INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

INFERENCE ON THE DIET OF PREDATORS USING FATTY ACID SIGNATURES

by
Connie Stewart

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

AT

DALHOUSIE UNIVERSITY
HALIFAX, NOVA SCOTIA
MAY 26, 2005

DALHOUSIE UNIVERSITY

To comply with the Canadian Privacy Act the National Library of Canada has requested that the following pages be removed from this copy of the thesis:

Preliminary Pages
- Examiners Signature Page (pii)
- Dalhousie Library Copyright Agreement (piii)

Appendices
- Copyright Releases (if applicable)

To James for your constant love and support.

Table of Contents

List of Tables

List of Figures

Abstract

Although methods of accurately estimating the diet of predators are of great ecological importance, prior to the work of Iverson *et al* (2004) the methods used were often unsatisfactory. Iverson *et al* (2004) proposed estimating the diet by matching fatty acid (FA) signatures of predators to those of their prey. Given the potential species in a predator's diet, they were able to use statistical methods to obtain estimates of the proportion of each species in the diet. To date, only point estimates of the diets of predators have been studied.

The primary focus of this thesis is interval estimation of the diet composition. As both the FA data and the diet estimates are compositional, and often with zeros, special techniques are required to handle this situation. Our proposed confidence interval methods include both parametric and nonparametric approaches, and mostly rely on bootstrapping techniques. We make use of mixture models as a device to eliminate the zeros for some of the procedures. A simulation study is carried out to evaluate and compare the coverage probabilities and interval lengths of our various confidence interval methods. Our recommended method is then applied to captive seabird data.

We also consider two related problems, namely the development of a measure of species contribution to the variability in the seal FA signatures and methods for testing for a difference in the diet. The motivation for this latter problem was real-life seal data that we use to illustrate one of our testing procedures.

Acknowledgements

The completion of this thesis would not have been possible without the assistance of many individuals. Firstly, I would like to thank my supervisor Dr. Chris Field for his invaluable expertise and guidance over the past few years. His genuine interest in my progress provided the motivation that I needed to finish. I cannot thank him enough for his patience and kindness. I would also like to express my gratitude to Dr. Duncan Murdoch, my external examiner, for attending my thesis defense and for his highly constructive suggestions. I appreciated his very thorough reading of my thesis. To Dr. Edward Susko and Dr. David Hamilton, my examining committee, thank you for also carefully reading my thesis and providing valuable feedback. Your helpful comments ensured a more polished end product.

Several other individuals at Dalhousie University contributed significantly to my thesis research. I would like to acknowledge Wade Blanchard with whom I spent a considerable amount of time collaborating. Thank you Kassiem Jacobs and Balagopal Pillai for your computer expertise, and biologists Dr. Sara Iverson, Dr. Don Bowen and Dr. Margi Cooper, for your help with my biological questions. Additionally, I would like to thank my officemate, Michele Millar, who, despite her busy schedule, was always ready to help out in any way that she could.

My family and friends were an integral part of this accomplishment and I sincerely appreciated their endless words of encouragement. Furthermore, I would like to thank family members Jill and Rob, and Kevin and Amanda for their generous hospitality in Halifax. To James, your love and support meant more to me than I can express.

Finally, my graduate work would not have been financially possible without funding from the National Sciences and Engineering Research Council of Canada and Dalhousie University, for which I am truly grateful.

Chapter 1

Introduction

In many areas of ecology, knowledge of the diet of predators is crucial. For some predators, direct observation of feeding is possible while for many others, including seals and seabirds, indirect methods are necessary. In the past, indirect methods consisted of estimating the diet by identifying prey structures that are resistant to digestion through the analysis of feces or stomach contents. Because, for example, not all prey have digestion resistant parts (or because these parts may not be consumed by the predator), estimates of diets based on these methods are known to be biased (Iverson *et al*, 2004). Furthermore, any parts recovered may only be representative of the latest meal and not the longer term diet.

More recently, fatty acid (FA) signatures have played a role in diet estimation (Iverson, 1993). In simplified biological terms, FAs are the main constituent of most lipids and are unique in that the FAs released from ingested lipid molecules are not degraded during digestion. Some of these FAs are deposited in the tissue of the predator with little modification. The outcome is that for some predators, the tissue may be a mirror of diet (Iverson *et al*, 1995). The FA signature is then the distribution of all the FAs measured in the predator or prey and the FA signature in the predator reflects the FA signature composition of the prey consumed.

Prior to the work of Iverson *et al* (2004), the use of FA signatures in the diet estimation of predators had been qualitative. In Iverson *et al* (2004), a statistical model was developed to estimate the proportion of prey species in the diet of a predator. Based on the results of a simulation study and some real-life examples (to be discussed), quantitative FA signature analysis (QFASA) was found to be a valuable diet estimation method with several advantages over previous methods.

In this thesis, QFASA is further explored and applications arising from QFASA (such as confidence intervals for the true diet) are examined. In Section 1.2 these

new applications are outlined and the layout of the thesis discussed. As much of this thesis relies on the developments in Iverson *et al* (2004), we begin, however, with a more detailed discussion of the essence of this paper.

1.1 QFASA

Given FA signatures from the species that could potentially be part of the predator's diet, the QFASA estimate of diet contains the estimated proportional contribution of each species to the predator's diet. To obtain the QFASA estimate, the potential prey species in the diet are first summarized by a single FA signature such as the sample mean, as used in Iverson *et al* (2004). The QFASA estimate is then given by the weights that minimize the "distance" between a weighted mixture of these FA signatures and the predator's FA signature. In Iverson *et al* (2004), various distance measures were considered with the Kulback-Liebler distance measure being preferred. Their method is discussed in detail in Section 3.2. For the QFASA method to be useful, a database of potential prey species must be carefully chosen and a few biological issues need to be addressed. Further, to examine properties of the estimates we require the ability to "generate" predators with a known diet. Since much of the QFASA presented in Iverson *et al* (2004) involved seals as the predators, they referred to these generated predators as *pseudo-seals*. These topics, discussed in detail in Iverson *et al* (2004), are now summarized.

Prey Base

A data base containing 28 prey species (954 FA signatures in total), collected along the Scotian Shelf off eastern Canada (see Budge *et al*, 2002), was used in Iverson *et al* (2004). This prey base has since been updated and we were provided with a current prey base in September, 2003 courtesy of Sara Iverson (Dalhousie University). This prey base contains 68 species and 2816 FA signatures in total. From this prey base, only prey from certain areas around the Scotian Shelf were selected for our analyses. (See Appendix A for the specific areas chosen.) After removal of the other FA signatures, a prey base containing 38 species and 1450 FA signatures remained.

Because obtaining a QFASA diet estimate requires optimizing in the dimension equal to the number of species and can be exceptionally time consuming in simulations when carried out in S-PLUS, most of our simulation studies use a reduced prey base containing 8 important species. The choice of the 8 species is discussed shortly under the heading "Simulations". Appendix A contains the 8 species and the sample size of FA signatures associated with each of these species.

It should be noted that in practice the prey base must be chosen to include all potential species that could be in the diet since the QFASA method will always find a "best" estimate. Also essential is that the potential species be distinguishable from their FA signatures. Various multivariate analysis methods can be used to examine the extent to which the FA signatures can be distinguished. While techniques such as hierarchical cluster analysis are useful in determining which species are similar, simulations can also be extremely helpful. Iverson *et al* (2004) suggested systematically removing each species from the prey base, estimating the diet without this species, and observing to which species the weight is re-assigned.

It should also be mentioned that for each FA signature in the prey base, a corresponding measurement of fat content is also recorded. Species with a high fat content will contribute a larger proportion to the FA signature of the predator than those with a low fat content. Consequently, given the average fat content for each species, in practice, the QFASA diet estimate is divided by the fat content and then re-normalized.

Biological Issues

In the development of QFASA, various biological issues arose. As discussed in more detail in Iverson *et al* (2004), it is known that for some FAs, the values in the predator may always be higher or lower than in the prey. For this reason, *calibration factors* were used to adjust the FAs before applying QFASA. In Iverson *et al* (2004), various sets of calibration factors were investigated, all of which were found by comparing the FAs of a predator with a known (experimentally fed) diet to the FAs of the diet (prey). For example, the calibration coefficients that we will use in this thesis

were found by comparing the FA signatures of 8 grey seals fed solely Herring for five months to the FA signatures of 30 randomly selected Herring.

An additional issue, related to calibration, is the choice of FA subset. Not all of the more than 70 FAs contribute equal information about the diet. Some FAs arise only from biosynthesis, some only from diet ("dietary FAs") and some from both ("extended-dietary FAs"). Based on the findings in Iverson *et al* (2004), we have chosen to use the extended-dietary FA subset throughout.

Finally, there are some issues related to predator sampling (see Iverson *et al* (2004) for more detail), but we will assume throughout that the samples of FA signatures obtained are reliable. Note that in seals, the sampling involves a blubber biopsy and is non-lethal. This is clearly an advantage of the QFASA method of estimating the diet versus earlier methods. For example, being non-lethal, QFASA allows for the diet of a predator to be analyzed over time.

Simulations

To investigate the effect of various factors (such as choice of FA subset, distance measure, performance of calibration factors, etc...), Iverson *et al* (2004) carried out a simulation study, using their full prey base containing 28 species, in which pseudo-seals (with and without calibration) were generated using an algorithm similar to that given in Appendix B. Essentially a pseudo-seal was created by choosing a true diet and sampling proportionately with replacement from the species in the diet. To account for the seal eating small amounts of prey not considered to be part of the diet, noise was added to the pseudo-seal by sampling from species not in the specified diet. Using 10% noise gave appropriate results and is the level we will use throughout this thesis.

Based on a hierarchical cluster analysis of the prey FA signatures, Iverson *et al* (2004) constructed four diets from which to generate pseudo seals. Their Diet 1 was considered to be a difficult case (some species were similar) while Diet 4 was meant to represent the diet of a free-ranging grey seal. Due to the computational intensity of the simulations carried out in this thesis, we could not investigate all four diets but

chose to examine Diets 1 and 4. The union of the species in these diets constitute the reduced prey base used in our simulation studies and consists of 8 species.

Experimental Studies

To validate the QFASA diet estimation method, estimates of diet were obtained for predators with a known diet. In one experiment, for example, the diet of captive grey seals which were experimentally fed a diet consisting of three different species was estimated using QFASA and the estimates were consistent with the true diet so long as appropriate calibration factors were used.

QFASA was also applied to free-ranging harbor seals whose true diet had previously been video recorded. In this case, the prey base consisted of all 28 potential species used in the prey base and the results using QFASA were similar to what was observed.

1.2 Thesis Overview

To date point estimates only of the diet of predators based on QFASA have been studied. Since the FA signatures of the prey (both within and between species) and of the predator can be highly variable (leading to a large amount of variability in the diet estimates), an important yet non-trivial problem is then interval estimation of the true diet and much of this thesis is devoted to this issue. We will also introduce a statistic analogous to "R^2" in regression analysis that measures the species contribution to the variability in the predator FA signatures as well as methods of testing for a difference in diet, based on QFASA. These methods involve the use of compositional analysis techniques since the FA signatures and QFASA diet estimates are vectors of proportions that sum to one. We now present an overview of the content of this thesis.

In Chapter 2, an introduction to compositional data analysis is presented. This chapter includes a discussion on parametric modeling of compositional data as well as interval estimation of certain measures of location for both finite and large sample problems. Note that the parametric models that we have encountered in our research

for dealing with compositional data have support which does not include zero and are not adequate for our applications. Consider, for example, modeling the diet estimates obtained using a prey base with only a few of the species actually contributing to the diet. We therefore propose parametric mixture models for modeling compositional data which may contain a significant number of zeros.

Chapter 3 begins with a detailed discussion of the QFASA diet estimation method and possible modified versions of these estimates. We then examine the closeness of various measures of location of these diet estimates to the true diet through a simulation study. In this chapter we also apply the discussions in Chapter 2 to the diet estimates and explore parametric modeling of the diet estimates. Given a sample of predator FA signatures, finite and large sample properties of point estimators of the diet are discussed as well. Finally, a maximum likelihood approach to quantitatively estimating the diet of a predator is investigated, but we will show that this method can be problematic for our application.

Confidence interval methods for the true diet of a predator or group of predators are discussed in Chapter 4. For reasons addressed in this chapter, the confidence intervals developed in Chapter 2 cannot be directly applied to the diet estimation problem and bootstrap procedures are needed for most of the methods. A simulation study is carried out in which the coverage probability and length of the confidence intervals are compared and discussed. The recommended confidence interval method is then applied to a real-life example consisting of FA signatures from captive seabirds.

A statistic that measures how well the variability in the predator FA signatures can be explained by the prey FA signatures is proposed in Chapter 5. Simulation results are presented to validate our our choice of statistic. A potential application of this statistic is reducing the number of possible species in the diet. Results of a simulation study in which our elimination procedure is applied to 27 species is discussed as well.

In Chapter 6 methods for testing for a difference in the diet of two independent or paired samples of predators are investigated. We consider the case where only FA signatures of the predators are given as well as the case when a prey base is also supplied so that, in this case, the test can be carried out using the diet estimates. In

addition to the data being compositional, a critical difficulty is that the dimension of our data may be much larger than the number of observations in our samples and typical multivariate analysis techniques cannot be used. We propose using nonparametric permutation tests and examine the probability of Type I errors and power associated with our tests through simulations. A real-life example containing the before and after FA signatures of two independent samples of seals is also presented.

Finally, in Chapter 7 we present a summary of our results and recommendations. Future work in this area is also addressed.

Chapter 2

Compositional Data Analysis

As described in Chapter 1, the estimated diet of a predator will be a vector of proportions determined from the FA signatures of the predator and its prey. These FA signatures are themselves vectors of proportions. Aitchison (1986) called data of this form *compositional*. In analyzing compositional data, he advised against applying standard multivariate techniques directly to the data, due to the unit sum constraint, and suggested using more appropriate modified methods. This chapter contains a general discussion of compositional data analysis based mostly on Aitchison (1986) and includes selected definitions and theorems applicable to the diet estimation problem.

2.1 Basic Definitions and Notation

Aitchison (1986) defined a *composition* and its *components* as follows:

Definition 2.1 *A* composition $\mathbf{X}$ *of* D parts *is a* $D \times 1$ *vector with positive* components $X_1, \ldots, X_D$ *whose sum is 1.*

As in Aitchison (1986) let $\mathbf{X}^{(c)} = (X_1, \ldots, X_c)$, $\mathbf{X}_{(c)} = (X_{c+1}, \ldots, X_D)$ and $\mathbf{X}_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_D)$. Note that a composition is completely specified by $(X_1, \ldots, X_d)$ where $d = D - 1$, since $X_D = 1 - X_1 - \cdots - X_d$.

The sample space of a D-part composition, termed the *simplex* by Aitchison (1986), is given in the following definition.

Definition 2.2 *The* d*-dimensional* simplex *is the set defined by*

$$\mathcal{S}^d = \{(X_1, \ldots, X_d) : X_1 > 0, \ldots, X_d > 0; X_1 + \cdots + X_d < 1\}.$$

If $\mathcal{R}^d$ and $\mathcal{R}^d_+$ denote d-dimensional real space and d-dimensional positive real space respectively then

$$\mathcal{S}^d \subset \mathcal{R}^d_+ \subset \mathcal{R}^d.$$

Any vector with positive components on the same measurement scale may be transformed into a composition by dividing by the sum of its components. This idea was formalized by Aitchison through Definitions 2.3-2.5:

Definition 2.3 *A* basis $\mathbf{W}$ *of* D *parts is a* $D \times 1$ *vector of positive components* $(W_1, \ldots, W_D)$ *all recorded on the same measurement scale.*

Definition 2.4 *The* size *of a basis* $\mathbf{W}$ *is* $T = W_1 + \cdots + W_D$.

Definition 2.5 *The* constraining operator C *transforms each vector* $\mathbf{W}$ *of* D *positive components into the unit-sum vector* $\mathbf{W}/T$.

Although every basis $\mathbf{W} \in \mathcal{R}_+^d$ yields a unique composition $\mathbf{X} = C(\mathbf{W}) = \mathbf{W}/T$, the converse is not true. That is, there are actually many bases corresponding to a given composition $\mathbf{X}$. In fact, the set of bases $\{T\mathbf{X} : T > 0\}$ all have common composition $\mathbf{X}$.

An additional definition will be needed, namely that of a *perturbation*. Aitchison (1986) defined the *perturbation* operation on compositions as follows:

Definition 2.6 *Let* $\mathbf{X}$ *be a* D*-part composition and* $\mathbf{U}$ *a* D *dimensional vector with positive elements. Then the operation*

$$\mathbf{V} = \mathbf{U} \circ \mathbf{X} = C(X_1U_1, \ldots, X_DU_D)$$

is termed a perturbation *with the original composition* $\mathbf{X}$ *being operated by the* perturbing vector $\mathbf{U}$ *to form a* perturbed composition $\mathbf{V}$.

Before leaving this section, a few remarks are needed concerning the positiveness of the components of a composition. In many applications, including the diet estimation problem, zeros may occur in data that would otherwise be considered compositional. In a recent paper by Martín-Fernández *et al* (2003), the "zero problem", as they called this issue, was addressed. They defined two types of zeros in compositional data, namely *essential* and *rounded* zeros. They defined *essential* zeros to be the "absolute absence of the part in the observation" and *rounded* zeros as the "presence

of a component, but below detection limit". They recommend treating the two types of zeros differently.

When essential zeros are present in a composition, they reasoned that for many problems, either the composition belongs to a different population, or perhaps the zero component is not useful for the study. In the former case, the sample of compositions may be divided into subsamples containing zeros in the same components. The statistical analysis would then be carried out separately on the subsamples.

For rounded zeros their approach was to replace each zero with a small value. They argued that the following *multiplicative replacement strategy* was an appropriate method of replacing the rounded zeros.

Definition 2.7 *Assume* $\mathbf{X}$ *has* Z *zeros. The multiplicative replacement strategy replaces* $\mathbf{X}$ *with* $\mathbf{R} \in \mathcal{S}^d$ *without zeros using the expression*

$$R_i = \begin{cases} \delta_i, & \text{if } X_i = 0, \\ (1 - \sum_{k|X_k=0} \delta_k) X_i, & \text{if } X_i > 0, \end{cases} \tag{2.1}$$

where δ_i *is the imputed value on the component* X_i.

2.2 Difficulties

Aitchison (1986) provided a thorough discussion of the various difficulties encountered in the analysis of compositional data. To aid in motivating the use of Aitchison's methods in the diet estimation problem, a few of these difficulties are outlined.

Typically compositional data occurs in high dimensions, as is the case in the diet estimation problem. This may be problematic as the graphical investigation of a few variables at a time may not be easily interpretable in the simplex. For example, the interpretation of patterns exhibited in say, $\mathcal{R}^2$, may not have the same interpretation in the simplex.

Perhaps a more crucial difficulty with compositional data is, as described by Aitchison (1986), the absence of an interpretable covariance structure. Several examples of why the standard covariance and correlation matrices are poor descriptions of the interdependence of the components of a composition are easily seen. For instance,

note that

$$\begin{aligned}\mathrm{Var}[X_1] + \mathrm{Cov}[X_1, X_2] + \cdots + \mathrm{Cov}[X_1, X_D] &= \mathrm{Cov}(X_1, X_1 + \cdots + X_D) \\ &= \mathrm{Cov}(X_1, 1) \\ &= 0,\end{aligned}$$

or

$$\mathrm{Cov}[X_1, X_2] + \cdots + \mathrm{Cov}[X_1, X_D] = -\mathrm{Var}[X_1].$$

Therefore at least of one of $\mathrm{Cov}[X_1, X_i]$, $i = 2, \ldots, D$ must be negative. Aitchison referred to this phenomena as the "negative bias difficulty".

An additional example is the lack of a relationship between the covariance or correlation matrix of a basis $\mathbf{W}$ and that of its composition $\mathbf{X} = \mathcal{C}(\mathbf{W})$. While it might be expected that $\mathrm{Cor}(X_i, X_j)$ be related to $\mathrm{Cor}(W_i, W_j)$, this is often not the case.

