2.918(1.120)	0.102	0.130	
12	0.346(0.098)	546.803(279.095)	0.100	0.174	
13	1.313(0.366)	24.512(8.162)	0.068	0.330	
14	1.382(0.589)	14.952(7.657)	0.050	0.519	
15	2.687(0.814)	0.062(0.019)	0.116	0.144	
16	3.102(1.712)	0.721(0.423)	0.060	0.420	
17	2.560(0.922)	3.187(1.210)	0.397	0.000	$*$
18	1.233(0.331)	6.478(1.814)	0.163	0.036	$*$
19	16.855(7.887)	0.783(0.370)	0.079	0.231	
20	3.235(1.233)	3.514(1.385)	0.052	0.551	
21	2.342(0.755)	3.467(1.154)	0.056	0.540	
22	12.990(6.055)	0.353(0.167)	0.026	0.917	
23	4.320(2.254)	0.660(0.354)	0.031	0.895	
24	13.680(6.066)	0.117(0.053)	0.067	0.355	
25	3.780(1.628)	0.496(0.226)	0.067	0.326	
26	26.664(12.517)	0.064(0.030)	0.038	0.726	
27	5.442(2.102)	0.936(0.367)	0.035	0.815	
28	4.147(1.915)	0.752(0.355)	0.051	0.579	
29	2.449(0.686)	1.211(0.351)	0.130	0.057	$*$
30	1.613(0.742)	1.775(0.940)	0.070	0.286	
31	5.524(2.557)	0.600(0.284)	0.029	0.889	
32	10.651(3.266)	0.077(0.024)	0.100	0.182	
33	35.112(15.636)	0.043(0.019)	0.050	0.514	
34	3.794(1.834)	1.459(0.740)	0.061	0.396	
35	1.215(0.424)	0.779(0.278)	0.584	0.000	$*$
36	1.603(0.465)	0.655(0.197)	0.038	0.771	
37	2.804(0.898)	0.726(0.243)	0.131	0.055	$*$
38	0.319(0.081)	4.116(1.154)	0.287	0.000	$*$
39	0.492(0.193)	132.040(78.797)	0.124	0.070	$*$
40	4.852(2.231)	6.010(2.842)	0.073	0.325	

Table A.1: (continued).

Case	Class	Family	Scientific name	Region	*n*
41	Asteroidea	Goniasteridae	Stellaster sp OPNO 118	West	6
42	Asteroidea	Goniasteridae	Stellaster sp OPNO 118	East	22
43	Ophiuroidea	Ophiuridae	Ophiuroidea OPNO 171b	East	9
44	Ophiuroidea	Ophiuridae	Ophiuroidea OPNO 171c	East	19
45	Ophiuroidea	Ophiuridae	Ophiuroidea OPNO 177a	East	7
46	Ophiuroidea	Ophiuridae	Ophiuroidea OPNO 006	East	24
47	Echinoidea	Temnopleuridae	Temnopleuridae OPNO 142	West	18
48	Echinoidea	Temnopleuridae	Temnopleuridae OPNO 203a	East	18
49	Echinoidea	Temnopleuridae	Temnopleuridae OPNO 203b	East	19
50	Echinoidea	Laganidae	Laganidae OPNO 142a	West	18
51	Echinoidea	Laganidae	Laganidae OPNO 142a	East	12
52	Echinoidea	Laganidae	Laganidae OPNO 142b	West	6
53	Echinoidea	Brissidae	Brissidae OPNO 006	East	14
54	Crustacea	Penaeidae	*Metapenaeopsis novaeguineae*	West	11
55	Crustacea	Penaeidae	*Parapenaeopsis cornuta*	West	14
56	Crustacea	Penaeidae	*Parapenaeopsis tenella*	West	8
57	Crustacea	Diogenidae	*Dardanus imbricatus*	West	17
58	Crustacea	Diogenidae	*Dardanus imbricatus*	East	12
59	Crustacea	Paguridae	Paguridae OPNO 142	West	11
60	Crustacea	Dorippidae	Dorippe sp OPNO 142a	West	9
61	Crustacea	Leucosiidae	*Leucosia whitei*	East	8
62	Crustacea	Leucosiidae	*Leucosia ocellata*	East	17
63	Crustacea	Leucosiidae	Leucosia sp OPNO 142	West	13
64	Crustacea	Leucosiidae	Arcania sp OPNO 008	East	16
65	Crustacea	Matutidae	*Matuta inermis*	West	11
66	Crustacea	Matutidae	*Matuta granulosa*	West	14
67	Crustacea	Majidae	Hyastenus sp OPNO 214	East	9
68	Crustacea	Majidae	Hyastenus sp OPNO 054	West	6
69	Crustacea	Majidae	Majidae OPNO 154b	West	8
70	Crustacea	Parthenopidae	*Aulacolambrus hoplonotus*	East	14
71	Crustacea	Parthenopidae	*Parthenope nodosus*	West	16
72	Crustacea	Parthenopidae	*Parthenope longispinus*	East	7
73	Crustacea	Parthenopidae	Parthenope sp OPNO 060	West	10
74	Crustacea	Portunidae	*Portunus (Portunus) pelagicus*	West	8
75	Crustacea	Portunidae	*Portunus (Monomia) rubromarginatus*	West	7
76	Crustacea	Portunidae	*Portunus (Xiphonectes) hastatoides*	East	7
77	Crustacea	Pilumnidae	*Pilumnus pugilator*	East	9
78	Ascidiacea	Clavelinidae	Clavelina sp OPNO 142	West	16
79	Ascidiacea	Ascidacea	Ascidian OPNO 211	East	11
80	Ascidiacea	Diazonidae	*Rhopalaea crassa*	East	9