Another challenge in the analysis of compositional data is that of parametric modeling on the simplex. Aitchison (1986) argued that the Dirichlet distribution, defined on $\mathcal{S}^d$, may not be suitable as compositions modeled by this distribution have a strong independence structure. An interesting paper on the history of the Dirichlet distribution is given by Gupta and Richards (2001).

Parametric modeling of compositional data is the topic of Section 2.3.

2.3 Parametric Models

Aitchison (1986) introduced various parametric models defined on $\mathcal{S}^d$. Two of the parametric models proposed by Aitchison are used in the diet estimation procedures, namely the additive and multiplicative logistic normal distributions. Both of these distributions are based on one-to-one transformations from $\mathcal{R}^d$ to $\mathcal{S}^d$. The additive and multiplicative logistic *transformations* were defined as follows by Aitchison (1986):

Definition 2.8 *The* additive logistic transformation *is the one-to-one transformation from* $\mathbf{Y} \in \mathcal{R}^d$ *to* $\mathbf{X} \in \mathcal{S}^d$ *defined by*

$$X_i = \frac{e^{Y_i}}{e^{Y_1} + \cdots + e^{Y_d} + 1}, \; i = 1, \ldots, d,$$

$$X_D = 1 - X_1 - \cdots - X_d = \frac{1}{e^{Y_1} + \cdots + e^{Y_d} + 1},$$

with inverse the additive logratio transformation

$$Y_i = \log(X_i/X_D),\ i = 1, \ldots, d$$

and Jacobian

$$\text{jac}(\mathbf{Y}|\mathbf{X}^{(d)}) = (X_1 \cdots X_D)^{-1}.$$

Definition 2.9 *The* multiplicative logistic transformation *is the one-to-one transformation from* $\mathbf{Y} \in \mathcal{R}^d$ *to* $\mathbf{X} \in \mathcal{S}^d$ *defined by*

$$\begin{aligned} X_i &= \frac{e^{Y_i}}{(1+e^{Y_1}) \cdots (1+e^{Y_i})},\ i = 1, \ldots, d, \\ X_D &= \frac{1}{(1+e^{Y_1}) \cdots (1+e^{Y_d})}, \end{aligned}$$

with inverse the multiplicative logratio transformation,

$$Y_i = \log\left(\frac{X_i}{1 - X_1 - \cdots - X_i}\right),\ i = 1, \ldots, d,$$

and Jacobian

$$\text{jac}(\mathbf{Y}|\mathbf{X}^{(d)}) = (X_1 \cdots X_D)^{-1}.$$

The additive and multiplicative logistic normal distributions are then derived by letting $\mathbf{Y} \in \mathcal{R}^d \sim \mathcal{N}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and considering the distribution of $\mathbf{X} \in \mathcal{S}^d$.

Definition 2.10 *A D-part composition* $\mathbf{X}$ *is said to have an* additive logistic normal *distribution* $\mathcal{L}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *when* $Y_i = \log(X_i/X_D),\ i = 1, \ldots, d$ *has a* $\mathcal{N}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *distribution.*

Definition 2.11 *A D-part composition* $\mathbf{X}$ *is said to have a* multiplicative logistic normal *distribution* $\mathcal{M}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *when* $Y_i = \log\left(\frac{X_i}{1 - X_1 - \cdots - X_i}\right),\ i = 1, \ldots, d$ *has a* $\mathcal{N}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *distribution.*

The forms of the densities $\mathcal{L}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathcal{M}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are easily derived using the $\mathcal{N}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ density function

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}.$$

If $\mathbf{X}$ has density $\mathcal{L}^d(\mu, \Sigma)$ then its density function is given by

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{\frac{1}{2}}(x_1 \cdots x_D)} e^{-\frac{1}{2}(\mathbf{y}-\mu)'\Sigma^{-1}(\mathbf{y}-\mu)},$$

where $\mathbf{y}$ has ith component $\log(x_i/x_D)$. Similarly, if $\mathbf{X}$ has density $\mathcal{M}^d(\mu, \Sigma)$ then its density function is given by

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{\frac{1}{2}}(x_1 \cdots x_D)} e^{-\frac{1}{2}(\mathbf{y}-\mu)'\Sigma^{-1}(\mathbf{y}-\mu)},$$

where $\mathbf{y}$ has ith component $\log\left(\frac{x_i}{1-x_1-\cdots-x_i}\right)$.

The connection between the $\mathcal{L}^d$ and $\mathcal{M}^d$ distributions with the $\mathcal{N}^d$ distribution results in several useful properties that are relatively easy to prove using known properties of the $\mathcal{N}^d$ distribution. Aitchison (1986) outlined many of these properties. In particular, two of these which will be later referenced are given in Properties 2.1 and 2.2.

Property 2.1 *Suppose that a D-part composition $\mathbf{X}$ is distributed as $\mathcal{M}^d(\mu, \Sigma)$ and let the $(c, D-C)$ partition of μ and Σ be*

$$\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Then $(\mathbf{X}^{(c)}, \mathbf{j}'_{D-C}\mathbf{X}_{(c)})$ is distributed as $\mathcal{M}^c(\mu_1, \Sigma_{11})$, where $\mathbf{j}'_{D-C} = [1, 1, \ldots, 1]$ of length $D-C$.

Property 2.2 *A D-part composition $\mathbf{X}$, which is $\mathcal{L}^d(\mu, \Sigma)$ distributed, is perturbed by a vector $\mathbf{U}$ of D positive components, distributed independently of $\mathbf{X}$. The distribution of the perturbed vector $\mathbf{V} = \mathbf{U} \circ \mathbf{X}$ is as given below for two different distributional assumptions about $\mathbf{U}$.*

Distribution of $\mathbf{U}$	Distribution of $\mathbf{V}$
Constant Vector	$\mathcal{L}^d(\mu + \log\left(\frac{\mathbf{U}_{-D}}{U_D}\right), \Sigma)$
$\mathcal{L}^d(\theta, \Theta)$	$\mathcal{L}^d(\mu + \theta, \Sigma + \Theta)$

In Mateu-Figueras *et al* (1998), the class of $\mathcal{L}^d$ distributions was extended to include a shape parameter to allow some skewness in the transformed data to be

present. Essentially, the logratio transformed data were modeled with Azzalini and Capitanio's (1999) multivariate skew-normal distribution instead of the $\mathcal{N}^d$ distribution. Specifically, the multivariate skew-normal and additive logistic skew-normal distributions were defined by Azzalini and Capitanio (1999) and Mateu-Figueras *et al* (1998) respectively as follows:

Definition 2.12 *A d-dimensional random vector* $\mathbf{Y}$ *is said to have a* multivariate skew-normal *distribution* $\mathcal{SN}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ *if it is continuous with density function*

$$f(\mathbf{Y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}) = 2\mathcal{N}^d(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi(\boldsymbol{\alpha}'\boldsymbol{\omega}^{-1}(\mathbf{y} - \boldsymbol{\mu})),$$

where $\Phi(\cdot)$ *denotes the* $N(0,1)$ *distribution function,* $\boldsymbol{\omega}$ *the diagonal matrix with standard deviations of the* $\boldsymbol{\Sigma}$ *diagonal and* $\boldsymbol{\alpha}$ *a d-dimensional shape parameter.*

In the univariate case, Y has a skew-normal distribution $\mathcal{SN}(\mu, \sigma^2, \alpha)$ if

$$f(Y; \mu, \sigma^2, \alpha) = 2\mathcal{N}(y; \mu, \sigma^2)\Phi\left(\frac{\alpha(y-\mu)}{\sigma}\right).$$

Definition 2.13 *A D-part composition* $\mathbf{X}$ *is said to have an* additive logistic skew-normal *distribution* $\mathcal{LS}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$, *when* $\mathbf{Y}$ *with ith component* $Y_i = \log(X_i/X_D)$, $i = 1, \ldots, d$ *has a* $\mathcal{SN}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ *distribution.*

We could similarly define the multiplicative logistic skew-normal distribution as follows:

Definition 2.14 *A D-part composition* $\mathbf{X}$ *is said to have a* multiplicative logistic skew-normal *distribution* $\mathcal{MS}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$, *when* $\mathbf{Y}$ *with ith component* $Y_i = \log\left(\frac{X_i}{1-X_1-\cdots-X_i}\right)$, $i = 1, \ldots, d$ *has a* $\mathcal{SN}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ *distribution.*

Observe that when $\boldsymbol{\alpha} = \mathbf{0}$, $\mathbf{Y} \sim \mathcal{N}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in Definitions 2.13 and 2.14. Also, in the univariate case, as the magnitude of α increases, so does the skewness of the distribution.

It should be remarked that various authors have examined extensions and generalizations to the $\mathcal{SN}^d$ distribution. Consider, for example, the recent work of Marc Genton and in particular, Ma and Genton (2004).

In many applications, including the diet estimation problem, essential zeros may be present and the parametric distributions defined thus far must be modified to accommodate these problems. Our overall strategy for dealing with such zeros involves dividing the compositions into populations according to where the zeros occur and defining separate parametric models within each population. We now consider this strategy in more detail and derive a mixture distribution for the general case of $\mathbf{X}$ being a composition with possibly some essential zero components.

Let $\mathbf{V}$ denote the vector of indices indexing the non-zero components of $\mathbf{X}$ and let $\mathbf{X_V}$ denote the vector containing the non-zero components of $\mathbf{X}$. Suppose that $f_{\mathbf{V}}(\mathbf{x_v})$ is the density of $\mathbf{X_V}$. Then $f_{\mathbf{V}}(\mathbf{x_v})$ may be $\mathcal{L}^d$ or $\mathcal{M}^d$, for example. To model $\mathbf{X}$, it is assumed that there are separate populations for every possible value of $\mathbf{V}$. Let $\theta_{\mathbf{v}} = P[\mathbf{V} = \mathbf{v}]$, the marginal probability that an observation comes from the population with non-zero components indexed by $\mathbf{v}$, where $\sum_{b=1}^{B} \theta_{\mathbf{v}_b} = 1$ and B denotes the number of populations.

Consider the joint density of $\mathbf{X}$ and $\mathbf{V}_b$

$$\begin{aligned} f_{\mathbf{X},\mathbf{V}_b}(\mathbf{x}, \mathbf{v}_b) &= f_{\mathbf{x}|\mathbf{V}_b}(\mathbf{x}|\mathbf{v}_b)\theta_{\mathbf{v}_b} \\ &= f_{\mathbf{V}_b}(\mathbf{x}_{\mathbf{v}_b})\theta_{\mathbf{v}_b}. \end{aligned}$$

Then

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}) &= \sum_{b=1}^{B} f_{\mathbf{x},\mathbf{V}_b}(\mathbf{x}, \mathbf{v}_b) \\ &= \sum_{b=1}^{B} f_{\mathbf{V}_b}(\mathbf{x}_{\mathbf{v}_b})\theta_{\mathbf{v}_b}. \end{aligned} \tag{2.2}$$

Observe that if the non-zero components of $\mathbf{x}$ are indexed by $\mathbf{v}$, then $f_{\mathbf{V}_b}(\mathbf{x}_{\mathbf{v}_b}) = 0\ \forall b$ such that $\mathbf{v}_b \neq \mathbf{v}$. The sum in Equation 2.2 then has only one non-zero term since only one of the populations will correspond to the non-zero components of $\mathbf{x}$.

It will be useful to also derive the marginal distributions of the components of $\mathbf{X}$ when $\mathbf{X}$ has the density in Equation 2.2. The derivation will be carried out by first integrating $f_{\mathbf{x},\mathbf{V}_b}(\mathbf{x}, \mathbf{v}_b)$ to obtain the joint density of x_i and $\mathbf{v}_b$, $f_i(x_i, \mathbf{v}_b)$, and then by summing this distribution over all possible $\mathbf{v}_b$.

Integrating $f_{\mathbf{x},\mathbf{v}_b}(\mathbf{x},\mathbf{v}_b)$ over the x_j, $j \neq i$ gives

$$f_i(x_i,\mathbf{v}_b) = \begin{cases} \theta_{\mathbf{v}_b} & \text{if } x_i = 0, i \notin \mathbf{v}_b, \\ \theta_{\mathbf{v}_b} \int \cdots \int f_{\mathbf{V}_b}(\mathbf{x}_{\mathbf{v}_b})\, d\mathbf{x}_{-i} & \text{if } 0 < x_i < 1, i \in \mathbf{v}_b, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\int \cdots \int f_{\mathbf{V}_b}(\mathbf{x}_{\mathbf{v}_b})\, d\mathbf{x}_{-i}$, is the marginal distribution of the ith component of $\mathbf{x}_{\mathbf{V}_b}$, if $i \in \mathbf{v}_b$. By property 2.1, recall that if $(x_1,\ldots,x_D) \sim \mathcal{M}(\mu,\Sigma)$, then $(x_1, \sum_{i=2}^{D} x_i) \sim \mathcal{M}(\mu_1,\sigma_{11})$. That is, the marginal distribution of x_1 is $\mathcal{M}(\mu_1,\sigma_{11})$. To apply this property to $\mathbf{x}_{\mathbf{v}_b}$, re-order the components of $\mathbf{x}$ as $(x_i,x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_D)$ so that if $i \in \mathbf{v}_b$, $\mathbf{x}_{\mathbf{v}_b} = (x_i,\mathbf{x}_{\mathbf{v}_{b-i}})$. Now assume that $(x_i,\mathbf{x}_{\mathbf{v}_{b-i}}) \sim \mathcal{M}(\boldsymbol{\mu}^{\mathbf{v}_b},\boldsymbol{\Sigma}^{\mathbf{v}_b})$. Let $\mu_i^{\mathbf{v}_b} = \mu^{\mathbf{v}_b}[1]$ and $\Sigma^{\mathbf{v}_b}[1,1] = \sigma_i^{2\mathbf{v}_b}$. Then

$$(x_i, 1-x_i) \sim \mathcal{M}(\mu_i^{\mathbf{V}_b}, \sigma_i^{2\mathbf{V}_b})$$

and

$$\begin{aligned} f_i(x_i) &= \sum_{b=1}^{B} f_i(x_i,\mathbf{v}_b) \\ &= \sum_{\{b:i\notin\mathbf{v}_b\}} f_i(x_i,\mathbf{v}_b) + \sum_{\{b:i\in\mathbf{v}_b\}} f_i(x_i,\mathbf{v}_b) \\ &= \begin{cases} \sum_{\{b:i\notin\mathbf{v}_b\}} \theta_{\mathbf{v}_b} & \text{if } x_i = 0, \\ \sum_{\{b:i\in\mathbf{v}_b\}} \theta_{\mathbf{v}_b} \mathcal{M}(\mu_i^{\mathbf{V}_b}, \sigma_i^{2\mathbf{V}_b}) & \text{if } 0 < x_i < 1, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{2.3}$$

Realize that although the expression for the discrete probability at zero may appear complicated, $P[X_i = 0]$ is simply the sum of the probabilities associated with those populations having ith component zero.

We will make the simplifying assumption that $\sigma_i^{2\mathbf{v}_b} = \sigma_i^2$ $\forall b = 1,\ldots,B$. Then for the ith component, a marginal mixture distribution that may be used to model compositional data with some zero components is given by

$$f_i(x_i) = \begin{cases} \sum_{\{b:i\notin\mathbf{v}_b\}} \theta_{\mathbf{v}_b} & \text{if } x_i = 0, \\ \sum_{\{b:i\in\mathbf{v}_b\}} \theta_{\mathbf{v}_b} \mathcal{M}(\mu_i^{\mathbf{V}_b}, \sigma_i^2) & \text{if } 0 < x_i < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

We will write $X_i \sim MixM(\theta_{\mathbf{v}_b}, \mu_i^{\mathbf{v}_b}, \sigma_i^2)$ (or say that X_i is $MixM$ distributed) if X_i has the density in Equation 2.4.

If it is further assumed that $\mu_i^{\mathbf{v}_b} = \mu_i \; \forall b = 1, \ldots, B$ then the density in Equation 2.4 becomes the simpler density

$$f_i(x_i) = \begin{cases} \sum_{\{b:i \notin \mathbf{v}_b\}} \theta_{\mathbf{v}_b} & \text{if } x_i = 0, \\ \mathcal{M}(\mu_i, \sigma_i^2) \left(\sum_{\{b:i \in \mathbf{v}_b\}} \theta_{\mathbf{v}_b} \right) & \text{if } 0 < x_i < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

Let $\theta_i = \sum_{\{b:i \notin \mathbf{v}_b\}} \theta_{\mathbf{v}_b}$ then since $\sum_{\{b:i \notin \mathbf{v}_b\}} \theta_{\mathbf{v}_b} + \sum_{\{b:i \in \mathbf{v}_b\}} \theta_{\mathbf{v}_b} = 1$, we have

$$f_i(x_i) = \begin{cases} \theta_i & \text{if } x_i = 0, \\ (1 - \theta_i)\mathcal{M}(\mu_i, \sigma_i^2) & \text{if } 0 < x_i < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.6}$$

If X_i has the density in Equation 2.6, we will use the notation, $X_i \sim SMixM(\theta_i, \mu_i, \sigma_i^2)$.

With the $SMixM$ model, we are, in effect, simply assuming that $X_i = 0$ with probability θ_i and that $\log\left(\frac{X_i}{1-X_i}\right) \sim N(\mu_i, \sigma_i^2)$, $X_i > 0$. While a disadvantage of using Equation 2.6 is that we are not utilizing any information provided by the other components, we will show that for the diet estimation problem, modeling with this simpler density often provides results that are very similar to those obtained with the more complicated density in Equation 2.4. As will also be shown, an improved fit can be obtained by replacing $\mathcal{M}(\mu_i, \sigma_i^2)$ in Equation 2.6 with its skew extension, $\mathcal{SM}(\mu_i, \sigma_i^2)$. We define the $SMixSM(\theta_i, \mu_i, \sigma_i^2, \alpha_i)$ density function as

$$f_i(x_i) = \begin{cases} \theta_i & \text{if } x_i = 0, \\ (1 - \theta_i)\mathcal{SM}(\mu_i, \sigma_i^2, \alpha_i) & \text{if } 0 < x_i < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.7}$$

Note that in our application, the $\mathcal{L}^d$ and $\mathcal{M}^d$ distributions (as well as their corresponding skew extensions) may be used to model the seal and/or prey FA signatures. In this case, D is the number of FAs in the signature and the notation n_{FA} (in place of D) will be adopted in Chapter 3. In addition to modeling the FA signatures themselves, we will also be interested in modeling the estimated diet, say $\mathbf{P}$. It is in this case that the mixture distributions will be needed since it will be possible for any of

the components of $\mathbf{P}$ to be zero. Based on the discussions in Section 2.1, these zeros correspond to essential zeros and represent the absence of the species in the diet.

Finally, observe that for all of the parametric densities discussed, $P[X_i = 1] = 0$, $i = 1, \ldots, D$. Although we could modify the densities to allow for a non-zero probability at one, in our applications either X_i will represent the ith FA in a signature or the diet estimate of the ith species. In both cases, we will likely have $P[X_i = 1] \approx 0$. If, however, it was thought that $P[X_i = 1]$ was non-zero and we were interested in, say, interval estimation of the diet of the ith species, then we could apply one of our nonparametric procedures to be discussed in Chapter 4.

2.4 Measures of Location

Before considering estimation procedures for compositional data modeled by one of the parametric densities defined in Section 2.3, it will be helpful to first specify parameters for which inferences will eventually be required. For the diet estimation application, the parameters of interest will be measures of location (MOLs) since it will be shown in Section 3.3 that certain MOLs tend to be close to the true diet. This is beneficial as it allows parametric inference procedures to be developed for an otherwise nonparametric quantity, namely the true diet of a predator. In this section, typical MOLs, such as the mean and median, are discussed, as well as some not so typical MOLs that may be better suited for compositional data. Recall that when the data represent diet estimates, I will replace D in the derivations below and will correspond to the number of species in the diet.