Case (again)	shape $\hat{\nu}$ (SE)	scale $\hat{\alpha}$ (SE)	$\tilde{W}_n^2\left(\hat{\theta}\right)$	p-value	
41	0.641(0.319)	31.123(20.705)	0.087	0.186	
42	0.468(0.126)	14.756(4.703)	0.142	0.043	∗
43	2.149(0.973)	0.849(0.397)	0.065	0.450	
44	5.041(1.622)	0.354(0.115)	0.151	0.027	∗
45	2.683(1.359)	2.718(1.497)	0.064	0.356	
46	1.415(0.394)	1.029(0.296)	0.058	0.499	
47	1.742(0.561)	0.320(0.106)	0.023	0.970	
48	0.684(0.207)	60.401(20.608)	0.128	0.079	∗
49	2.388(0.748)	5.049(1.645)	0.075	0.293	
50	1.773(0.571)	0.613(0.203)	0.327	0.000	∗
51	6.750(2.722)	0.048(0.019)	0.051	0.610	
52	5.018(2.858)	0.259(0.149)	0.035	0.868	
53	1.007(0.354)	3.693(1.343)	0.041	0.826	
54	1.728(0.695)	0.194(0.084)	0.110	0.097	∗
55	1.213(0.420)	0.582(0.227)	0.128	0.070	∗
56	4.142(2.010)	0.066(0.033)	0.030	0.899	
57	1.382(0.443)	1.510(0.523)	0.100	0.162	
58	4.953(1.966)	0.479(0.198)	0.058	0.421	
59	3.405(1.411)	0.126(0.054)	0.053	0.559	
60	0.528(0.211)	6.585(3.644)	0.053	0.544	
61	27.462(13.660)	0.049(0.025)	0.071	0.293	
62	69.968(23.958)	0.038(0.013)	0.130	0.056	∗
63	2.144(0.812)	2.001(0.779)	0.082	0.285	
64	6.768(2.356)	0.050(0.018)	0.136	0.052	∗
65	1.851(0.745)	4.211(1.844)	0.171	0.009	∗
66	0.743(0.245)	26.130(11.074)	0.184	0.012	∗
67	1.760(0.770)	1.748(0.862)	0.037	0.767	
68	10.607(6.030)	0.031(0.018)	0.027	0.924	
69	4.163(2.024)	0.044(0.022)	0.040	0.721	
70	1.136(0.395)	2.426(0.940)	0.050	0.576	
71	1.414(0.474)	2.895(1.026)	0.064	0.425	
72	6.572(3.441)	0.816(0.439)	0.091	0.159	
73	7.953(3.498)	0.226(0.101)	0.038	0.770	
74	0.229(0.091)	95.524(68.991)	0.266	0.002	∗
75	0.690(0.314)	3.327(2.141)	0.091	0.172	
76	22.983(12.203)	0.027(0.014)	0.036	0.781	
77	5.297(2.443)	0.204(0.096)	0.045	0.648	
78	0.443(0.147)	3.419(1.145)	0.049	0.699	
79	2.655(1.093)	2.501(1.064)	0.058	0.508	
80	2.825(1.265)	9.045(4.392)	0.117	0.060	∗

Appendix B

Table B.1: Parameters and results of goodness-of-fit tests for Model 2.

Case	$\hat{r}_0$(SE)	$\hat{r}_1$(SE)	$\hat{r}_2$(SE)	$\hat{\delta}_1$(SE)	$\hat{\sigma}_1$(SE)	d	$\hat{\delta}_2 = \hat{\delta}_1$+d	$\hat{\sigma}_2$(SE)	$W_n^{(d)2}\left(\hat{\boldsymbol{\theta}}\right)$	p-value
3	0.000*	0.799 (0.099)	0.201 (0.099)	8.465 (0.600)	1.539 (0.429)	2	10.465	2.835 (1.334)	0.005	0.778
7	0.000*	0.165 (0.032)	0.835 (0.032)	5.261 (0.346)	1.368 (0.208)	2	7.261	3.295 (0.223)	0.033	0.230
12	0.027 (0.021)	0.852 (0.041)	0.121 (0.034)	4.311 (0.337)	2.250 (0.294)	6	10.311	1.660 (0.435)	0.010	0.756
14	0.079 (0.013)	0.786 (0.021)	0.135 (0.018)	6.014 (0.132)	1.396 (0.094)	4	10.014	1.498 (0.181)	0.001	0.984
17	0.322 (0.159)	0.368 (0.067)	0.310 (0.113)	5.099 (0.479)	2.093 (0.413)	4	9.099	2.752 (0.455)	0.021	0.218
18	0.304 (0.072)	0.145 (0.079)	0.551 (0.093)	5.733 (0.836)	1.307 (0.570)	2	7.733	2.359 (0.576)	0.024	0.210
19	0.270 (0.122)	0.552 (0.110)	0.178 (0.139)	8.895 (0.253)	0.611 (0.215)	2	10.895	0.805 (0.456)	0.005	0.464

Acknowledgment

I would like to express my sincere gratitude to my supervisor, Professor Ritei Shibata, and Professor Mihoko Minami for their supports of my Ph.D. studies and completion of the thesis. I have been extremely fortunate to have a supervisor who cared so much about my research, guided to work on my Ph.D. studies from a theoretical and practical point of view, and gave me various opportunities to develop my research. Professor Mihoko Minami has supported me by giving me not only valuable suggestions which incented me to widen my research but also warm and consistent encouragement.

I am also very grateful to the rest of my thesis committee: Professor Makoto Maejima, Professor Hiroshi Shiraishi, and Professor Akihisa Tamura, for giving me valuable comments and corrections, which improved this thesis significantly.

It has been a pleasure having valuable research experience at Keio University. I would like to thank Professor Kunio Shimizu and Professor Masaaki Sibuya for giving me insightful comments and suggestions. I cannot thank enough to members of Data Science laboratory. I am especially grateful to Dr. Hideyasu Shimadzu, Dr. Natsuhiko Kumasaka, and Dr. Yuki Sugaya for their unconditional supports.

Special thanks go to Dr. Ross Darnell, who gave me an opportunity to do an internship at CSIRO (Commonwealth Scientific and Industrial Research Organisation, Australia) as well as valuable comments on the analyses of the two case studies. I would like to thank Mr. Mick Haywood and Dr. Charis Burridge of CSIRO Marine Research for data and helpful suggestions in the analysis of the survey data for the assessment of the trawling effect. Also I would like to thank Mr. Ian Halliday of DEEDI (Department of Employment, Economic Development and Innovation, Australia) for data and valuable suggestions in the analysis of the survey data for the assessment of the effect of freshwater flows. I am grateful to Dr. Peter Thomson for his warm support and encouragement.