Consider first the mean, E[X], a popular measure of location (MOL). Assume first that $\mathbf{X} \sim \mathcal{L}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $\mathbf{Y}$ has ith component $Y_i = \log(X_i/X_D)$, $i = 1, \ldots, d$, then

$$\begin{aligned} \mathrm{E}[\mathbf{Y}] &= \boldsymbol{\mu} && (2.8) \\ &= \left[\mathrm{E}\left[\log\left(\frac{X_1}{X_D}\right)\right], \ldots, \mathrm{E}\left[\log\left(\frac{X_d}{X_D}\right)\right]\right]. && (2.9) \end{aligned}$$

Let

$$\mathrm{MOL}^{AL} = \left[\frac{e^{\mu_1}}{e^{\mu_1} + \cdots + e^{\mu_d} + 1}, \cdots, \frac{e^{\mu_d}}{e^{\mu_1} + \cdots + e^{\mu_d} + 1}, \frac{1}{e^{\mu_1} + \cdots + e^{\mu_d} + 1}\right],$$

then by the Delta method,

$$\mathrm{E}[\mathbf{X}] \approx \mathrm{MOL}^{AL}.$$

It is interesting to note that MOL^{AL} is essentially the population version of the sample MOL recommended in Aitchison (1989). He argued that for a sample of compositions, the appropriate sample MOL is simply MOL^{AL} with μ_i replaced by $\bar{Y}_i$. Based on Aitchison's recommendation, when $\mathbf{X} \sim \mathcal{L}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, MOL^{AL} may be favoured over $\text{E}[\mathbf{X}]$ though they should be similar by the Delta method.

If $\mathbf{X} \sim \mathcal{M}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ instead, and if $\mathbf{Y}$ now has ith component $Y_i = \log\left(\frac{X_i}{1-X_1-\cdots-X_i}\right)$, $i = 1, \ldots, d$, then a similar argument would lead to

$$\text{E}[\mathbf{X}] \approx \text{MOL}^{ML},$$

where

$$\text{MOL}^{ML} = \left[\frac{e^{\mu_1}}{1+e^{\mu_1}}, \cdots, \frac{e^{\mu_d}}{(1+e^{\mu_1})+\cdots+(1+e^{\mu_d})}, \frac{1}{(1+e^{\mu_1})\cdots(1+e^{\mu_d})}\right].$$

An appealing feature of MOL^{ML} is that the first component, $\text{MOL}^{ML}[1] = \frac{e^{\mu_1}}{1+e^{\mu_1}}$, doesn't depend on any other parameters. In contrast, each component of MOL^{AL} contains at least two parameters which will be unknown in practice. If the components are ordered so that the component of interest is first, then inference procedures for $\text{MOL}^{ML}[1]$ should be less complex than for $\text{MOL}^{AL}[1]$.

For compositional data modeled by the density in Equation 2.4, (that is, $X_i \sim Mix\mathcal{M}(\theta_{\mathbf{v}_b}, \mu_i^{\mathbf{v}_b}, \sigma_i^{\mathbf{v}_b})$) an analogous approach yields a MOL that is roughly a weighted sum of $\text{MOL}^{ML}[1]$ within each population. To see this note first that

$$f_i(x_i|\mathbf{v}_b) = \begin{cases} 1 & \text{if } x_i = 0, i \notin \mathbf{v}_b, \\ \mathcal{M}(\mu_i^{\mathbf{v}_b}, \sigma_i^2) & \text{if } 0 < x_i < 1, i \in \mathbf{v}_b, \\ 0 & \text{otherwise,} \end{cases} \tag{2.10}$$

so that for the bth population a natural MOL would be

$$\eta_i(\mathbf{v}_b) = \begin{cases} 0 & \text{if } i \notin \mathbf{v}_b \\ \frac{e^{\mu_i^{\mathbf{v}_b}}}{1+e^{\mu_i^{\mathbf{v}_b}}} & \text{if } i \in \mathbf{v}_b \end{cases}$$

Since

$$\text{E}[X_i] = \sum_{b=1}^{B} \theta_{\mathbf{v}_b} \text{E}[X_i|\mathbf{v}_b],$$

and $\mathrm{E}[X_i|\mathbf{v}_b] \approx \eta_i(\mathbf{v}_b)$ by the Delta method, the intuitive overall MOL is

$$\eta_i = \sum_{b=1}^{B} \theta_{\mathbf{v}_b} \eta_i(\mathbf{v}_b).$$

In a similar manner, it could be argued that if $X_i \sim SMix\mathcal{M}(\theta_i, \mu_i, \sigma_i^2)$ then the appropriate MOL would be

$$\lambda_i = \begin{cases} 0 & \text{if } \theta_i = 1 \\ (1-\theta_i)\frac{e^{\mu_i}}{1+e^{\mu_i}} & \text{if } \theta_i < 1. \end{cases}$$

We have also defined an MOL based on the skew distribution in Equation 2.7. From Azzalini and Dalla Valle (1996), if $Y \sim \mathcal{SN}(\mu, \sigma^2, \alpha)$,

$$\mathrm{E}[Y] = \sigma \left(\frac{2}{\pi}\right)^{\frac{1}{2}} \frac{\alpha}{(1+\alpha^2)^{\frac{1}{2}}} + \mu.$$

Denote this expected value by $\xi(\mu, \sigma^2, \alpha)$, then if $X_i \sim SMix\mathcal{SM}(\theta_i, \mu_i, \sigma_i^2, \alpha_i)$, an appropriate MOL would be

$$\lambda_i^s = \begin{cases} 0 & \text{if } \theta_i = 1 \\ (1-\theta_i)\frac{e^{\xi_i(\mu_i, \sigma_i^2, \alpha_i)}}{1+e^{\xi_i(\mu_i, \sigma_i, \alpha_i)}} & \text{if } \theta_i < 1. \end{cases}$$

Note that if X_i is $SMix\mathcal{SM}$ distributed, then $\lambda_i = \lambda_i^s$.

In Section 3.3, the median will also prove to be a useful MOL under certain circumstances. We will attempt to derive the median when X_i is modeled by the $\mathcal{M}$, $Mix\mathcal{M}$ and the $SMix\mathcal{M}$ distributions. Consider first the median, say M_i, of a component X_i from a composition $\mathbf{X}$ if $(X_i, (1-X_i)) \sim \mathcal{M}(\mu_i, \sigma_i^2)$. We must solve $P[X_i \leq M_i] \geq \frac{1}{2}$ and $P[X_i \geq M_i] \geq \frac{1}{2}$ for M_i. Since, in this case, X_i is continuous, M_i satisfies

$$\int_0^{M_i} \mathcal{M}(\mu_i, \sigma_i^2) = \int_{M_i}^1 \mathcal{M}(\mu_i, \sigma_i^2) = \frac{1}{2}.$$

We have

$$\int_0^{M_i} \mathcal{M}(\mu_i, \sigma_i^2) = P\left[Y_i \leq \log\left(\frac{M_i}{1-M_i}\right)\right] = \frac{1}{2},$$

where $Y_i \sim N(\mu_i, \sigma_i^2)$ and therefore,

$$\log\left(\frac{M_i}{1-M_i}\right) = \mu_i \Rightarrow M_i = \frac{e^{\mu_i}}{1+e^{\mu_i}}.$$

The median in this case is actually $\mathrm{MOL}^{ML}[1]$.

Now suppose that $X_i \sim Mix\mathcal{M}(\theta_{\mathbf{v}_b}, \mu_i^{\mathbf{v}_b}, \sigma_i^2)$. Note that

$$M_i = 0 \Leftrightarrow P[X_i = 0] = \sum_{\{b:i\notin\mathbf{v}_b\}} \theta_{\mathbf{v}_b} \geq \frac{1}{2}.$$

(If $M_i = 0$ then $P[X_i \leq 0] = P[X_i = 0] \geq \frac{1}{2}$. Conversely, if $P[X_i = 0] \geq \frac{1}{2}$, then $M_i = 0$ satisfies the equations $P[X_i \leq M_i] = P[X_i = 0] + P[0 < X_i < M_i] \geq \frac{1}{2}$ and $P[X_i \geq M_i] \geq \frac{1}{2}$.) When $P[X_i = 0] < \frac{1}{2}$ (and $M_i > 0$), the case is not so trivial and there is no closed form solution. M_i must satisfy the following Equations

$$\begin{aligned} P[X_i \leq M_i] &= P[X_i = 0] + P[0 < X_i \leq M_i] \\ &= \sum_{\{b:i\notin\mathbf{v}_b\}} \theta_{\mathbf{v}_b} + \sum_{\{b:i\in\mathbf{v}_b\}} \theta_{\mathbf{v}_b} \int_0^{M_i} \mathcal{M}(x_i, \mu_i^{\mathbf{v}_b}, \sigma_i^2) dx_i \\ &= \sum_{\{b:i\notin\mathbf{v}_b\}} \theta_{\mathbf{v}_b} + \sum_{\{b:i\in\mathbf{v}_b\}} \theta_{\mathbf{v}_b} P\left[Y_i^{\mathbf{v}_b} \leq \log\left(\frac{M_i}{1-M_i}\right)\right] \\ &\geq \frac{1}{2}, \end{aligned}$$

and

$$\begin{aligned} P[X_i \geq M_i] &= \sum_{\{b:i\in\mathbf{v}_b\}} \theta_{\mathbf{v}_b} \int_{M_i}^1 \mathcal{M}(x_i, \mu_i^{\mathbf{v}_b}, \sigma_i^2) dx_i \\ &= \sum_{\{b:i\in\mathbf{v}_b\}} \theta_{\mathbf{v}_b} P[Y_i^{\mathbf{v}_b} \geq \log\left(\frac{M_i}{1-M_i}\right)] \\ &\geq \frac{1}{2}, \end{aligned}$$

where $Y_i^{\mathbf{v}_b} \sim N(\mu_i^{\mathbf{v}_b}, \sigma_i^2)$.

Finally, let X_i be $SMix\mathcal{M}$ distributed. Then similarly,

$$M_i = 0 \Leftrightarrow P[X_i = 0] = \theta_i \geq \frac{1}{2}.$$

When $\theta_i < \frac{1}{2}$ (or $M_i > 0$), we must solve for M_i in the following Equations

$$\begin{aligned} P[X_i \leq M_i] &= P[X_i = 0] + P[0 < X_i \leq M_i] \\ &= \theta_i + (1-\theta_i) \int_0^{M_i} \mathcal{M}(\mu_i, \sigma_i^2) dx_i \\ &= \theta_i + (1-\theta_i) P\left[Y_i \leq \log\left(\frac{M_i}{1-M_i}\right)\right] \\ &\geq \frac{1}{2}, \end{aligned}$$

and

$$\begin{aligned} P[X_i \geq M_i] &= (1-\theta_i)\int_{M_i}^{1} \mathcal{M}(x_i, \mu_i, \sigma_i^2)dx_i \\ &= (1-\theta_i)P\left[Y_i \geq \log\left(\frac{M_i}{1-M_i}\right)\right] \\ &\geq \frac{1}{2}. \end{aligned}$$

It is straightforward to show that $M_i = \frac{e^{\mu_i}}{1+e^{\mu_i}}$ satisfies both equations. Overall, we have

$$M_i = \begin{cases} 0 & \text{if } \theta_i \geq \frac{1}{2} \\ (1-\theta_i)\frac{e^{\mu_i}}{1+e^{\mu_i}} & \text{if } \theta_i < \frac{1}{2}. \end{cases}$$

Although M_i appears similar to λ_i, recall that λ_i is zero for $\theta_i = 1$.

While $\sum_{i=1}^{D} \mathrm{E}[X_i] = 1$, note that $\sum_{i=1}^{D} \eta_i \neq 1$ (and similarly for λ_i and M_i). Although we could simply normalize these parameters (by using $\frac{\eta_i}{\sum_{j=1}^{D}\eta_j}$, for example), deriving a CI for the normalized parameters is often more complicated as they contain several unknown parameters. As will be detailed in Chapter 4, we will adjust our CIs for this potential bias.

2.5 Inference with Compositional Data

2.5.1 Point Estimation

Suppose the observations in a compositional data set, say $\mathbf{X} = [X_{ri} : r = 1, \ldots, N, ; i = 1, \ldots, d]$, are modeled by one of the parametric distributions from Section 2.3. There are then several unknown parameters. For the $\mathcal{L}^d$ and $\mathcal{M}^d$ distributions, μ and Σ are unknown and must be estimated. In the mixture distribution in Equation 2.4 the unknowns are the actual populations themselves and corresponding probabilities $\theta_{\mathbf{v}_b}$ as well as $\mu_i^{\mathbf{v}_b}$ and σ_i^2, $b = 1, \ldots, B$ and $i = 1, \ldots, d$. The unknown parameters in the simpler mixture distribution (Equation 2.6) are θ_i, μ_i and σ_i^2, $i = 1, \ldots, d$ and for the skew distribution (Equation 2.7) there is the additional parameter, α_i, $i = 1, \ldots, d$. Furthermore, the MOLs discussed in Section 2.4 must be estimated. These, however, will be functions of the parameters just mentioned. It is important to realize that when the arguments that follow are applied to the diet estimation problem, N will

correspond to the number of seal (or predator) FA signatures in the sample and the notation n_s will be used (in place of N) in the chapters to follow.

In the analysis of compositional data modeled by the $\mathcal{L}^d$ or $\mathcal{M}^d$ distributions, the general strategy is to transform the data using the corresponding logratio transformations and to apply standard multivariate statistical tools. If $\mathbf{Y}$ (of dimension $N \times d$) contains the transformed data, then it is well-known that the maximum likelihood estimators (MLEs) of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given by $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ with components

$$\begin{aligned} \hat{\mu}_i &= \frac{1}{N}\sum_{r=1}^{N} Y_{ri},\ i = 1, \ldots d, \text{ and} \\ \hat{\sigma}_{ij} &= \frac{1}{N}\sum_{r=1}^{N}(Y_{ri} - \hat{\mu}_i)(Y_{rj} - \hat{\mu}_j),\ i, j = 1, \ldots, d. \end{aligned}$$

Note that in the estimation procedures discussed in the chapters to follow, $N - 1$ is used in estimating σ_{ij} so that the estimate is unbiased. We will let $\mathbf{S}$ have i,jth component $S_{ij} = \frac{N\hat{\sigma}_{ij}}{N-1}$ and $S_i^2 = S_{ii}$.

Now suppose that $X_{1i}, \ldots, X_{Ni}$ are independent and identically distributed (iid) $Mix\mathcal{M}(\theta_{\mathbf{v}_b}, \mu_i^{\mathbf{v}_b}, \sigma_i^2)$ random variables. We will obtain the MLEs using the joint distribution of $\mathbf{X}$ given in Equation 2.2, where

$$f_{\mathbf{V}_b}(\mathbf{x}_{\mathbf{v}_b}) = \begin{cases} \mathcal{M}^d(\boldsymbol{\mu}^{\mathbf{v}_b}, \boldsymbol{\sigma}^{\mathbf{v}_b}), & \text{if } \mathbf{v}_b = \mathbf{v} \\ 0 & \text{otherwise,} \end{cases} \tag{2.11}$$

and $\mathbf{v}$ indexes the non-zero components of $\mathbf{x}$. For a sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$ that are iid with joint density given by Equations 2.2 and 2.11, let $N_{\mathbf{v}_b}$ denote the number of observations in the sample from group b. The MLEs maximize the likelihood function,

$$\begin{aligned} L(\theta_{\mathbf{v}_b}, \boldsymbol{\mu}^{\mathbf{v}_b}, \boldsymbol{\Sigma}^{\mathbf{v}_b}, b = 1, \ldots, B) &= \prod_{\{r:\mathbf{x}_r \in \text{ Group } 1\}} \theta_{\mathbf{v}_1} f_{\mathbf{V}_1}(\mathbf{x}_{r\mathbf{v}_1}) \\ &\cdots \prod_{\{r:\mathbf{x}_r \in \text{ Group } B\}} \theta_{\mathbf{v}_p} f_{\mathbf{V}_p}(\mathbf{x}_{r\mathbf{v}_p}) \\ &= \prod_{\{r:\mathbf{x}_r \in \text{ Group } 1\}} \theta_{\mathbf{v}_1} \mathcal{M}(\mathbf{x}_{r\mathbf{v}_1}, \boldsymbol{\mu}^{\mathbf{v}_1}, \boldsymbol{\Sigma}^{\mathbf{v}_1}) \\ &\cdots \prod_{\{r:\mathbf{x}_r \in \text{ Group } B\}} \theta_{\mathbf{v}_B} \mathcal{M}(\mathbf{x}_{r\mathbf{v}_B}, \boldsymbol{\mu}^{\mathbf{v}_B}, \boldsymbol{\Sigma}^{\mathbf{v}_B}), \end{aligned}$$

or equivalently the log likelihood function

$$\log L(\theta_{\mathbf{v}_b}, \boldsymbol{\mu}^{\mathbf{v}_b}, \boldsymbol{\Sigma}^{\mathbf{v}_b}, b = 1, \ldots, B) = N_{\mathbf{v}_1} \log \theta_{\mathbf{v}_1} \tag{2.12}$$

$$+ \sum_{\{r:\mathbf{x}_r \in \text{ Group } 1\}} \log \mathcal{M}(\mathbf{x}_{r\nu_1}, \boldsymbol{\mu}^{\nu_1}, \boldsymbol{\Sigma}^{\nu_1})$$
$$+ \cdots + N_{\nu_B} \log \theta_{\nu_B}$$
$$+ \sum_{\{r:\mathbf{x}_r \in \text{ Group } B\}} \log \mathcal{M}(\mathbf{x}_{r\nu_B}, \boldsymbol{\mu}^{\nu_B}, \boldsymbol{\Sigma}^{\nu_B}),$$

subject to $\sum_{b=1}^{B} \theta_{\nu_b} = 1$ and $\sum_{b=1}^{B} N_{\nu_b} = N$.

With the populations essentially separated in Equation 2.12 as well as θ_{ν_b} being separated from $\boldsymbol{\mu}^{\nu_b}$ and $\boldsymbol{\Sigma}^{\nu_b}$, $b = 1, \ldots, B$, the MLEs are readily obtainable. Let $\mathbf{Y}^{\nu_b}$ be the matrix containing the transformed, non-zero observations from group b. Then the MLEs are

$$\begin{aligned} \hat{\theta}_{\nu_b} &= \frac{N_{\nu_b}}{N} \qquad (2.13) \\ \hat{\mu}_i^{\nu_b} &= \frac{1}{N_{\nu_b}} \sum_{r=1}^{N_{\nu_b}} Y_{ri}^{\nu_b}, \; i = 1, \ldots d, \text{ and} \\ \hat{\sigma}_{ik}^{\nu_b} &= \frac{1}{N_{\nu_b}} \sum_{r=1}^{N_{\nu_b}} (Y_{ri}^{\nu_b} - \hat{\mu}_i^{\nu_b})(Y_{rk}^{\nu_b} - \hat{\mu}_k^{\nu_b}), \; i, k = 1, \ldots, d. \end{aligned}$$

Note that we are implicitly estimating the number of populations, B, since $\hat{\theta}_{\nu_b} = 0$ if population b does not occur in the sample.

For the *marginal* mixture density in Equation 2.3, the MLEs of $\mu_i^{\nu_b}$ and $\sigma_i^{2\nu_b}$ are then respectively

$$\begin{aligned} \hat{\mu}_i^{\nu_b} &= \frac{1}{N_{\nu_b}} \sum_{r=1}^{N_{\nu_b}} \log\left(\frac{X_{ri}^{\nu_b}}{1 - X_{ri}^{\nu_b}}\right). \qquad (2.14) \\ \hat{\sigma}_i^{2\nu_b} &= \frac{1}{N_{\nu_b}} \sum_{r=1}^{N_{\nu_b}} \left[\log\left(\frac{X_{ri}^{\nu_b}}{1 - X_{ri}^{\nu_b}}\right) - \hat{\mu}_i^{\nu_b}\right]^2 \end{aligned}$$

As before, we will actually estimate $\sigma_i^{2\nu_b}$ by $s_i^{2\nu_b} = \frac{N_{\nu_b}}{N_{\nu_b}-1} \hat{\sigma}_i^{2\nu_b}$.

As in Equation 2.4, where $\sigma_i^{\nu_b} = \sigma_i^2 \; \forall b$ we will estimate the common variance by

$$S_i^{2\text{pool}} = \frac{(N_{\nu_1} - 1)S_i^{2\nu_1} + \cdots + (N_{\nu_B} - 1)S_i^{2\nu_B}}{N_{\nu_1} + \cdots + N_{\nu_B} - B}.$$

For an iid sample, $X_{1i}, \ldots, X_{Ni}$ from $SMix\mathcal{M}(\theta_i, \mu_i, \sigma_i^2)$, the log likelihood equation is

$$\log L(\theta_i, \mu_i, \sigma_i^2) = (N - N') \log \theta_i + N' \log \theta_i \sum_{\{r:x_{ri} > 0\}} \log \mathcal{M}(x_{ri}, \mu_i, \sigma_i^2), \qquad (2.15)$$

where N' = number of non-zero observations in the sample. The MLEs can be observed directly and are given by

$$\hat{\theta}_i = \frac{N - N'}{N} \tag{2.16}$$

$$\hat{\mu}_i = \frac{1}{N'} \sum_{r=1}^{N'} \log\left(\frac{X'_{ri}}{1 - X'_{ri}}\right)$$

$$\hat{\sigma}_i^2 = \frac{1}{N'} \sum_{r=1}^{N'} \left[\log\left(\frac{X'_{ri}}{1 - X'_{ri}}\right) - \hat{\mu}_i\right]^2, \tag{2.17}$$

where X'_{ri} denotes the rth non-zero sample observation. (We will prefer to use the estimate $S_i^2 = \frac{N'}{N'-1}\hat{\sigma}_i^2$ of σ_i^2.)

It should be mentioned that ML estimation of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}$ in the $\mathcal{SM}^d$ distribution requires numerical methods. The library *sn*, written by Adelchi Azzalini for S-PLUS, contains functions to compute the ML estimates. The extension to the $SMix\mathcal{SM}$ distribution in Equation 2.7 is straightforward. We will estimate θ_i as in Equation 2.16 and then use the relevant *sn* S-PLUS functions to fit the $\mathcal{SN}$ distribution (by ML estimation) to the non-zero transformed observations.

Having presented the MLEs for the parameters in the parametric models of Section 2.3, we may now specify point estimators for the MOLs discussed in Section 2.4. Table 2.1 contains the MOLs of interest and corresponding choices of estimators. For notational convenience, note that in $\hat{\eta}_k$, the sum is over all groups but $\hat{\theta}_{\mathbf{v}_b} = 0$ for groups not occurring in the sample. Observe also that $\hat{\eta}_i$, $\hat{\lambda}_i$ and $\hat{\lambda}_i^s$, being functions of MLEs, are the MLEs of η_i, λ_i and λ_i^s respectively when X_{ri}, $r = 1, \ldots N$ are assumed to be $Mix\mathcal{M}$, $SMix\mathcal{M}$, or $SMix\mathcal{SM}$ distributed. Additionally, often when $N' \leq 2$, S-PLUS has difficulty computing the MLEs of μ_i, σ_i, and α_i in the $SMix\mathcal{SM}$ distribution. We will therefore let $\hat{\lambda}_k^s = \hat{\lambda}_k$ when $N' \leq 2$. Finally, we will estimate M_i by the sample median, m_i. While we could have alternatively chosen, say,

$$\hat{M}_i = \begin{cases} 0 & \text{if } \hat{\theta}_i \geq \frac{1}{2} \\ (1 - \hat{\theta}_i)\frac{e^{\hat{\mu}_i}}{1+e^{\hat{\mu}_i}} & \text{if } \hat{\theta}_i < \frac{1}{2}, \end{cases}$$

m_i might be considered to be more nonparametric than $\hat{M}_i$ since $\hat{M}_i$ is the MLE of the population median with observations from the $SMix\mathcal{M}$ distribution.

Parameters	Point Estimators
$\mu_{X_i} = \mathrm{E}[X_i]$	$X_i = \frac{1}{N}\sum_{r=1}^{N} X_{ri}$
$\eta_i = \sum_{b=1}^{B} \theta_{\mathbf{v}_b}\eta_i(\mathbf{v}_b)$ where $\eta_i(\mathbf{v}_b) = \begin{cases} 0 & \text{if } i \notin \mathbf{v}_b \\ \frac{e^{\mu_i^{\mathbf{v}_b}}}{1+e^{\mu_i^{\mathbf{v}_b}}} & \text{if } i \in \mathbf{v}_b \end{cases}$	$\hat{\eta}_i = \sum_{b=1}^{B} \hat{\theta}_{\mathbf{v}_b}\hat{\eta}_i(\mathbf{v}_b)$ where $\hat{\eta}_i(\mathbf{v}_b) = \begin{cases} 0 & \text{if } i \notin \mathbf{v}_b \\ \frac{e^{\hat{\mu}_i^{\mathbf{v}_b}}}{1+e^{\hat{\mu}_i^{\mathbf{v}_b}}} & \text{if } i \in \mathbf{v}_b \end{cases}$
$\lambda_i = \begin{cases} 0 & \text{if } \theta_i = 1 \\ (1-\theta_i)\frac{e^{\mu_i}}{1+e^{\mu_i}} & \text{if } \theta_i < 1. \end{cases}$	$\hat{\lambda}_i = \begin{cases} 0 & \text{if } \hat{\theta}_i = 1 \\ (1-\hat{\theta}_i)\frac{e^{\hat{\mu}_i}}{1+e^{\hat{\mu}_i}} & \text{if } \hat{\theta}_i < 1. \end{cases}$
$\lambda_i^s = \begin{cases} 0 & \text{if } \theta_i = 1 \\ (1-\theta_i)\frac{e^{\xi_i(\mu_i,\sigma_i^2,\alpha_i)}}{1+e^{\xi_i(\mu_i,\sigma_i^2,\alpha_i)}} & \text{if } \theta_i < 1. \end{cases}$	$\hat{\lambda}_i^s = \begin{cases} 0 & \text{if } \hat{\theta}_i = 1 \\ (1-\hat{\theta}_i)\frac{e^{\hat{\xi}_i(\hat{\mu}_i,\hat{\sigma}_i^2,\hat{\alpha}_i)}}{1+e^{\hat{\xi}_i(\hat{\mu}_i,\hat{\sigma}_i^2,\hat{\alpha}_i)}} & \text{if } \hat{\theta}_i < 1. \end{cases}$
$\mathrm{M}_i = \mathrm{median}[X_i]$ (Population Median)	$\mathrm{m}_i = \mathrm{median}[X_i]$ (Sample Median)

Table 2.1: MOLs and their point estimators

2.5.2 Interval Estimation

To assist with the discussions in Chapter 4, some insight into interval estimation for the simplified case where certain nuisance parameters are known will be the topic of the remainder of this section.

Before detailing the interval estimation methods, a few comments are needed. Firstly, depending on the underlying assumed distribution, confidence intervals (CIs) for certain MOLs will be more intuitive than for others. Accordingly, we will attempt to derive CIs for these MOLs. Also, individual CIs for the components (versus a confidence region (CR)) will be of most use as the components will eventually correspond to species and D (or I, in this case) will be large. For this reason a CR would be difficult to interpret and impractical. For some CI methods, however, the extension to a CR is fairly straightforward. To obtain simultaneous CIs, we will simply adjust α to α/D to yield Bonferroni type intervals. This correction is conservative and the overall coverage probability will be at least $1-\alpha$. Lastly, we have divided the discussion on CIs into finite and large sample interval estimation.

Finite Sample Intervals

If $\mathbf{X} \sim \mathcal{L}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ or if $\mathbf{X} \sim \mathcal{M}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then CRs for MOL^{AL} or MOL^{ML} follow almost immediately from standard multivariate theory. For example, the $100(1-\alpha)\%$ CR for $\boldsymbol{\mu}$ is given by the set

$$\left\{\boldsymbol{\mu} : n(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \leq \frac{d(N-1)}{N-d} F_{d,N-d}(1-\alpha)\right\}, \tag{2.18}$$

where $F_{d,N-d}$ denotes the F-distribution with d and $N-d$ degrees of freedom (df). Then if $\mathbf{X} \sim \mathcal{M}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, say, a $100(1-\alpha)\%$ CR for MOL^{ML} is the set

$$\left\{\text{MOL}^{ML} : \left[\log\left(\frac{\text{MOL}^{ML}[i]}{1 - \text{MOL}^{ML}[1] - \cdots - \text{MOL}^{ML}[i]}\right)\right] \in \text{CR for } \boldsymbol{\mu}\right\}. \tag{2.19}$$

(And similarly for the additive logistic transformation.) As remarked earlier, individual CIs for the components of the MOLs will be of more practical use. Although CIs for μ_i are apparent (that is, since on the transformed scale $Y_i \sim N(\mu_i, \sigma_i^2)$), it is only with the *multiplicative* logistic transformation that a CI on the composition scale may be easily obtained since, as previously mentioned, each component of MOL^{AL} contains at least two unknowns. A $100(1-\alpha)\%$ CI for μ_i and for $\text{MOL}^{ML}[1] = \frac{e^{\mu_i}}{1+e^{\mu_i}}$ are respectively

$$\hat{\mu}_i \pm t_{[(N-1),1-\alpha/2]} \frac{s_i}{\sqrt{N}}, \tag{2.20}$$

and

$$\left[\frac{e^{\hat{\mu}_i - t_{[(N-1),1-\alpha/2]}\frac{s_i}{\sqrt{N}}}}{1 + e^{\hat{\mu}_i - t_{[(N-1),1-\alpha/2]}\frac{s_i}{\sqrt{N}}}}, \frac{e^{\hat{\mu}_i + t_{[(N-1),1-\alpha/2]}\frac{s_i}{\sqrt{N}}}}{1 + e^{\hat{\mu}_i + t_{[(N-1),1-\alpha/2]}\frac{s_i}{\sqrt{N}}}}\right], \tag{2.21}$$

where t_{N-1} denotes the t distribution with $N-1$ df. Note that in this case, Equation 2.21 is also a CI for for the median M_i since $M_i = \text{MOL}^{ML}[1]$, as was shown in Section 2.4.

When zeros are present in the data set, inference procedures may use the $Mix\mathcal{M}$, $SMix\mathcal{M}$ or $SMix\mathcal{SM}$ distributions given in Equations 2.4, 2.6 and 2.7. Since CIs based on the $SMix\mathcal{M}$ distribution are the simplest to derive, these intervals will be considered first.

For observations modeled by the $SMix\mathcal{M}$ distribution, a CI for the MOL

$$\lambda_i = \begin{cases} 0 & \text{if } \theta_i = 1 \\ (1-\theta_i)\frac{e^{\mu_i}}{1+e^{\mu_i}} & \text{if } \theta_i < 1 \end{cases}$$

is most evident since a CI for $\frac{e^{\mu_i}}{1+e^{\mu_i}}$ is already given in Equation 2.21. Assuming that the nuisance parameter, θ_i, is known, clearly a CI is needed only if $\theta_i < 1$. It follows that if $\theta_i < 1$, a $100(1-\alpha)\%$ CI for λ_i is

$$\left[(1-\theta_i)\frac{e^{\hat{\mu}_i - t_{[(N_i'-1),1-\alpha/2]}\frac{s_i}{\sqrt{N_i'}}}}{1+e^{\hat{\mu}_i - t_{[(N_i'-1),1-\alpha/2]}\frac{s_i}{\sqrt{N_i'}}}}, (1-\theta_i)\frac{e^{\hat{\mu}_i + t_{[(N_i'-1),1-\alpha/2]}\frac{s_i}{\sqrt{N_i'}}}}{1+e^{\hat{\mu}_i + t_{[(N_i'-1),1-\alpha/2]}\frac{s_i}{\sqrt{N_i'}}}}\right]. \tag{2.22}$$

CIs computed as above may be problematic when N is small and, in particular, when θ_i is close to 1, as there may be no non-zero observations in the sample. For example, suppose $N = 10$ and $\theta_i = 0.9$ then $P[x_{ir} = 0,\ r = 1,\ldots,N] = (0.9)^{10} \approx 0.35$. In this case, in more than 1/3 of samples, a CI for λ_i is not possible. In Chapter 4, a bootstrap approach is introduced that may be used when N is small. This approach will also allow for the more realistic setting in which θ_i is unknown.

It should be noted that we will not attempt to derive an exact CI for observations modeled by the $SMixS\mathcal{M}$ distribution. We have, however, implemented a bootstrap algorithm (to be discussed in Chapter 4) that yields an approximate CI for λ_i^s based on this distribution.

When $X_{ir} \sim Mix\mathcal{M}^d(\mu_i^{\mathbf{v}_b}, \sigma_i^2)$, our MOL of interest will be $\eta_i = \sum_{b=1}^{B} \theta_{\mathbf{v}_b}\eta_i(\mathbf{v}_b)$ where recall that

$$\eta_i(\mathbf{v}_b) = \begin{cases} 0 & \text{if } i \notin \mathbf{v}_b \\ \frac{e^{\mu_i^{\mathbf{v}_b}}}{1+e^{\mu_i^{\mathbf{v}_b}}} & \text{if } i \in \mathbf{v}_b. \end{cases}$$

For group $\mathbf{v}_b$, if $i \in \mathbf{v}_b$ then Equation 2.21 essentially gives a CI for $\eta_i(\mathbf{v}_b)$. (If $i \notin \mathbf{v}_b$ then a CI is not needed as $\eta_i(\mathbf{v}_b) = 0$). The difficulty is then how to combine the separate CIs for $\eta_i(\mathbf{v}_b)$ from the various populations into a single CI for η_i. A conceivably easier task would be to obtain a P- value, say $p(\eta_{i0}(\mathbf{v}_b))$, for the test of

$$\begin{aligned} \mathrm{H}_0 &: \ \eta_i(\mathbf{v}_b) = \eta_{i0} \\ \mathrm{H}_1 &: \ \eta_i(\mathbf{v}_b) \neq \eta_{i0}, \end{aligned} \tag{2.23}$$

and then to pool the P-values (preferably using as weights $\theta_{\mathbf{v}_b}, b = 1,\ldots,B$) to obtain a single P-value, say $p(\eta_{i0})$, for the test of

$$\begin{aligned} \mathrm{H}_0 &: \ \eta_i = \eta_{i0} \\ \mathrm{H}_1 &: \ \eta_i \neq \eta_{i0}. \end{aligned} \tag{2.24}$$

The $100(1-\alpha)\%$ CI would then be the set

$$\{\eta_{i0} : p(\eta_{i0}) \geq \alpha\}.$$

We will now illustrate a method of computing the P-values under some idealized assumptions. Assume that $\theta_{\mathbf{v}_b}$, $b = 1, \ldots, B$ are known (and therefore the populations are known as well). First observe that

$$\begin{aligned} \eta_i = 0 &\Leftrightarrow \nexists b \text{ such that } i \in \mathbf{v}_b \\ &\Leftrightarrow P[X_i = 0] = 1. \end{aligned}$$

A P-value is therefore not needed for the test in Equation 2.24 when $\eta_{i0} = 0$ since $P[X_i = 0] = \sum_{\{\mathbf{v}: i \notin \mathbf{v}_b\}} \theta_{\mathbf{v}_b}$, and the $\theta_{\mathbf{v}_b}$'s are assumed to be known.

When $\eta_{i0} > 0$, a P-value for the test in Equation 2.23 may be obtained for each population in a relatively straightforward manner. For population $\mathbf{v}_b$, if $i \notin \mathbf{v}_b$, then $\eta_i(\mathbf{v}_b) = 0$ and a P-value is not needed. Otherwise, if $i \in \mathbf{v}_b$ then since

$$\eta_i(\mathbf{v}_b) = \frac{e^{\mu_i^{\mathbf{v}_b}}}{1 + e^{\mu_i^{\mathbf{v}_b}}},$$

we have

$$\eta_i(\mathbf{v}_b) = \eta_{i0} \Leftrightarrow \mu_i^{\mathbf{v}_b} = \log\left(\frac{\eta_{i0}}{1-\eta_{i0}}\right).$$

Our test statistic is then

$$T_{\mathbf{v}_b} = \left| \frac{\hat{\mu}_i^{\mathbf{v}_b} - \log\left(\frac{\eta_{i0}}{1-\eta_{i0}}\right)}{\frac{s_i^{\text{pool}}}{\sqrt{N_{\mathbf{v}_b}}}} \right|, \tag{2.25}$$

which is t-distributed with $\sum_{\{b: i \in \mathbf{v}_b\}}^{B} N_{\mathbf{v}_b} - \#\{b : i \in \mathbf{v}_b\}$ df under the null, and

$$p(\eta_{i0}(\mathbf{v}_b)) = 2P\left[T_{\mathbf{v}_b} > T_{\mathbf{v}_b,\text{obs}}\right].$$

An overall P-value for the test in Equation 2.24 is

$$p(\eta_{i0}) = \sum_{\{b: i \in \mathbf{v}_b\}} \theta_{\mathbf{v}_b} p(\eta_{i0}(\mathbf{v}_b)). \tag{2.26}$$

As with the intervals based on the simpler mixture density (Equation 2.6), there are some problems when this method is applied in practice. The most evident is

that there will likely not be observations from each of the populations. If, as is the case in practice, $\theta_{\mathbf{v}_b}, b = 1, \ldots, B$ were unknown and estimated by their MLEs, then $p(\eta_{i0}(\mathbf{v}_b))$ is only needed for b such that $\hat{\theta}_{\mathbf{v}_b} > 0$ and we would have observations in the necessary groups. A problem still exists when N is fairly small since $\hat{\theta}_{\mathbf{v}_b}$ will often be a very poor estimate of $\theta_{\mathbf{v}_b}$ (the populations themselves are also not well estimated in this case) and its variability should be taken into account in order to generate sensible intervals. An algorithm involving bootstrapping is discussed in Chapter 4 to deal with these difficulties.

It is important to observe that the P-value in Equation 2.26 is not uniformly distributed under the null hypothesis in Equation 2.24 as we would like. We have therefore considered an alternative P-value computed as follows

$$p(\eta_{i0}) = P\left[T > T_{\text{obs}}\right],$$

where

$$T = -2 \sum_{\{b: i \in \mathbf{v}_b\}} \log\left[p(\eta_{i0}(\mathbf{v}_b))\right]$$

and $T \dot{\sim} \chi^2$ distributed with $2 \times \#\{b : i \in \mathbf{v}_b\}$ df under the null hypotheses specified in Equation 2.23. This approximation assumes that the $\#\{b : i \in \mathbf{v}_b\}$ tests are independent and we assume this to be approximately the case. Note that this test statistic does not make use of the weights and further work is needed to incorporate them.

Large Sample Intervals

Observe that the estimators presented in Table 2.1 are all asymptotically normal by the following theorems (where $\rightarrow_d$ symbolizes "converges in distribution"):

Theorem 2.1 *Let $X_1, X_2, \ldots$ be a sequence of iid random variables with $\mathrm{E}[X_i] = \mu$ and $0 < \mathrm{Var}[X_i] = \sigma^2 < \infty$. Then*

$$\frac{\bar{X}_N - \mu}{\frac{\sigma}{\sqrt{N}}}$$

has a limiting standard normal distribution. (Casella and Berger, 1990)

Theorem 2.2 *Let $X_1, X_2, \ldots,$ be iid $f(x|\boldsymbol{\theta})$ and let $\hat{\boldsymbol{\theta}}$ denote the MLE of $\boldsymbol{\theta}$. Under the certain regularity conditions on $f(x|\boldsymbol{\theta})$,*

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \rightarrow_d N\left(0, \mathbf{I}(\boldsymbol{\theta})^{-1}\right),$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information. (Welsh, 1996)

Theorem 2.3 *Let $X_1, \ldots, X_N$ be a sample from a population with pdf f (assumed to be differentiable). Let m_N be the sample median and M the population median. Then*

$$\sqrt{N}(m_N - M) \rightarrow_d N\left(0, \frac{1}{[2f(M)]^2}\right).$$

(Casella and Berger, 2002)

Slutsky's Theorem ensures that the asymptotic normality in the above theorems also holds for consistent estimators of the variances.