I would like to thank Japan Society for the Promotion of Science International Training Program for supporting my stays in CSIRO to develop my research.

Bibliography

Basu, A., Shioya, H. and Park, C. (2011). *Statistical Inference: the Minimum Distance Approach*, CRC Press, Boca Raton.

Beran, R. (1984). Minimum distance procedures, *in* P. Krishnaiah and P. Sen (eds), *Handbook of Statistics*, Vol. 4, Elsevier, Amsterdam, pp. 741–754.

Beutner, E. and Bordes, L. (2011). Estimators based on data-driven generalized weighted Cramér-von Mises distances under censoring–with applications to mixture models, *Scandinavian Journal of Statistics* **38**(1): 108–129.

Billingsley, P. (1968). *Convergence of Probability Measures*, Wiley, New York.

Bishop, J., Die, D. and Wang, Y.-G. (2000). A generalized estimating equations approach for analysis of the impact of new technology on a trawl fishery, *Australian & New Zealand Journal of Statistics* **42**(2): 159–177.

Bolthausen, E. (1977). Convergence in distribution of minimum-distance estimators, *Metrika* **24**(1): 215–227.

Boos, D. D. (1981). Minimum distance estimators for location and goodness of fit, *Journal of the American Statistical Association* **76**(375): 663–670.

Burke, M. D., Csorgo, M., Csorgo, S. and Révész, P. (1979). Approximations of the empirical process when parameters are estimated, *Annals of Probability* **7**(5): 790–810.

Burridge, C. Y., Pitcher, C. R., Wassenberg, T. J., Poiner, I. R. and Hill, B. J. (2003). Measurement of the rate of depletion of benthic fauna by prawn (shrimp) otter trawls: an experiment in the Great Barrier Reef, Australia, *Fisheries Research* **60**(2-3): 237–253.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*, Wadsworth, Belmont.

Choulakian, V. and Stephens, M. A. (2001). Goodness-of-fit tests for the generalized Pareto distribution, *Technometrics* **43**(4): 478–484.

Choulakian, V., Lockhart, R. A. and Stephens, M. A. (1994). Cramér-von Mises statistics for discrete distributions, *Canadian Journal of Statistics* **22**(1): 125–137.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests, *Biometrics* **10**(4): 417–451.

Collie, J. S., Hall, S. J., Kaiser, M. J. and Poiner, I. R. (2000). A quantitative analysis of fishing impacts on shelf-sea benthos, *Journal of Animal Ecology* **69**(5): 785–798.

Conover, W. J. (1972). A Kolmogorov goodness-of-fit test for discontinuous distributions, *Journal of the American Statistical Association* **67**(339): 591–596.

D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-fit Techniques*, Marcel Dekker, New York.

Darling, D. A. (1955). The Cramer-Smirnov test in the parametric case, *Annals of Mathematical Statistics* **26**(1): 1–20.

Davidson, J. (1938). On the ecology of the growth of the sheep population in South Australia, *Transactions of the Royal Society of South Australia* **62**: 141–148.

Donoho, D. L. and Liu, R. C. (1988). The “automatic” robustness of minimum distance functionals, *Annals of Statistics* **16**(2): 552–586.

Duchesne, P. and De Micheaux, P. L. (2010). Computing the distribution of quadratic forms: Further comparisons between the Liu–Tang–Zhang approximation and exact methods, *Computational Statistics & Data Analysis* **54**(4): 858–862.

Duchesne, T., Rioux, J. and Luong, A. (1997). Minimum Cramér-von Mises distance methods for complete and grouped data, *Communications in Statistics - Theory and Methods* **26**(2): 401–420.

Durbin, J. (1973). *Distribution Theory for Tests based on the Sample Distribution Function,* Society for Industrial and Applied Mathematics, Philadelphia.

Feldman, M. W. and Roughgarden, J. (1975). A population's stationary distribution and chance of extinction in a stochastic environment with remarks on the theory of species packing, *Theoretical Population Biology* **7**(2): 197–207.

Gan, F. F. and Koehler, K. J. (1990). Goodness-of-fit tests based on P-P probability plots, *Technometrics* **32**(3): 289–303.

García-Dorado, A. and Marin, J. M. (1998). Minimum distance estimation of mutational parameters for quantitative traits, *Biometrics* **54**(3): 1097–1114.

Gihman, I. I. and Skorohod, A. V. (1979). *The Theory of Stochastic Processes III*, Springer, New York, pp. 113–219.

Goel, N. and Richter-Dyn, N. (1974). *Stochastic Models in Biology*, Academic Press, New York.

Gürtler, N. and Henze, N. (2000). Recent and classical goodness-of-fit tests for the Poisson distribution, *Journal of Statistical Planning and Inference* **90**(2): 207–225.

Guttorp, P. and Lockhart, R. A. (1988). On the asymptotic distribution of quadratic forms in uniform order statistics, *Annals of Statistics* **16**(1): 433–449.

Halliday, I. and Robins, J. (2007). *Environmental Flows for Sub-tropical Estuaries: Understanding the Freshwater Needs of Estuaries for Sustainable Fisheries Production and Assessing the Impact of Water Regulation*, Departmentof Primary Industries and Fisheries.

Haywood, J. and Khmaladze, E. (2008). On distribution-free goodness-of-fit testing of exponentiality, *Journal of Econometrics* **143**(1): 5–18.

Haywood, M. D. E. and Staples, D. J. (1993). Field estimates of growth and mortality of juvenile banana prawns (*Penaeus merguiensis*), *Marine Biology* **116**(3): 407–416.

Haywood, M., Hill, B., Donovan, A., Rochester, W., Ellis, N., Welna, A., Gordon, S., Cheers, S., Forcey, K., McLeod, I., Moeseneder, C., Smith, G., Manson, F., Wassenberg, T., Thomas, S., Kuhnert, P., Laslett, G., Burridge, C. and Thomas, S. (2005). *Quantifying the Effects of Trawling on Seabed Fauna in Northern Prawn Fishery*, FRDC Project No. 2002/102, CSIRO Marine Research, Cleveland.

Henze, N. (1996). Empirical-distribution-function goodness-of-fit tests for discrete models, *Canadian Journal of Statistics* **24**(1): 81–93.

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case, *Annals of Mathematical Statistics* **42**(6): 1977–1991.

Holmgren, E. B. (1995). The P-P plot as a method for comparing treatment effects, *Journal of the American Statistical Association* **90**(429): 360–365.

Horn, S. D. (1977). Goodness-of-fit tests for discrete data: a review and an application to a health impairment scale, *Biometrics* **33**(1): 237–247.

Huber, P. J. (1964). Robust estimation of a location parameter, *Annals of Mathematical Statistics* **35**(1): 73–101.

Kaiser, M. J. and Spencer, B. E. (1996). The effects of beam-trawl disturbance on infaunal communities in different habitats, *Journal of Animal Ecology* **65**(3): 348–358.

Karlis, D. and Xekalaki, E. (2000). A simulation comparison of several procedures for testing the Poisson assumption, *The Statistician* **49**(3): 355–382.

Khmaladze, E., Brownrigg, R. and Haywood, J. (2007). Brittle power: On Roman Emperors and exponential lengths of rule, *Statistics & Probability Letters* **77**(12): 1248–1257.

Khmaladze, E. V. (1981). Martingale approach in the theory of goodness-of-fit tests, *Theory of Probability & Its Applications* **26**(2): 240–257.

Kojadinovic, I. and Yan, J. (2012). Goodness-of-fit testing based on a weighted bootstrap: A fast large-sample alternative to the parametric bootstrap, *Canadian Journal of Statistics* **40**(3): 480–500.