Theorem 2.4 *(Slutsky's Theorem) If $X_n \rightarrow_d X$ and $Y_n \rightarrow_p a$, a constant, then*

1. $Y_n X_n \rightarrow_d aX$.
2. $X_n + Y_n \rightarrow_d X + a$.

(Casella and Berger, 1990.)

For example, since S_N^2 is consistent for σ^2, it is straightforward to show that $S_n \rightarrow_p \sigma$ and that $\frac{\sigma}{S_N} \rightarrow_p 1$. Slutsky's Theorem gives

$$\frac{\bar{X}_N - \mu}{\frac{S_N}{\sqrt{N}}} = \frac{\sigma}{S_N} \frac{\bar{X}_N - \mu}{\frac{\sigma}{\sqrt{N}}} \rightarrow_d N(0,1),$$

since, by Theorem 2.1, $\frac{\bar{X}_N - \mu}{\frac{\sigma}{\sqrt{N}}} \rightarrow_d N(0,1)$.

Note that Slutsky's Theorem may similarly be applied to Theorem 2.2 since $\mathbf{I}(\hat{\boldsymbol{\theta}})$ is a consistent estimator of $\mathbf{I}(\boldsymbol{\theta})$. We have the following large sample $100(1-\alpha)\%$ CIs for μ_{X_i} and η_i,

$$\bar{X}_i \pm z_{1-\alpha/2} \frac{S_{X_i}}{\sqrt{N}} \tag{2.27}$$

$$\hat{\eta}_i \pm z_{1-\alpha/2}\sqrt{\frac{I_{\hat{\eta}_i}(\hat{\theta}_{\nu_b}, \hat{\mu}_i^{\nu_b}, \hat{\sigma}_i^{2\nu_b})}{N}} \tag{2.28}$$

and similarly for λ_k and λ_k^s.

Application of the asymptotic results to the diet estimation problem is discussed in Section 3.5. Realize that in the diet estimation application, the large sample intervals require both the number of predators (n_s) and the number of prey (n_k) to be large. Additionally, although the asymptotic variance in Equation 2.28 could be derived using the log likelihood functions in Equations 2.12 and 2.15 respectively, Equation 2.27 is much simpler and should adequately estimate the true diet of the predator when n_s and n_k are large. Similarly, while there are various ways of estimating the asymptotic variance of m_N in Theorem 2.3, including a kernel density estimate, we will again prefer the simpler large sample result in Equation 2.27.

Chapter 3

Diet Point Estimation

Having examined general methods of compositional data analysis in Chapter 2, we now consider our particular application of interest, namely the diet estimation problem. In this chapter we detail several methods of obtaining quantitative estimates of the diet of a predator by matching its FA signature to its prey signatures. We also consider estimating the diet of a group of predators. In the latter case, we require a sample of predator FA signatures from the group of predators of interest and a suitable aggregate point estimator of diet. Appropriate point estimators are discussed as well as some of their properties when the sample size of predator and prey FA signatures is large. Additionally, sources of variability in the diet estimates, the bias in our estimation procedure and possible methods of modeling the estimates are discussed.

Although the discussions in this chapter may be applied to various types of predators, we will assume that our predator of interest is seals to illustrate our procedures.

3.1 Notation

At the root of the diet estimation problem there are essentially two populations, namely the set of all FA signatures from seals from a specified region, and the set of all prey FA signatures from the I prey types known to be part of the seals' diet. To set the notation, let Y_{ij} denote the jth FA of the ith seal and $\mathbf{Y}$ the FA signature of a single seal. When referring to the prey, let X_{klj} denote the jth FA of the lth prey from the kth prey type, $\mathbf{X}_{kl}$ the lth FA signature from the kth prey type, and $\mathbf{X}_k$ the $n_k \times n_{FA}$ matrix of FA signatures for prey type k. Usually prey type will correspond to species but could more generally represent any appropriate grouping of the prey. In practice, a sample of size n_s of FA signatures of the seals, $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_s}$, $n_s \geq 1$, and a sample of FA signatures from the various prey types, $\mathbf{X}_1 \ldots \mathbf{X}_I$, are provided. We will usually assume that the n_s FA signatures are independent as should be the

case if they correspond to different seals.

Let $f_{\mathbf{Y}}$ denote the distribution of the seal FA signatures and let $f_{\mathbf{X}_k}$ be the distribution of the prey FA signatures from prey type k. Assuming that the prey FA signatures between species are independent, $f_{\mathbf{X}_1,\ldots,\mathbf{X}_I}(\mathbf{x}_1,\ldots,\mathbf{x}_I) = f(\mathbf{x}_1)\cdots f(\mathbf{x}_I)$. Let $\boldsymbol{\pi}_0$ denote the true diet of a seal where the kth component, π_0^k, gives the true long-run proportion of the kth prey type in the seal's diet. Let $\boldsymbol{\pi}$ denote the average diet of the group of seals in the region of interest. In some situations, it may be reasonable to assume that the seals in the region of interest have, at least approximately, a common diet. In this case, we let $\boldsymbol{\pi} = \boldsymbol{\pi}_0$ for all seals in the region. (This issue is discussed in more detail in Section 3.3.)

As will be addressed in Section 3.2, some of the point estimation algorithms require that the prey FA signatures be summarized. For prey type k, let $\mathbf{Q}_k$ be a statistic representing the FA signatures $\mathbf{X}_k$ and $\boldsymbol{\psi}_k$, the population version. As a simple example, $\mathbf{Q}_k$ may be taken to be $\bar{\mathbf{X}}_k$, the sample mean vector for species k, so that $\boldsymbol{\psi}_k = \mathrm{E}[\mathbf{X}_{kl}]$. (Note that we will often drop the "l" and simply write $\mathrm{E}[\mathbf{X}_{kl}]$ as $\mathrm{E}[\mathbf{X}_k]$.) It will also be possible for $\mathbf{Q}_k$ to represent a selected sample quantile of $\mathbf{X}_k$ (and $\boldsymbol{\psi}_k$ the corresponding population quantile). Two methods for determining appropriate sample quantiles of the $\mathbf{X}_k$, $k = 1,\ldots,I$ are given in Section 3.2. Any further notation will be defined as needed in the sections that follow.

3.2 Diet Estimation Method

In this section we present our primary diet estimation method which we have named the distance minimization (DM) algorithm. While other methods of estimation are discussed in Section 3.6, we have found these methods to be problematic and to yield less accurate estimates of the diet. We will consider a few variations of the DM algorithm including the QFASA method developed in Iverson *et al* (2004). As will be apparent, the DM algorithm requires a single seal FA signature only, so that a different estimate of diet is possible for each seal signature in the sample. Overall point estimators of diet are given in Section 3.4 and are based on the discussions in Section 3.3.

In the distance minimization (DM) algorithm, the estimator of $\boldsymbol{\pi}_0$ is the set of

weights, $p_1, \ldots, p_I$, that minimize the distance between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$, for some suitable distance measure, where

$$\hat{\mathbf{Y}} = \sum_{k=1}^{I} p_k \mathbf{Q}_k.$$

More specifically, define

$$\mathbf{p}(\mathbf{Y}, \mathbf{Q}) = \mathbf{p}(\mathbf{Y}, \mathbf{Q}_1, \ldots, \mathbf{Q}_I) = \arg\min_{p_1 \ldots, p_I} \text{dist}(\mathbf{Y}, \hat{\mathbf{Y}}),$$

to be the DM estimate of diet for seal with FA signature $\mathbf{Y}$ when the kth species is represented by $\mathbf{Q}_k$ and where $\mathbf{Q}$ denotes $\mathbf{Q}_1, \ldots, \mathbf{Q}_I$.

Two distance measures will be considered, both of which take into account the compositional nature of $\mathbf{Y}$ and $\hat{\mathbf{Y}}$. One will be the symmetric Kulback-Leibler (KL) distance measure which was investigated in Iverson *et al* (2004), along with several other distance measures, and found to be the preferred choice. The KL distance between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ is defined as

$$\text{KL}(\mathbf{Y}, \hat{\mathbf{Y}}) = \sum_{j=1}^{n_{FA}} (Y_j - \hat{Y}_j) \log(Y_j / \hat{Y}_j). \tag{3.1}$$

Note that the KL distance, as defined in Equation 3.1, is actually an average of the forward and backward KL distances (that is, $\sum_{j=1}^{n_{FA}} Y_j \log(Y_j/\hat{Y}_j)$ and $\sum_{j=1}^{n_{FA}} \hat{Y}_j \log(\hat{Y}_j/Y_j)$) defined in DiCiccio and Romano (1990).

An alternative compositional distance measure not considered in Iverson *et al* (2004), but recommended in Aitchison (1992) and discussed further in Aitchison (2000), will also prove to be useful. Aitchison's distance between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ is defined as

$$\text{AIT}(\mathbf{Y}, \hat{\mathbf{Y}}) = \sum_{j=1}^{n_{FA}} \{\log[Y_j / g(\mathbf{Y})] - \log[\hat{Y}_j / g(\hat{\mathbf{Y}})]\}^2, \tag{3.2}$$

where $g(\mathbf{Y}) = (Y_1 \cdots Y_{n_{FA}})^{\frac{1}{n_{FA}}}$ represents the geometric mean.

What follows is a discussion of various choices for $\mathbf{Q}_k$, $k = 1, \ldots, I$.

1. *Current (MEAN) Method*

   In Iverson *et al* (2004), the diet of a seal with FA signature $\mathbf{Y}$ is estimated by the set of p_k that minimize $\text{KL}(\mathbf{Y}, \hat{\mathbf{Y}})$ where

   $$\hat{\mathbf{Y}} = \sum_{k=1}^{I} p_k \bar{\mathbf{X}}_k, \tag{3.3}$$

or, in terms of the above defined notation, by $\mathbf{p}(\mathbf{Y}, \bar{\mathbf{X}})$. This method will be referred to as the MEAN method because $\mathbf{Q}_k = \bar{\mathbf{X}}_k$. (Note that we could also use the AIT distance.)

The sample of prey signatures may vary substantially within a species, particularly if the sample contains prey from various regions. To capture this variability, Iverson *et al* (2004) carried out a nonparametric bootstrap of the prey, the details of which are given in Appendix C.

An estimate of the standard error of p_k is then

$$se(p_k) = \sqrt{\frac{\sum_{r=1}^{R}(p_k^{*r} - \bar{p}_k^{*r})^2}{R-1}}$$

where $\bar{p}_k^{*r} = \frac{1}{R}\sum_{r=1}^{R} p_k^{*r}$ and the p_k^{*r} are the simulated estimates.

In an analogous manner, a *parametric* bootstrap procedure is possible if it is assumed that the rows of $\mathbf{X}_k$ form a random sample from one of the parametric distributions discussed in Chapter 2 such as the $\mathcal{L}(\mu_k, \Sigma)$ distribution. The algorithm is also given in Appendix C. (Note the assumed common covariance matrix, Σ.) We will generally prefer the nonparametric bootstrap procedure since n_{FA} will be fairly large making it difficult to assess whether the $\mathcal{L}(\mu_k, \Sigma)$ distribution is an appropriate distributional assumption.

It is important to realize that these bootstrap procedures yield estimates of the variability of the (MEAN method) diet estimates due to the variability in the prey FA signatures only and are conditional on $\mathbf{Y}$. When the variability due to the seal FA signatures is taken into account, the situation is more complex and requires a more detailed discussion. This discussion is given Section 3.3.

2. *Random Sampling Method*

In addition to producing estimates of π_0, the random sampling (RS) method provides an alternative to using bootstrap procedures to gain insight into the variability due to the prey.

In the RS method, $\bar{\mathbf{X}}_k$ in Equation 3.3 is replaced by a randomly selected FA signature from prey type k. (That is, $\mathbf{Q}_k$ is simply a randomly selected FA

signature from prey type k.) If $M^{(\max)} = n_1 n_2 \cdots n_I$ is the total possible number of distinct random selections, then using the RS method, $M^{(\max)}$ estimates could be generated. For computational purposes, due to the large number of possible combinations, usually a smaller number of estimates, say M estimates, are computed. The RS method generates $\mathbf{p}^{(RS)}(\mathbf{Y}, \mathbf{Q}^{(m)})$, $m = 1 \ldots M$.

3. *Multivariate Quantile Method*

Because in practice the number of distinct samples will be extremely large, it seemed useful to first choose n_{quant} representative prey FA signatures from each prey type and then to sample from this smaller prey base. Note that in the MEAN method, $n_{\text{quant}} = 1$ and in the RS method, $n_{\text{quant}} = n_k$ for the kth species.

In the Multivariate Quantile (MQ) method, the representative prey are chosen to be specified multivariate quantiles of the prey data. Multivariate quantiles are defined in Chaudhuri (1996) as an extension to one dimensional quantiles. That is, for a sample $x_1, \ldots, x_n$ in one dimension, the αth quantile q, $0 < \alpha < 1$, is given by

$$\arg\min_{q \in \mathcal{R}} \sum_{i=1}^{n} \{|x_i - q| + u(x_i - q)\},$$

where $u = 2\alpha - 1$. In d-dimensions, Chaudhuri indexes multivariate quantiles by elements of the open unit ball $\{\mathbf{u} | \mathbf{u} \in \mathcal{R}^d, \ ||\mathbf{u}|| < 1\}$. He then defines the geometric quantile $\mathbf{q}$, indexed by $\mathbf{u}$, for a multivariate sample, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathcal{R}^d$ as

$$\arg\min_{\mathbf{q} \in \mathcal{R}^d} \sum_{i=1}^{n} \{||\mathbf{x}_i - \mathbf{q}|| + \mathbf{u}^{\mathrm{T}}(\mathbf{x}_i - \mathbf{q})\}. \tag{3.4}$$

If the l_p norm ($1 \leq p < \infty$) is used in (3.4), the geometric quantile will be referred to as the l_p-quantile and $||\mathbf{x}_i - \mathbf{q}||_p = (\sum_{j=1}^{n} |x_{ij} - q_j|^p)^{1/p}$.

Chakraborty (2002) noted that the multivariate quantiles determined by Equation 3.4 are not affine equivariant and therefore proposed an affine equivariant modification to the l_p-quantile. Informally, his procedure consisted of forming a data-driven coordinate system based on $(d + 1)$ of the observations. Then letting α contain the $d+1$ indices corresponding to the chosen observations and

$\mathbf{X}(\alpha)$ the transformation matrix to the new coordinate system, each data point $\mathbf{X}_j$ is transformed to $\mathbf{Y_j} = [\mathbf{X}(\alpha)^{-1}]\mathbf{X}_j$. Let $\hat{\mathbf{R}}$ be the $\mathbf{v}(\alpha)th$ l_p-quantile based on the $\mathbf{Y}_j$'s, where

$$\begin{aligned} \mathbf{v}(\alpha) &= \frac{[\mathbf{X}(\alpha)]^{-1}\mathbf{u}}{||[\mathbf{X}(\alpha)]^{-1}\mathbf{u}||_q}||\mathbf{u}||_q \quad \text{for } \mathbf{u} \neq 0 \\ &= 0 \qquad\qquad\qquad\qquad\quad \text{for } \mathbf{u} = 0. \end{aligned}$$

Chakraborty then defined the multivariate transformation retransformation (TR) l_p-quantile for the original data by $\hat{\mathbf{Q}}^{(\alpha,p)}(\mathbf{u}) = [\mathbf{X}(\alpha)]\hat{\mathbf{R}}$.

To apply the above definitions and algorithm to the problem of finding n_{quant} multivariate quantiles of the prey, consider first the quantile contours described by the sets $\{\hat{\mathbf{Q}}^{(\alpha,p)}(\mathbf{u}) : ||\mathbf{u}||_q = r\}$ where $0 < r < 1$, $1/p+1/q = 1$. Chakraborty (2002) stated that for a certain optimal selection of $\boldsymbol{X}(\alpha)$, the *population* quantile contours for $p = 2$ $(q = 2)$ correspond to the probability density contours if the underlying density is elliptically symmetric with density of the form $[\det(\Sigma)]^{-1/2} f((\mathbf{x} - \boldsymbol{\theta})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\theta}))$. Note that if the prey FA signatures were assumed to be $\mathcal{L}^d$ distributed, then transformed using the additive logratio transformation, they are multivariate normally distributed and hence satisfy the above criteria. It would then seem appropriate to consider indexing the multivariate quantiles of the transformed data by vectors whose magnitudes correspond to typical univariate quantiles of interest. Matlab code was provided by Chakraborty (personal communication) to compute TR l_2-quantiles, indexed by a vector $\mathbf{u}$, of a multivariate data set. Some difficulties were encountered in computing the TR l_2-quantiles as the algorithm requires $n_k > (n_{FA}+1)$ and this is not always the case in the prey data. Further, for the prey types for which this inequality was satisfied, the computational time was significant. For these reasons, the l_2-quantiles (which were computable from the Matlab code) were used instead.

In summary, the MQ method used to generate n_{quant} multivariate quantiles for each species is as follows:

(a) Transform $\mathbf{X}_1 \ldots \mathbf{X}_I$ to $\mathbf{Z}_1 \ldots \mathbf{Z}_I$ using the additive logratio transformation.

(b) Choose a set of n_{quant} quantiles, say $S = \{s_1, \ldots, s_{n_{\text{quant}}}\}$. (For example, $S = \{0.25, 0.5, 0.75\}$.)

(c) For each element, s_i, in S compute $\mathbf{u}_{s_i}$ by choosing its components from a U[0,1] distribution and then normalizing the vector so that $||\mathbf{u}_{s_i}|| = s_i$.

i. *for* $k = 1, \ldots, I$

A. Compute the l_2-quantiles of $\mathbf{Z}_k$, indexed by $\mathbf{u}_{s_i}$, $i = 1, \ldots, n_{\text{quant}}$. Note that if $n_k \leq 6$, the l_2-quantiles could not be computed and the first n_{quant} observations are used instead.

B. Transform the n_{quant} l_2-quantiles using the additive logistic transformation and let $\mathcal{Q}_k$ contain the transformed quantiles.

By randomly selecting from the rows of $\mathcal{Q}_k$, we may generate M MQ method estimates, $\mathbf{p}^{(MQ)}(\mathbf{Y}, \mathbf{Q}^{(m)})$, $m = 1 \ldots M$.

4. *KL Quantile Method*

The interpretation of a multivariate quantile in several dimensions is not straightforward and for this reason the KL Quantile (KLQ) method was developed. Essentially the KLQ method first reduces each prey FA signature, $\boldsymbol{X}_{kl}$, to a univariate quantity, R_{kl}, computes the quantiles of the R_{kl} and then relates those quantiles back to the corresponding prey FA signatures.

The algorithm is as follows:

(a) Choose a set of n_{quant} quantiles, say $S = \{s_1, \ldots, s_{n_{\text{quant}}}\}$.

i. *for* $k = 1, \ldots, I$

A. Calculate

$$R_{kl} = \text{KL}(\mathbf{Y}, \mathbf{X}_{kl}), l = 1, \ldots, n_k.$$

B. Compute the (univariate) quantiles specified in S of the R_{kl}.

C. Determine the FA signature of the prey associated with each quantile. Let $\mathcal{Q}_k$ contain the matrix of these FA signatures.

The KLQ estimates, $\mathbf{p}^{(KLQ)}(\mathbf{Y}, \mathbf{Q}^{(m)}), m = 1, \ldots, M$ are then generated by randomly selecting FA signatures from the rows of $\mathcal{Q}_k$.

It should be noted that unlike the MEAN, RS and MQ methods, the representative prey from each prey type chosen by the KLQ method depends on the FA signature of the seal. For a sample of seals, one seal is selected at random to be used in Step A. Realize also that we could have chosen to use the AIT distance measure in Step A. When the AIT distance measure is used, we will refer to this method as the AITQ method.

3.3 Parameterization of the Diet

Section 3.2 described various methods of estimating the diet of a single seal given its FA signature and a representative prey ($\mathbf{Q}$) from each species. For a seal with true diet π_0, its FA signature, and therefore the distance minimization algorithm (DMA) diet estimate, will vary over time. While we might expect that over time, the average DMA estimate of diet would equal π_0, we show in this section using pseudo-seals that this is not the case. If we generate many pseudo-seals with diet π_0 and estimate the diet of each pseudo-seal using the DMA, the average of the diet estimates is not π_0. We call this difference between the average of the diet estimates and π_0 the bias in our estimation procedure. Possible factors contributing to this bias include the similarity between the signatures of certain species, the choice of FA subsets and the unknown calibration factors. Additionally, nonlinear estimation techniques are used. While in Chapter 4 we attempt to estimate this bias, in this section we are interested in determining which DMA estimate we should use (that is, the choice of ψ, the population version of $\mathbf{Q}$) and how to average the diet estimates that would arise over time so that the chosen measure of location (MOL) is close to the true diet. It is then this parameter that we attempt to estimate. We call this process the "parameterization of the diet".

While we will initially discuss parameterization of the diet of a single seal we will also consider parameterizing π, the true average diet of a group of seals. We will argue that the parameterization of π follows once we have found a MOL that is close to π_0.

Understanding the sources of variability in the seal FA signatures (which is related to the variability in the prey signatures) is crucial in developing useful DMA based

parameters. For a single seal, while the diet may be assumed to remain roughly constant over time, the FA signature of the seal will vary with location and time. A major source of this variability is due to the prey signatures also varying with location and time. That is, in a large region of the ocean we can imagine there to be various clusters of prey with similar FA signatures and the seal travelling from cluster to cluster over time. Its FA signature at a given point in time (such as when the seal FA signature is recorded), will reflect mostly the prey signatures in the cluster from which it is eating. A second source of variability in the seal's FA signature is due to the prey signatures varying within a cluster so that even if the seal remained in a given region of the ocean, over time its FA signature would change. For a group of seals in an area of interest, there is also the variability due to the seals having slightly different diets. In summary, the variability in the FA signatures of seals in a given region of the ocean can be primarily explained by the following three elements:

1. The seals may not have identical diets.

2. The variability between clusters in the prey signatures.

3. The variability within clusters in the prey signatures.

For each π_0, we assume that there is a corresponding population of FA signatures. Note that if two seals each have true diet π_0 then we are assuming that the population of possible FA signatures are the same for both seals. For a seal with true diet π_0, depending on the order in which the summarizing in the prey signatures is carried out, there are essèntially two reasonable ways of parameterizing its diet:

1. $\mathrm{MOL}_{\mathbf{Y}}[\mathbf{p}(\mathbf{Y}, \mathrm{MOL}_{\mathbf{X}}[\mathbf{X}])] \equiv \mathrm{MOL}_{\mathbf{Y}}[\mathbf{p}(\mathbf{Y}, \mathrm{MOL}_{\mathbf{X}_1}[\mathbf{X}_1], \ldots, \mathrm{MOL}_{\mathbf{X}_I}[\mathbf{X}_I])]$, or

2. $\mathrm{MOL}_{\mathbf{Y},\mathbf{X}_1,\ldots,\mathbf{X}_I}[\mathbf{p}(\mathbf{Y}, \mathbf{X})] \equiv \mathrm{MOL}_{\mathbf{Y},\mathbf{X}_1,\ldots,\mathbf{X}_I}[\mathbf{p}(\mathbf{Y}, \mathbf{X}_1, \ldots, \mathbf{X}_I)]$,

where MOL[·] is an appropriate MOL computed using the distribution of the FA signatures that arise when the true diet is π_0.

The argument for which of these two parameters is better is slightly arbitrary. The first parameter, however, will be preferred for the reason that the seal would eat many fish from each species so that an estimate based on average prey FA signatures should

conceptually be more accurate than the average of many estimates based on individual prey FA signatures. For simplicity, we will choose $\text{MOL}[\mathbf{X}_k] = \text{E}[\mathbf{X}_k] = \boldsymbol{\mu}_{\mathbf{X}_k}$. Then our preferred DMA based estimate of diet for a single seal is

$$\mathbf{p}(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}}) = \mathbf{p}(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}_1}, \ldots, \boldsymbol{\mu}_{\mathbf{X}_I}).$$

Accordingly, $\boldsymbol{\pi}_0$ will be parameterized by

$$\text{MOL}_{\mathbf{Y}}[\mathbf{p}(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})],$$

where the expectations are taken over all possible prey signatures and the MOL over the distribution of FA signatures corresponding to $\boldsymbol{\pi}_0$. Note that if we knew the actual clusters of prey signatures then we could instead use

$$\mathbf{p}(\mathbf{Y}, \text{MOL}_{\mathbf{X}^{\text{clust}(1)}}[\mathbf{X}], \text{MOL}_{\mathbf{X}^{\text{clust}(2)}}[\mathbf{X}], \ldots)$$

to estimate the diet.

It should be mentioned that an alternative approach which we have not considered would have been to parameterize the diet as $\mathbf{p}(\boldsymbol{\mu}_{\mathbf{Y}}, \boldsymbol{\mu}_{\mathbf{X}})$ where $\boldsymbol{\mu}_{\mathbf{Y}}$ is the average over all FA signatures associated with $\boldsymbol{\pi}_0$. We prefer our approach as the biologists may want to examine the diet estimates of the individual seals as well as an aggregate of the diet estimates. Furthermore, in comparing the diet of two or more groups of seals (as is the case in Chapter 6) we are able to obtain a better sense of the variability in the diet estimates across seal FA signatures if we estimate the diet of each seal, in each group, rather than obtain one average diet estimate for each group.

For a group of seals with possibly different diets, $\boldsymbol{\pi}_0$ will vary from seal to seal and we can consider $\boldsymbol{\pi}_0$ to be a random variable with a corresponding distribution. We would then be interested $\boldsymbol{\pi} = \text{MOL}[\boldsymbol{\pi}_0]$, the average diet of the seals, and therefore $\boldsymbol{\pi}$ will be parameterized by $\text{MOL}_{\boldsymbol{\pi}_0}[\text{MOL}_{\mathbf{Y}}[\mathbf{p}(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})]|\boldsymbol{\pi}_0]$. We will, however, make the assumption that all of the seals in the region of interest have, at least approximately, the same diet so that $\boldsymbol{\pi} \approx \boldsymbol{\pi}_0$ and our parameter of interest is $\text{MOL}_{\mathbf{Y}}[\mathbf{p}(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})]$, where the MOL is taken over all FA signatures that may arise when the true diet is $\boldsymbol{\pi}_0$. In practice we must estimate $\text{MOL}_{\mathbf{Y}}[\mathbf{p}(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})]$. Point estimation is discussed in Section 3.4 while interval estimation is discussed in Chapter 4.

To assess the above parameterization of π_0 or π and to determine an appropriate choice of a MOL, we will use the pseudo-seals described in Chapter 1 and in Appendix B. We will generate many pseudo-seals with a given diet and examine our parameterization for various choices of MOLs. While this will give us insight into the distribution of FA signatures that arise from a specified diet, pseudo-seals generated in this manner will be less variable than seals in the wild if our assumption of a common diet is not valid. In attempting to parameterize π_0, this issue is irrelevant but could matter in Chapter 4 where CIs for π will be constructed if it is believed that the diets of seals in the wild, in a given region, differ considerably. This issue is further discussed in Section 4.5.

Another difference in the variability, also discussed in Section 4.5, occurs because the pseudo-seals are essentially generated as if the seals are random sampling from the prey signatures instead of cluster sampling, as previously discussed. Since the diet estimates, $\mathbf{p}(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})$, depend on the overall mean, $\boldsymbol{\mu}_{\mathbf{X}}$ (and not the cluster means), the estimates based on the pseudo-seals might be less variable than the estimates computed using seals that were cluster sampling, as in the wild. In Chapter 4 this difference in variability may be important but in determining a parameterization of the diet, the distinction is less crucial. Note that it is in dealing with this difference in variability that the other DM methods (that is, the RS, MQ and KLQ methods) will be useful.

Recall from Section 2.4 the various MOLs that were reasoned to be appropriate for compositional data (with or without zeros). Since the diet estimates, $\mathbf{p}(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})$, are compositions (with some zeros likely to be present), the MOLs of Section 2.4 are possible parameters of interest. Using the pseudo-seals, the extent to which one or more of these MOLs are close to the true diet will now be investigated more concretely. These MOLs, applied to the diet estimates, are as follows

- $\mu_{p_k} = \mathrm{E}_{\mathbf{Y}}[p_k(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})]$
- $\eta_k = \sum_{b=1}^{B} \theta_{\mathbf{v}_b} \eta_k(\mathbf{v}_b)$ where $\eta_k(\mathbf{v}_b) = \begin{cases} 0 & \text{if } k \notin \mathbf{v}_b \\ \frac{e^{\mu_k^{\mathbf{v}_b}}}{1+e^{\mu_k^{\mathbf{v}_b}}} & \text{if } k \in \mathbf{v}_b \end{cases}$

  and $\mu_k^{\mathbf{v}_b} = \mathrm{E}_{\mathbf{Y}}\left[\log\left(\frac{p_k^{\mathbf{v}_b}(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})}{1-p_k^{\mathbf{v}_b}(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})}\right)\right]$

- $\lambda_k = \begin{cases} 0 & \text{if } \theta_k = 1 \\ (1-\theta_k)\frac{e^{\mu_k}}{1+e^{\mu_k}} & \text{if } \theta_k < 1 \end{cases}$ where $\mu_k = \mathrm{E}_{\mathbf{Y}}\left[\log\left(\frac{p'_k(\mathbf{Y},\mu_{\mathbf{X}})}{1-p'_k(\mathbf{Y},\mu_{\mathbf{X}})}\right)\right]$ and p'_k denotes the non-zero diet estimates.
- $\lambda_k^s = \begin{cases} 0 & \text{if } \theta_k = 1 \\ (1-\theta_k)\frac{e^{\xi_k(\mu_k,\sigma_k^2,\alpha_k)}}{1+e^{\xi_k(\mu_k,\sigma_k^2,\alpha_k)}} & \text{if } \theta_k < 1 \end{cases}$
  where $\xi_k(\mu_k,\sigma_k^2,\alpha_k) = \sigma_k\left(\frac{2}{\pi}\right)^{\frac{1}{2}}\frac{\alpha_k}{(1+\alpha_k^2)^{\frac{1}{2}}} + \mu_k$ and
  $\log\left(\frac{p'_k(\mathbf{Y},\mu_{\mathbf{X}})}{1-p'_k(\mathbf{Y},\mu_{\mathbf{X}})}\right) \sim \mathcal{SN}(\mu_k,\sigma_k^2,\alpha_k)$
- $\mathrm{M}_{p_k} = \mathrm{median}_{\mathbf{Y}}[p_k(\mathbf{Y},\mu_{\mathbf{X}})]$

We have also considered the parameter $\mathrm{E}_{\mathbf{Y},\mathbf{X}}[p_k(\mathbf{Y},\mathbf{X})]$, though we do not expect this parameter to be an adequate parameterization as previously discussed. Note that all of the parameters will be normalized so that their components sum to one.

An obvious difficulty in comparing and evaluating these parameters is that we have only a sample of prey signatures. We will therefore examine the closeness of the parameters to the true diet by defining our population of the prey signatures to be the sample, $\mathbf{X}_1,\ldots,\mathbf{X}_I$.

To compute the MOLs given above the following algorithm was carried out:

1. *for* $r = 1:1000$
   (a) Generate a pseudo-seal, $\mathbf{Y}^r$, with diet π, and without splitting $\mathbf{X}_1,\ldots,\mathbf{X}_I$. (See Appendix B.)
   (b) Compute $\mu_{\mathbf{X}_k} \equiv \bar{\mathbf{X}}_k$, $k = 1,\ldots,I$.
   (c) Compute the diet estimate for the rth pseudo-seal: $\mathbf{p}^r(\mathbf{Y}^r,\mu_{\mathbf{X}})$.
2. Compute the MOLs using $\mathbf{p}^r(\mathbf{Y}^r,\mu_{\mathbf{X}})$, $r = 1,\ldots,1000$.

Note that we applied the ML estimation functions in the *sn* library to $p_k^r(\mathbf{Y}^r,\mu_{\mathbf{X}})$ to compute λ_k^s. To compute $\mathrm{E}_{\mathbf{Y},\mathbf{X}}[p_k(\mathbf{Y},\mathbf{X})]$, $\mu_{\mathbf{X}_k}$ in the above algorithm is replaced by a randomly selected prey signature from $\mathbf{X}_k$. Also, we have chosen π in (a) to be Diet 1 and Diet 4 from Iverson *et al* (2004). Recall from Section 1.1 that Diet 1 is considered to be a difficult diet to estimate while Diet 4 should be similar to

the true diet of seals in the region of interest. Realize that because we are using 10% noise, we actually expect the "zero species" in the diet to contribute (together) 10% of the diet. In Diet 1, for example, while we specify that the true proportion of Plaice, Sandlance, WinterFlounder and YellowTail in the diet is zero, in actuality, it is roughly $0.10/4 = 0.025$ for each of these species. We therefore expect a slight positive bias for these species. For Diet 4, there are only two zero species, Haddock and Plaice, and we thus expect a bias of about 0.05 as we have specified their true contribution to be zero. If the prey base is large, as is the usual case in practice, this source of bias for zero species should be negligible for pseudo-seals generated with 10% noise.

The results are presented in Tables 3.1 and 3.2 and in Figures 3.1 and 3.2. The tables give the actual difference between the MOL and π: MOL - π. We will call this difference the "bias" in the MOL though it is only for μ_{p_k} and for $\mathrm{E}_{\mathbf{Y},\mathbf{X}}[p_k(\mathbf{Y},\mathbf{X})]$ that this difference truly represents the statistical definition of bias. The figures also illustrate the bias in the parameters and may be used to more easily compare the performance of the various MOLs and the distance measures. Note that missing data on the figures (for example in Figure 3.2 there are no MOLs based on AIT distance) correspond to bias results that are larger than 0.10 in magnitude. Since $\sqrt{\frac{s_{p_k}}{1000}} \times 1.96 \approx 0.01$ for all species we expect μ_{p_k}, for example, to be within ± 0.01 (with 95% confidence) of the results shown.

As was surmised earlier, the results show that, in general, $\mathrm{E}_{\mathbf{Y},\mathbf{X}}[\mathbf{p}(\mathbf{Y},\mathbf{X})]$ is not a good parameterization. The performance of the remaining MOLs appears to depend on π_k and on the distance measure used. Surprisingly the distance measures often give very different results and even produce opposite signs in the bias for some species. The effect of the distance measures is most obvious when the true diet is fairly large, with Sandlance in Diet 1 and Haddock and Pollock in Diet 4 being exceptions. Pollock in Diet 4 appears to be particularly troublesome and is overestimated by all of the MOLs. Even if we considered its true contribution to be approximately 0.05 (due to the 10% noise that was used), there would still be a large bias. Note that Sandlance (Diet 4) is almost always underestimated suggesting that the algorithm is having trouble distinguishing between the FA signatures of Sandlance and of Pollock. To

	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
	π	0.3	0.3	0	0.15	0	0.15	0	0
$E_{\mathbf{Y},\mathbf{X}}$	AIT	-0.100	-0.134	0.080	0.054	0.082	-0.001	0.057	0.062
$[p_k(\mathbf{Y},\mathbf{X})]$	KL	-0.105	-0.115	0.087	0.004	0.087	0.014	0.058	0.069
$E_{\mathbf{Y}}$	AIT	-0.017	-0.055	0.023	0.074	0.045	-0.016	0.017	0.029
$[p_k(\mathbf{Y},\mu_{\mathbf{X}})]$	KL	-0.082	0.051	0.010	-0.014	0.085	-0.009	0.020	0.040
$\frac{\eta_k}{\sum_{i=1}^{I}\eta_i}$	AIT	-0.003	-0.053	0.018	0.078	0.043	-0.023	0.013	0.026
	KL	-0.088	0.069	0.008	-0.009	0.078	-0.010	0.018	0.034
$\frac{\lambda_k}{\sum_{i=1}^{I}\lambda_i}$	AIT	0.007	-0.039	0.018	0.078	0.029	-0.029	0.014	0.022
	KL	-0.075	0.089	0.008	-0.024	0.070	-0.018	0.018	0.033
$\frac{\lambda_k^s}{\sum_{i=1}^{I}\lambda_i^s}$	AIT	0.001	-0.037	0.018	0.082	0.032	-0.031	0.014	0.022
	KL	-0.077	0.093	0.008	-0.033	0.072	-0.012	0.018	0.032
$\frac{M_k}{\sum_{i=1}^{I}M_i}$	AIT	0.054	-0.022	0	0.087	0	-0.025	0.006	0
	KL	-0.022	0.127	0	0.013	0	-0.022	0.005	0

Table 3.1: Diet 1: Bias results where Bias = MOL - π. (Based on 1000 pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
	π	0.09	0	0.09	0	0.45	0.09	0.09	0.09
$E_{\mathbf{Y},\mathbf{X}}$	AIT	0.072	0.099	0.022	0.157	-0.242	0.013	0.003	-0.025
$[p_k(\mathbf{Y},\mathbf{X})]$	KL	0.051	0.098	0.036	0.122	-0.194	0.015	-0.006	-0.023
$E_{\mathbf{Y}}$	AIT	0.014	0.043	-0.021	0.155	-0.081	0.027	-0.016	-0.021
$[p_k(\mathbf{Y},\mu_{\mathbf{X}})]$	KL	-0.004	0.098	-0.009	0.105	-0.049	-0.013	-0.012	-0.016
$\frac{\eta_k}{\sum_{i=1}^{I}\eta_i}$	AIT	0.005	0.036	-0.030	0.151	-0.048	0.024	-0.014	-0.024
	KL	-0.012	0.093	-0.018	0.102	-0.019	-0.018	-0.010	-0.019
$\frac{\lambda_k}{\sum_{i=1}^{I}\lambda_i}$	AIT	0.005	0.037	-0.028	0.145	-0.033	0.016	-0.011	-0.031
	KL	-0.012	0.085	-0.017	0.103	-0.003	-0.019	-0.010	-0.028
$\frac{\lambda_k^s}{\sum_{i=1}^{I}\lambda_i^s}$	AIT	0.004	0.037	-0.029	0.152	-0.040	0.017	-0.011	-0.030
	KL	-0.012	0.087	-0.017	0.098	0.004	-0.019	-0.014	-0.027
$\frac{M_k}{\sum_{i=1}^{I}M_i}$	AIT	0.020	0	-0.031	0.133	0.027	-0.017	0	-0.031
	KL	0.009	0.030	-0.023	0.054	0.126	-0.088	0.030	-0.040

Table 3.2: Diet 4: Bias results where Bias = MOL - π. (Based on 1000 pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

investigate this and similar occurrences further, a hierarchical cluster analysis was performed on $\bar{\mathbf{X}}_k$ and the resulting tree is given in Figure 3.3. For both the AIT and KL distances, the tree does show Sandlance to be somewhat similar to Pollock. The tree also helps to explain the opposite bias effect occurring in Haddock and Plaice for both diets, since these species appear to be very similar.

With the exception of Pollock (Diet 4), the bias in the median, $\frac{M_k}{\sum_{i=1}^{I} M_i}$, is zero or near zero when the AIT distance measure is used and $\pi_k = 0$. (For Diet 1, the distance measures are almost equivalent for this case.) Observe also the zero bias in $\frac{M_k}{\sum_{i=1}^{I} M_i}$ in Winter Flounder (Diet 4) whose true diet is 0.09. Consequently when $\pi_k \approx 0$, we expect $\frac{M_k}{\sum_{i=1}^{I} M_i}$ to be a very good parameterization of the diet.