Koul, H. and DeWet, T. (1983). Minimum distance estimation in a linear regression model, *Annals of Statistics* **11**(3): 921–932.

LePage, R., Woodroofe, M. and Zinn, J. (1981). Convergence to a stable distribution via order statistics, *Annals of Probability* **9**(4): 624–632.

Levitan, B. M. and Sargsjan, I. S. (1991). *Sturm-Liouville and Dirac operators*, Kluwer Academic Publishers, Dordrecht.

Lockhart, R. A. and Stephens, M. A. (1994). Estimation and tests of fit for the three-parameter Weibull distribution, *Journal of the Royal Statistical Society. Series B (Methodological)* **56**(3): 491–500.

Lockhart, R. A., Spinelli, J. J. and Stephens, M. A. (2007). Cramér-von Mises statistics for discrete distributions with unknown parameters, *Canadian Journal of Statistics* **35**(1): 125–133.

Loynes, R. M. (1980). The empirical distribution function of residuals from generalised regression, *Annals of Statistics* **8**(2): 285–298.

Lucas, C., Kirkwood, G. and Somers, I. (1979). An assessment of the stocks of the banana prawn *Penaeus merguiensis* in the Gulf of Carpentaria, *Australian Journal of Marine and Freshwater Research* **30**(5): 639–652.

Luceño, A. (2007). A universal QQ-plot for continuous non-homogeneous populations, *Journal of Applied Statistics* **34**(10): 1207–1223.

Martynov, G. (2010). Note on the Cramér-von Mises test with estimated parameters, *Publicationes Mathematicae Debrecen* **76**(3-4): 341–364.

Marubini, E., Resele, L. F., Tanner, J. M. and Whitehouse, R. H. (1972). The fit of Gompertz and logistic curves to longitudinal data during adolescence on height, sitting height and biacromial diameter in boys and girls of the Harpenden Growth Study, *Human Biology* **44**(3): 511–523.

May, R. M. (1973). Stability in randomly fluctuating versus deterministic environments, *American Naturalist* **107**(957): 621–650.

Millar, P. W. (1981). Robust estimation via minimum distance methods, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **55**(1): 73–89.

Naka, M. and Shibata, R. (2016). Asymptotic distribution of Cramer-von Mises statistic when contamination exists, *International Journal of Statistics and Probability* **5**(1): 90–97.

Naka, M., Shibata, R. and Darnell, R. (2012). Detection of ecological disturbances to seabed fauna through change of weight distribution, *Journal of the Japan Statistical Society* **42**(2): 185–206.

Owen, D. B. (1980). A table of normal integrals, *Communications in Statistics - Simulation and Computation* **9**(4): 389–419.

Parr, W. C. (1981). Minimum distance estimation: a bibliography, *Communications in Statistics - Theory and Methods* **10**(12): 1205–1224.

Parr, W. C. and Schucany, W. R. (1980). Minimum distance and robust estimation, *Journal of the American Statistical Association* **75**(371): 616–624.

Patel, J. K. and Read, C. B. (1996). *Handbook of the Normal Distribution*, Vol. 150, 2nd edn, Marcel Dekker, New York.

Pierce, D. A. and Kopecky, K. J. (1979). Testing goodness of fit for the distribution of errors in regression models, *Biometrika* **66**(1): 1–5.

Puig, P. and Stephens, M. A. (2000). Tests of fit for the Laplace distribution, with applications, *Technometrics* **42**(4): 417–424.

Puig, P. and Stephens, M. A. (2001). Goodness-of-fit tests for the hyperbolic distribution, *Canadian Journal of Statistics* **29**(2): 309–320.

Richards, F. J. (1969). The quantitative analysis of plant growth, *in* F. C. Steward (ed.), *Analysis of Growth: Behavior of Plants and Their Organs*, Vol. 5A of *Plant Physiology : A treatise*, Academic Press, New York, chapter 1, pp. 3–76.

Riesz, F. and Sz.-Nagy, B. (1990). *Functional analysis*, Dover, New York.