Recall from Section 2.4 that by the Delta Method, μ_{p_k}, η_k, λ_k, and λ_k^s should be similar and furthermore, if $p_k(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}}) \sim SMixSM(\mu_k, \sigma_k^2, \alpha_k)$ then $\lambda_k = \lambda_k^s$. (The distribution of the DMA estimates is discussed in Section 3.4.) For a given distance measure, this appears to be the case and μ_{p_k}, $\frac{\eta_k}{\sum_{i=1}^{I} \eta_i}$, $\frac{\lambda_k}{\sum_{i=1}^{I} \lambda_i}$, and $\frac{\lambda_k^s}{\sum_{i=1}^{I} \lambda_i^s}$ behave roughly similarly and adequately.

Overall, there does not appear to be one best parameterization though $\frac{M_k}{\sum_{i=1}^{I} M_i}$ may be the preferred choice when $\pi_k \approx 0$. Apart from $\mathrm{E}_{\mathbf{Y},\mathbf{X}}[\mathbf{p}(\mathbf{Y}, \mathbf{X})]$, the MOLs are all usually reasonably close to π_k. As in Subsection 2.5.2, the CI method will usually determine the choice of MOL.

3.4 Distribution of the DM Algorithm Estimates

While the DMA estimates are not based on any distributional assumptions (as are the diet estimation methods discussed in Section 3.6), having a parametric model is useful in interval estimation. (Note however that some nonparametric CI methods have been developed as well. These, along with the parametric CIs, are discussed in Chapter 4). In this section, we first consider parametric modeling of the DMA estimates and subsequently modeling estimates obtained by aggregating over the sample of seals.

We have chosen to attempt to model the DMA estimates directly. That is, while it may be possible to model the seal and prey FA signatures and then to derive, at least approximately, the distribution of the DMA estimates, we will show that our simpler approach appears to work sufficiently well. Recall from Section 3.3 that

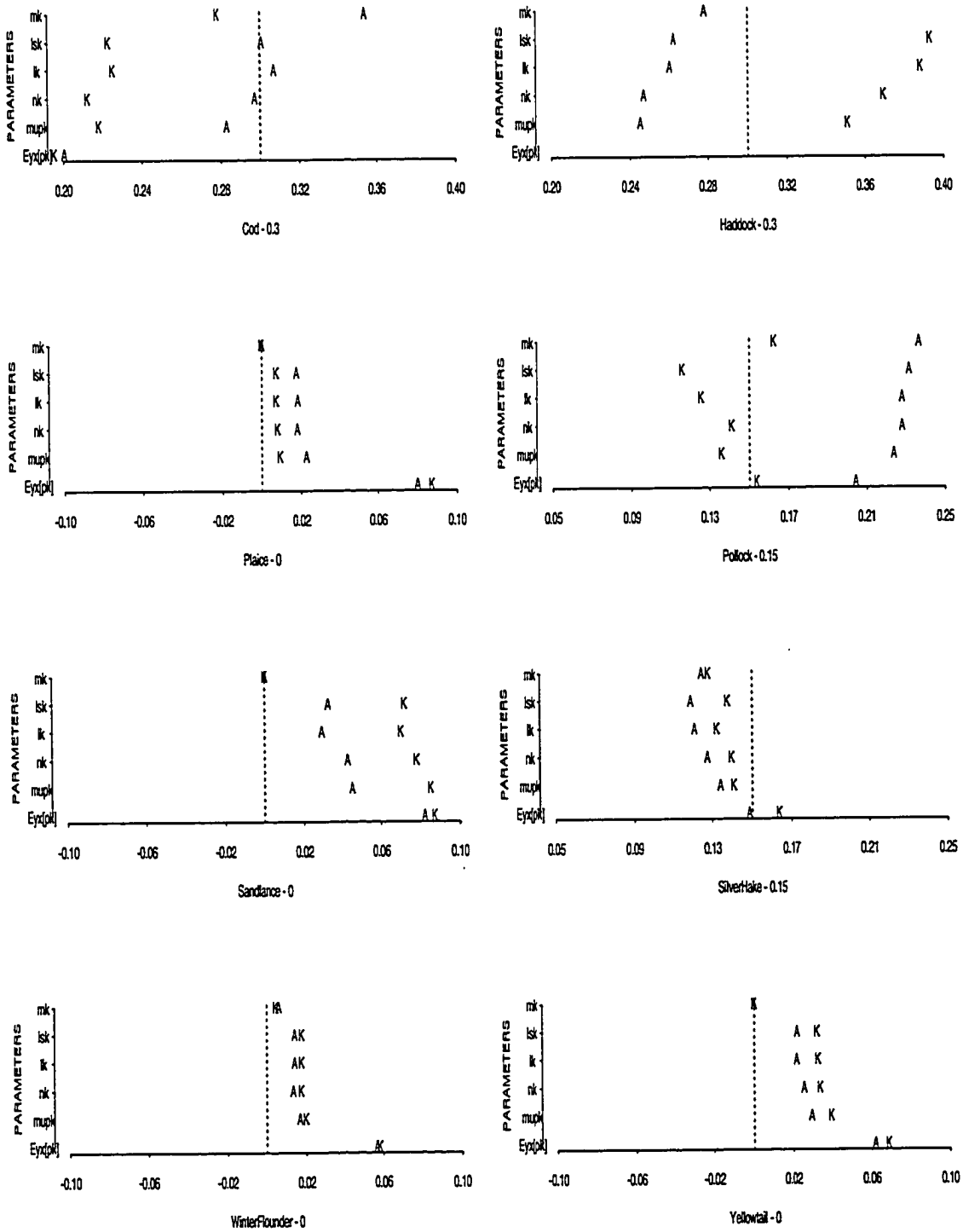

Figure 3.1: Diet 1: Comparison of MOLs and true diet, π (indicated by dashed line). Eyx[pk]$\equiv \mu_{p_k}$, mupk$\equiv \mathrm{E}_{\mathbf{Y}}[p_k(\mathbf{Y}, \mu_{\mathbf{X}})]$, nk $\equiv \frac{\eta_k}{\sum_{i=1}^{I} \eta_i}$, lk$\equiv \frac{\lambda_k}{\sum_{i=1}^{I} \lambda_i}$, lsk$\equiv \frac{\lambda_k^s}{\sum_{i=1}^{I} \lambda_i^s}$, mk$\equiv \frac{m_k}{\sum_{i=1}^{I} m_i}$, K$\equiv$ KL Distance, A$\equiv$AIT Distance. (Based on 1000 pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

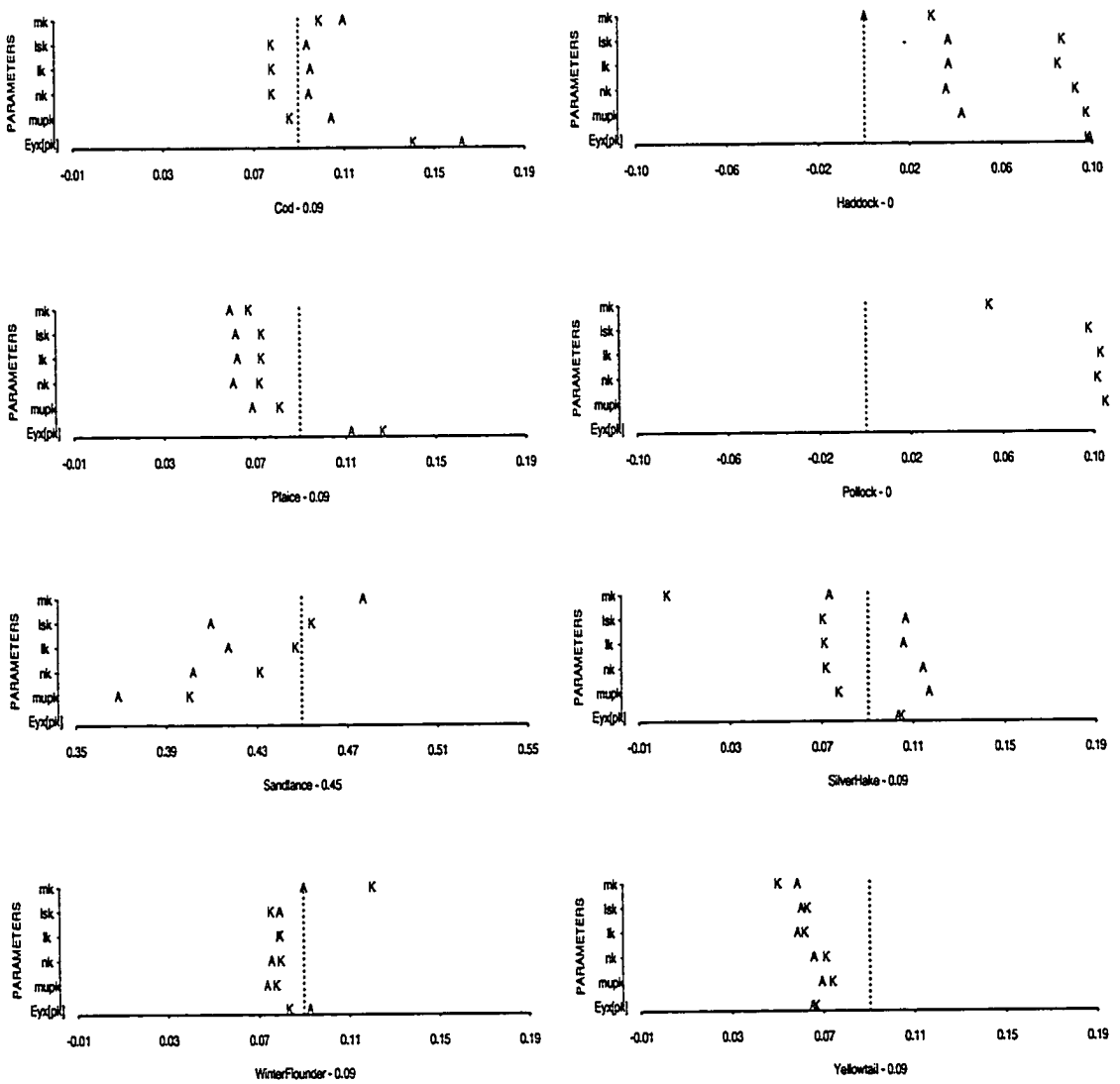

Figure 3.2: Diet 4: Comparison of MOLs and true diet, π (indicated by dashed line). Eyx[pk]$\equiv \mu_{p_k}$, mupk$\equiv \mathrm{E}_{\mathbf{Y}}[p_k(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})]$, nk $\equiv \frac{\eta_k}{\sum_{i=1}^{I} \eta_i}$, lk$\equiv \frac{\lambda_k}{\sum_{i=1}^{I} \lambda_i}$, lsk$\equiv \frac{\lambda_k^s}{\sum_{i=1}^{I} \lambda_i^s}$, mk$\equiv \frac{m_k}{\sum_{i=1}^{I} m_i}$, K$\equiv$ KL Distance, A$\equiv$AIT Distance. (Based on 1000 pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

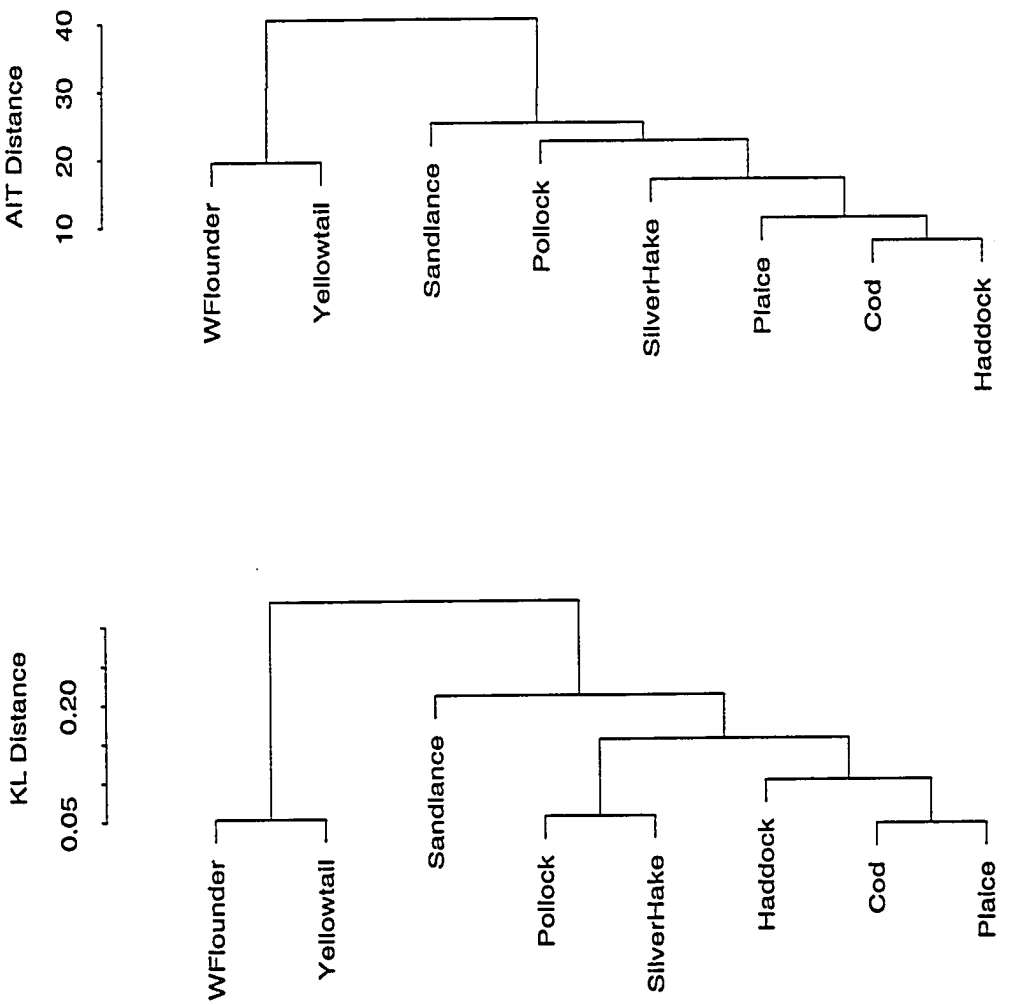

Figure 3.3: Hierarchical cluster analysis on $\bar{\mathbf{X}}_k$, $k = 1, \ldots, 8$, using *hclust* in S-PLUS with both the AIT and KL distance measures, and the average linkage method.

the recommended DMA based estimate of diet for a seal with FA signature $\mathbf{Y}$ is $\mathbf{p}(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})$. The population of diet estimates would then be all possible diet estimates that could be obtained from all possible seal FA signatures when the true diet is $\boldsymbol{\pi}_0$. A subtleness in modeling these estimates in practice occurs because $\boldsymbol{\mu}_{\mathbf{X}}$ is unknown, in which case the population of estimates can be defined in different ways. If we ignore for the moment the issue of $\boldsymbol{\mu}_{\mathbf{X}}$ being unknown in practice, we may consider modeling the estimates with one of the mixture distributions discussed in Section 2.3. In our current notation, the $MixM$, $SMixM$ and $SMixSM$ distributions are

$$f_k(p_k) = \begin{cases} \sum_{\{b:k\notin \mathbf{v}_b\}} \theta_{\mathbf{v}_b} & \text{if } p_k = 0, \\ \sum_{\{b:k\in \mathbf{v}_b\}} \theta_{\mathbf{v}_b} \mathcal{M}(\mu_k^{\mathbf{v}_b}, \sigma_k^2) & \text{if } 0 < p_k < 1, \\ 0 & \text{otherwise,} \end{cases} \tag{3.5}$$

$$f_k(p_k) = \begin{cases} \theta_k & \text{if } p_k = 0, \\ (1-\theta_k)\mathcal{M}(\mu_k, \sigma_k^2) & \text{if } 0 < p_k < 1, \\ 0 & \text{otherwise,} \end{cases} \tag{3.6}$$

and

$$f_k(p_k) = \begin{cases} \theta_k & \text{if } p_k = 0, \\ (1-\theta_k)\mathcal{SM}(\mu_k, \sigma_k^2, \alpha_k) & \text{if } 0 < p_k < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7}$$

To assess the fit of the continuous part of these distributions, we have fit these distributions to the (non-zero) diet estimates that were used to obtain Figures 3.1-3.2. The parameters in densities 3.5, 3.6 and 3.7 were estimated by their MLEs (as derived in Subsection 2.5.1). The estimated densities along with a histogram of the (non-zero) diet estimates are given in Figures 3.4-3.7. For comparison purposes, we have included the fit of the normal distribution with mean $\hat{\mu}_{p_k} = \frac{1}{1000}\sum_{i=1}^{1000} p_{k,i}(\mathbf{Y}_i, \boldsymbol{\mu}_{\mathbf{X}})$ and variance $s_{p_k} = \frac{1}{999}\sum_{i=1}^{1000}(p_{k,i}(\mathbf{Y}_i, \boldsymbol{\mu}_{\mathbf{X}}) - \hat{\mu}_{p_k})^2$.

The figures show the overall shape and appropriateness of the fits to be roughly the same for both distance measures. One noticeable difference is the larger spread in the distribution for Haddock with the KL distance (both diets) than with the AIT distance. Conversely, the diet estimates of Pollock are more variable with the AIT distance, in both diets, than with the KL distance.

Perhaps what is most obvious from the figures is the poor fits by the normal distribution, particular when π_k is small. Also apparent is the similarity between the fits of the $MixM$ and $SMixM$ distributions with the exception being Sandlance in Diet 4. In this case, the $MixM$ distribution attempts to capture the seemingly multimodal nature of the distribution. The closeness of the $MixM$ and $SMixM$ distributions for the other species suggest that the more complicated $MixM$ distribution is not always needed. Particularly when π_k is not large, both mixture distributions appear to provide moderate fits.

The $SMixSM$ distribution quite often provides a considerable improvement in the fit over the $MixM$ and $SMixM$ distributions and appears to fit the data quite well.

In practice $\boldsymbol{\mu}_{\mathbf{X}}$ is unknown and the definition of the population of the diet estimates must be altered slightly. One option is to consider the population of the diet estimates conditional on $\bar{\mathbf{X}}$. Another allows the population of diet estimates to contain all possible estimates when the both the seal and sample of prey FA signatures vary. As will be discussed in Chapter 4, parametric CIs will tend to be more easily derived when we consider the distribution of the estimates conditional on $\bar{\mathbf{X}}$, mainly because our sample estimates will be independent (if the seal FA signatures are independent). Our approach will be to assume that our diet estimates conditional on $\bar{\mathbf{X}}$ may be modeled by one of the mixture distributions and to derive any parametric CIs accordingly. To incorporate the variability of the prey into the intervals we will either use a bootstrapping procedure or one of the other DMAs (that is, apart from the MEAN method DMA).

Given a sample of n_s seal and prey FA signatures, we now consider aggregate point estimators of the DMA based parameters discussed in Section 3.3, and examine their distributions.

For a sample $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_s}$ and $\mathbf{X}_1, \ldots, \mathbf{X}_I$, Table 3.3 contains point estimators of the MOLs discussed in Section 3.3. (These point estimators are analogous to those given in Table 2.1.) We will adopt the notation, $\bar{p}_k(\cdot)$ (and similarly for the other point estimators), to allow for the specification of $\bar{\mathbf{X}}$ or $\boldsymbol{\mu}_{\mathbf{X}}$.