Robins, J. B., Halliday, I. A., Staunton-Smith, J., Mayer, D. G. and Sellin, M. J. (2005). Freshwater-flow requirements of estuarine fisheries in tropical Australia: a review of the state of knowledge and application of a suggested approach, *Marine and Freshwater Research* **56**(3): 343–360.

Rupšys, P. (2007). The relationships between the diameter growth and distribution laws, *WSEAS Transactions on Biology and Biomedicine* **4**(11): 172–191.

Russo, T., Baldi, P., Parisi, A., Magnifico, G., Mariani, S. and Cataudella, S. (2009). Lévy processes and stochastic von Bertalanffy models of growth, with application to fish population analysis, *Journal of Theoretical Biology* **258**(4): 521 – 529.

Sahler, W. (1970). Estimation by minimum-discrepancy methods, *Metrika* **16**(1): 85–106.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples), *Biometrika* **52**(3-4): 591–611.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical processes with applications to statistics*, Wiley, New York.

Smith, F. E. (1963). Population dynamics in *Daphnia magna* and a new model for population growth, *Ecology* **44**(4): 651–663.

Spinelli, J. J. (2001). Testing fit for the grouped exponential distribution, *Canadian Journal of Statistics* **29**(3): 451–458.

Spinelli, J. J. and Stephens, M. A. (1987). Tests for exponentiality when origin and scale parameters are unknown, *Technometrics* **29**(4): 471–476.

Spinelli, J. J. and Stephens, M. A. (1997). Cramér-von Mises tests of fit for the Poisson distribution, *Canadian Journal of Statistics* **25**(2): 257–268.

Staples, D. J. and Heales, D. S. (1991). Temperature and salinity optima for growth and survival of juvenile banana prawns *Penaeus merguiensis*, *Journal of Experimental Marine Biology and Ecology* **154**(2): 251–274.

Steele, M. and Chaseling, J. (2006). Powers of discrete goodness-of-fit test statistics for a uniform null against a selection of alternative distributions, *Communications in Statistics-Simulation and Computation* **35**(4): 1067–1075.

Sukhatme, S. (1972). Fredholm determinant of a positive definite kernel of a special type and its application, *Annals of Mathematical Statistics* **43**(6): 1914–1926.

Székely, G. J. and Rizzo, M. L. (2004). Mean distance test of Poisson distribution, *Statistics & Probability Letters* **67**(3): 241–247.

Tong, L., Yang, J. and Cooper, R. S. (2010). Efficient calculation of p-value and power for quadratic form statistics in multilocus association testing, *Annals of Human Genetics* **74**(3): 275–285.

Tovar-Ávila, J., Troynikov, V. S., Walker, T. I. and Day, R. W. (2009). Use of stochastic models to estimate the growth of the Port Jackson shark, *Heterodontus portusjacksoni,* off eastern Victoria, Australia, *Fisheries Research* **95**(2-3): 230 – 235.

Vance, D. J. and Pendrey, R. C. (2008). Vertical migration of postlarval penaeid prawns in two Australian estuaries: the effect of tide and day/night, *Marine and Freshwater Research* **59**(8): 671–683.

Verhulst, P. F. (1838). Notice sur la loi que la population suit dans son accroissement, *Correspondance Mathematique et Physique* **10**: 113–121.

von Bertalanffy, L. (1960). Principles and theory of growth, *in* W. W. Nowinski (ed.), *Fundamental Aspects of Normal and Malignant Growth*, Elsevier, New York, pp. 137–259.

Wang, Y.-G. and Haywood, M. D. E. (1999). Size-dependent natural mortality of juvenile banana prawns *Penaeus merguiensis* in the Gulf of Carpentaria, Australia, *Marine and Freshwater Research* **50**(4): 313–317.

Wood, C. L. and Altavela, M. M. (1978). Large-sample results for Kolmogorov-Smirnov statistics for discrete distributions, *Biometrika* **65**(1): 235–239.

Woodward, W. A., Parr, W. C., Schucany, W. R. and Lindsey, H. (1984). A comparison of minimum distance and maximum likelihood estimation of a mixture proportion, *Journal of the American Statistical Association* **79**(387): 590–598.