Observe that these estimators are not unbiased for their corresponding parameters

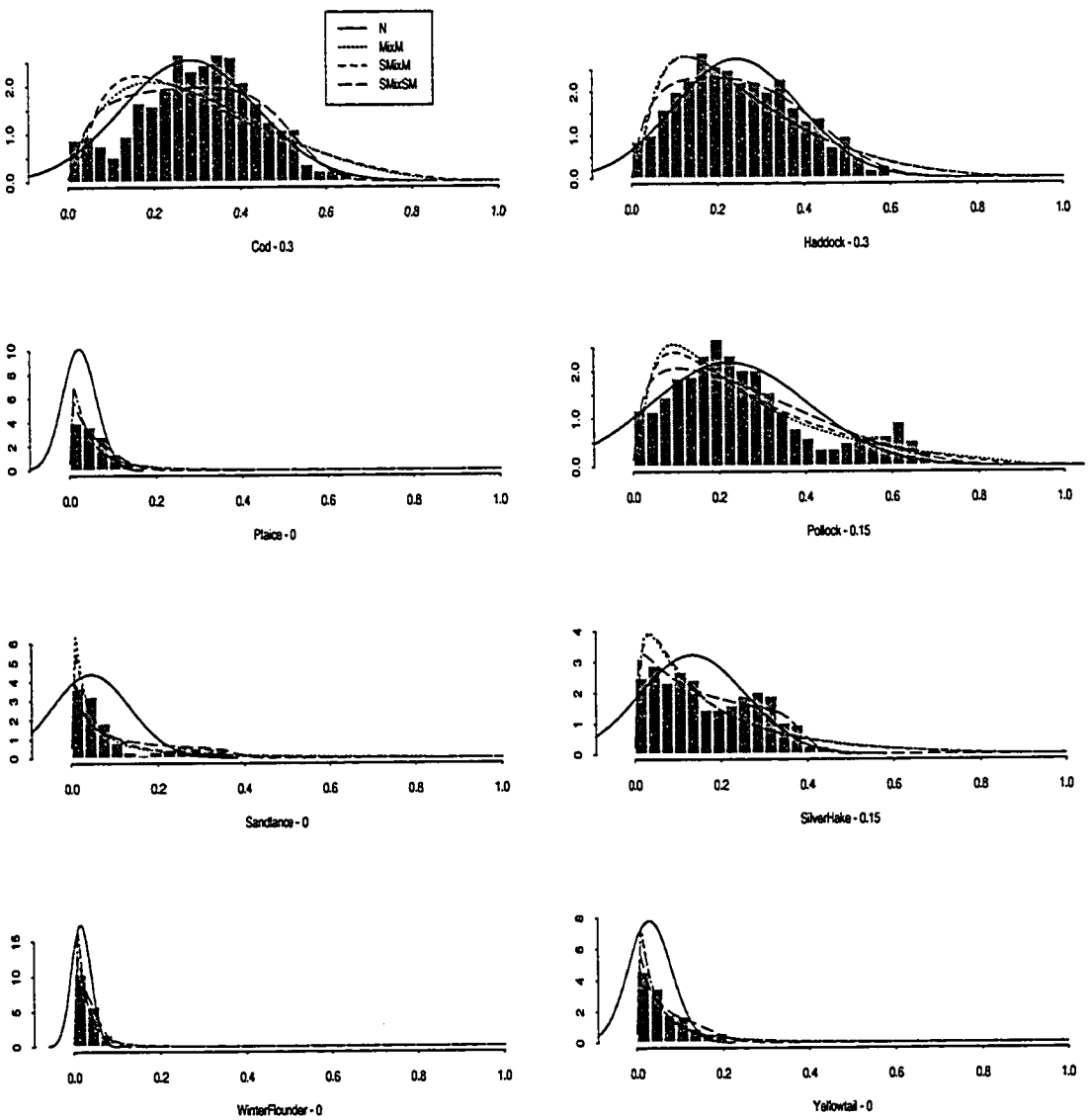

Figure 3.4: Diet 1, AIT distance: Histograms of (non-zero) diet estimates and various estimated distributions. (Based on 1000 pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

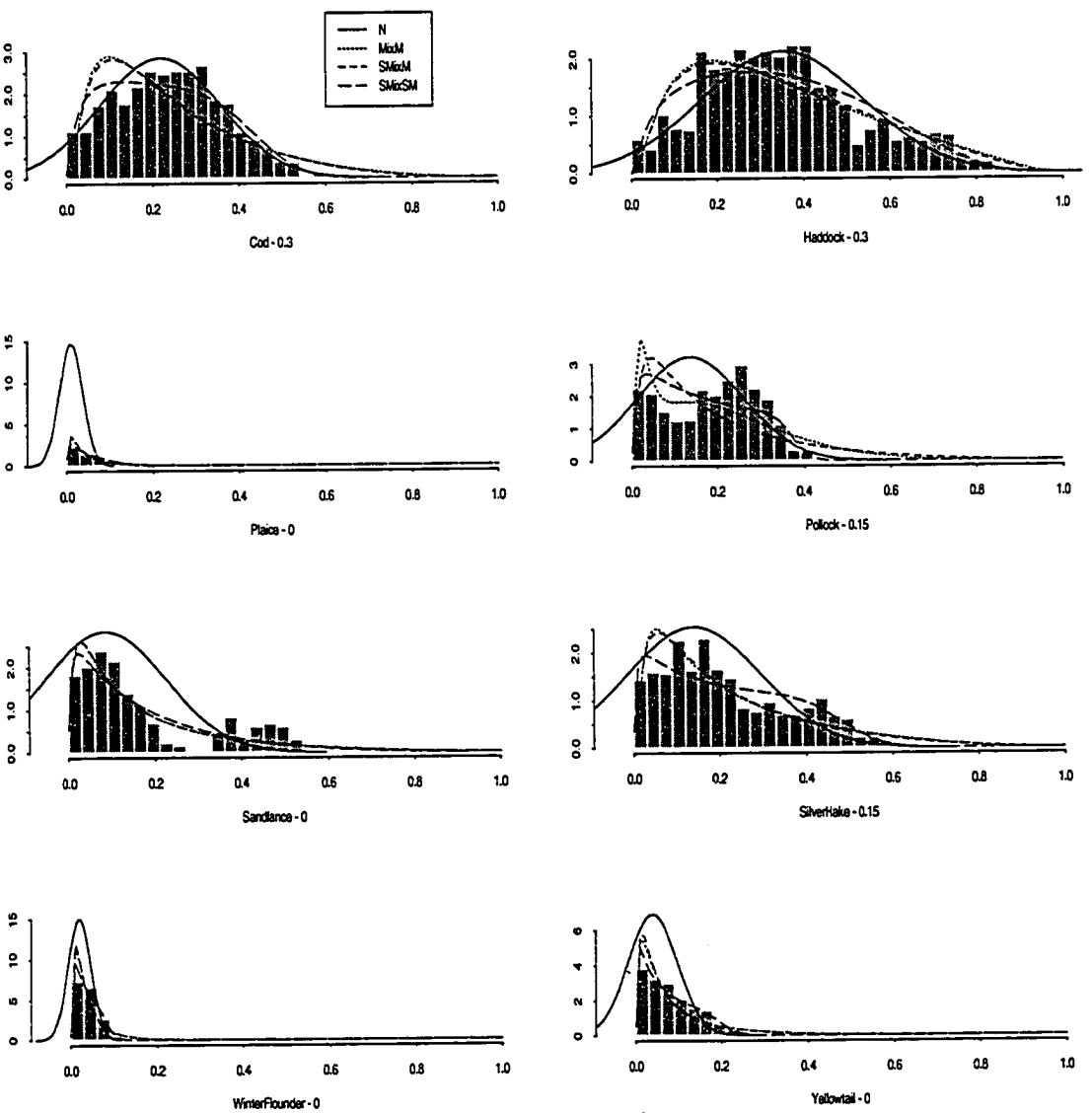

Figure 3.5: Diet 1, KL distance: Histograms of (non-zero) diet estimates and various estimated distributions. (Based on 1000 pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

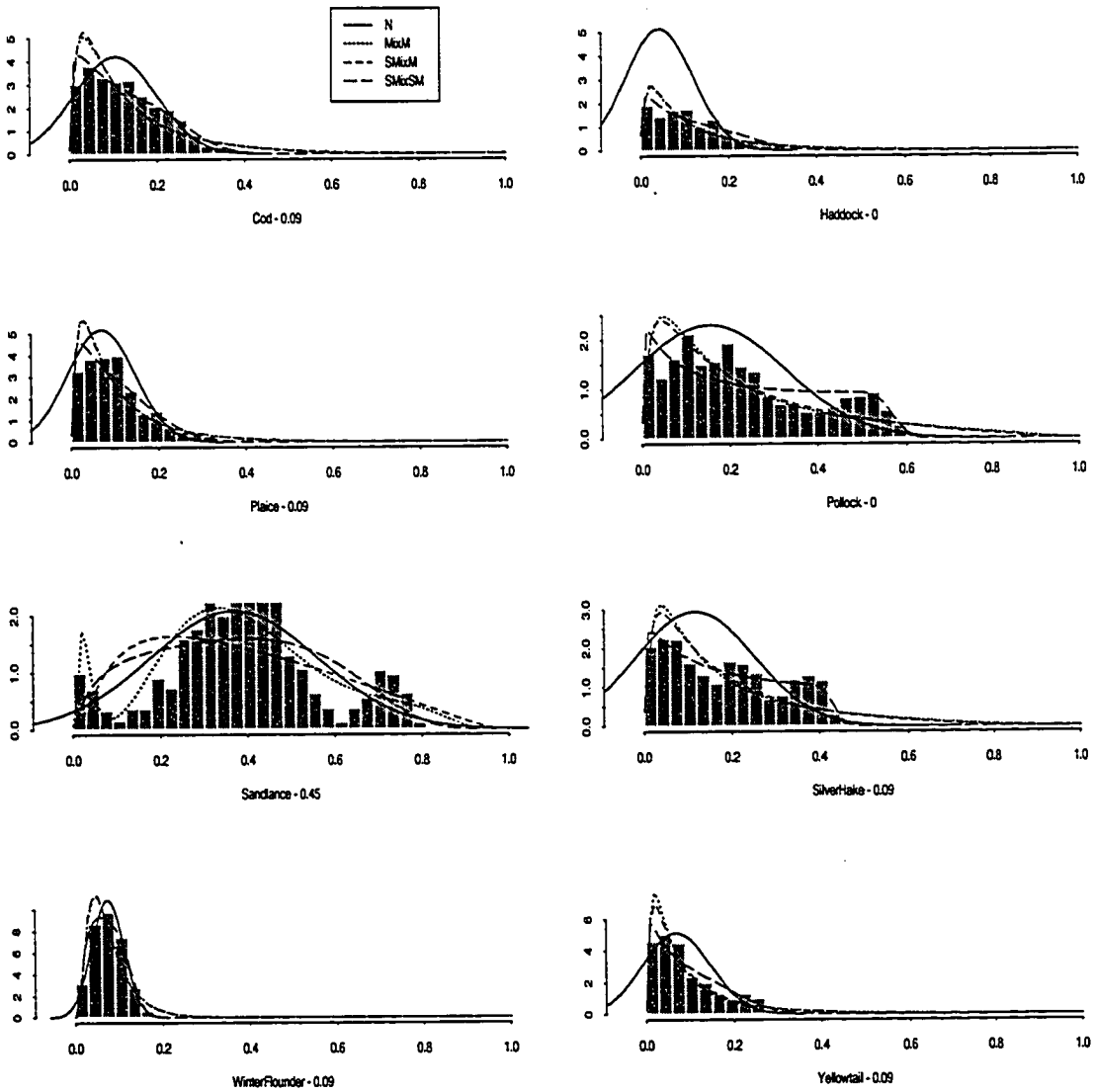

Figure 3.6: Diet 4, AIT distance: Histograms of (non-zero) diet estimates and various estimated distributions. (Based on 1000 pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

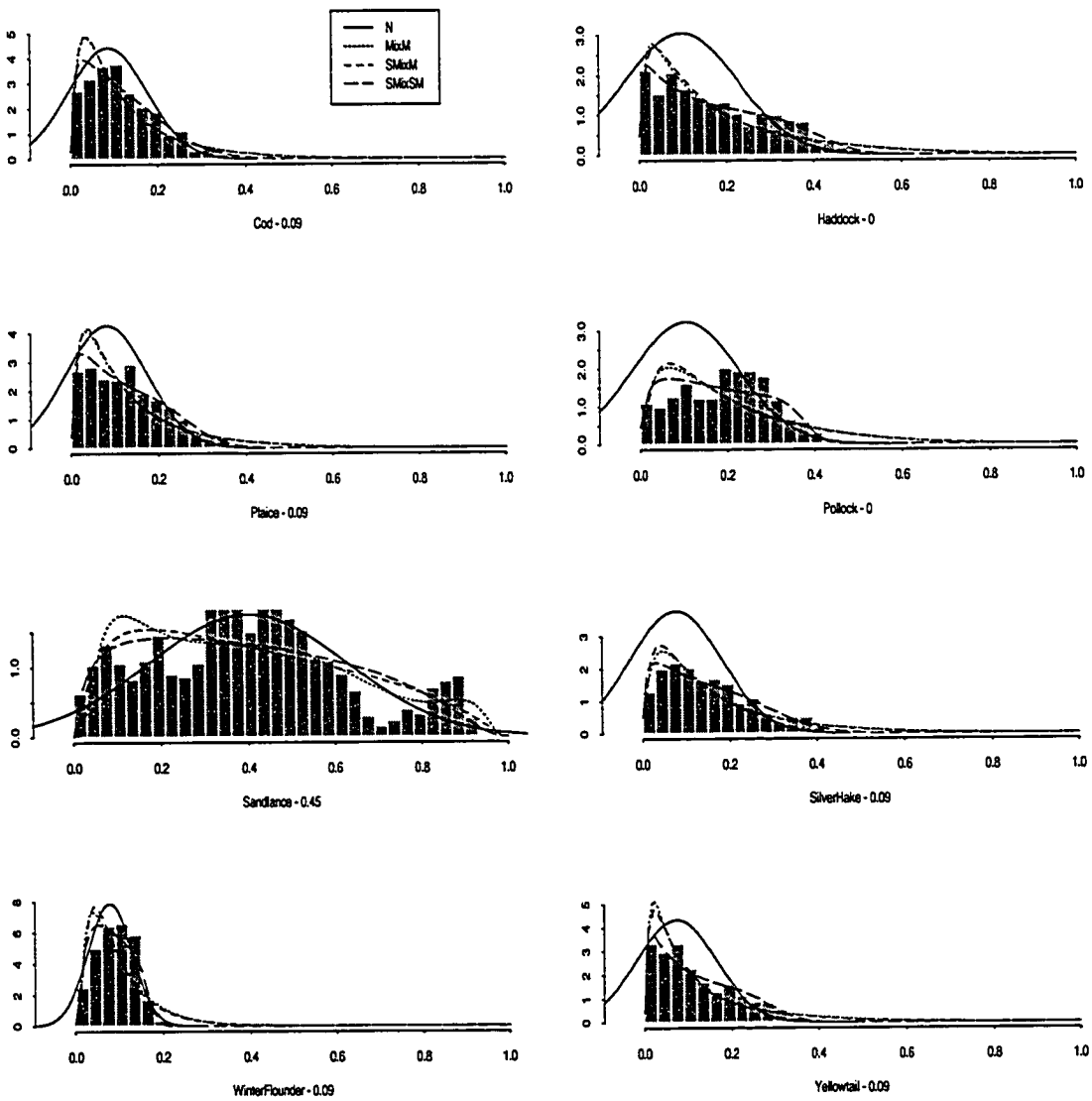

Figure 3.7: Diet 4, KL distance: Histograms of (non-zero) diet estimates and various estimated distributions. (Based on 1000 pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

essentially because

$$\mathrm{E}_{\mathbf{Y},\bar{\mathbf{X}}}[p_k(\mathbf{Y},\bar{\mathbf{X}})] \neq \mathrm{E}_{\mathbf{Y}}[p_k(\mathbf{Y},\boldsymbol{\mu}_{\mathbf{X}})].$$

Also with $\bar{\mathbf{X}}$ in place of $\boldsymbol{\mu}_{\mathbf{X}}$, $\mathrm{p}_i(\mathbf{Y}_i,\bar{\mathbf{X}})$, $i = 1,\ldots,n_s$ are not independent and, consequently, we do not have simple variance expressions for the point estimators.

Conditional on $\bar{\mathbf{X}}$, however, we have, for example,

$$\begin{aligned}\mathrm{E}_{\mathbf{Y}}[\bar{p}_k(\bar{\mathbf{X}})|\bar{\mathbf{X}}] &= \mathrm{E}_{\mathbf{Y}}[\mathrm{p}(\mathbf{Y},\bar{\mathbf{X}})|\bar{\mathbf{X}}] = \mu_{p_k|\bar{\mathbf{X}}},\\ \mathrm{VAR}_{\mathbf{Y}}[\bar{p}_k(\bar{\mathbf{X}})|\bar{\mathbf{X}}] &= \frac{\mathrm{VAR}_{\mathbf{Y}}[\mathrm{p}(\mathbf{Y},\bar{\mathbf{X}})|\bar{\mathbf{X}}]}{n_s} = \frac{\sigma^2_{p_k|\bar{\mathbf{X}}}}{n_s}.\end{aligned}$$

Parameters	Point Estimators
$\mu_{p_k} = \mathrm{E}_{\mathbf{Y}}[p_k(\mathbf{Y},\boldsymbol{\mu}_{\mathbf{X}})]$	$\bar{p}_k(\bar{\mathbf{X}}) = \frac{1}{n_s}\sum_{i=1}^{n_s} p_{k,i}(\mathbf{Y}_i,\bar{\mathbf{X}})$
$\eta_k = \sum_{b=1}^{B}\theta_{\mathbf{v}_b}\eta_k(\mathbf{v}_b)$ where $\eta_k\|\mathbf{v}_b = \begin{cases} 0 & \text{if } k \notin \mathbf{v}_b \\ \frac{e^{\mu_k^{\mathbf{v}_b}}}{1+e^{\mu_k^{\mathbf{v}_b}}} & \text{if } k \in \mathbf{v}_b \end{cases}$	$\hat{\eta}_k(\bar{\mathbf{X}}) = \sum_{b=1}^{B}\hat{\theta}_{\mathbf{v}_b}\hat{\eta}_k(\mathbf{v}_b,\bar{\mathbf{X}})$ where $\hat{\eta}_k(\bar{\mathbf{X}})\|\mathbf{v}_b = \begin{cases} 0 & \text{if } k \notin \mathbf{v}_b \\ \frac{e^{\hat{\mu}_k^{\mathbf{v}_b}(\bar{\mathbf{X}})}}{1+e^{\hat{\mu}_k^{\mathbf{v}_b}(\bar{\mathbf{X}})}} & \text{if } k \in \mathbf{v}_b \end{cases}$
$\lambda_k =$ $\begin{cases} 0 & \text{if } \theta_k = 1 \\ (1-\theta_k)\frac{e^{\mu_k}}{1+e^{\mu_k}} & \text{if } \theta_k < 1. \end{cases}$	$\hat{\lambda}_k(\bar{\mathbf{X}}) =$ $\begin{cases} 0 & \text{if } \hat{\theta}_k = 1 \\ (1-\hat{\theta}_k)\frac{e^{\hat{\mu}_k(\bar{\mathbf{X}})}}{1+e^{\hat{\mu}_k(\bar{\mathbf{X}})}} & \text{if } \hat{\theta}_k < 1. \end{cases}$
$\lambda_k^s =$ $\begin{cases} 0 & \text{if } \theta_k = 1 \\ (1-\theta_k)\frac{e^{\xi_k(\mu_k,\sigma_k^2,\alpha_k)}}{1+e^{\xi_k(\mu_k,\sigma_k^2,\alpha_k)}} & \text{if } \theta_k < 1. \end{cases}$	$\hat{\lambda}_k^s(\bar{\mathbf{X}}) =$ $\begin{cases} 0 & \text{if } \hat{\theta}_k = 1 \\ (1-\hat{\theta}_k)\frac{e^{\hat{\xi}_k(\hat{\mu}_k(\bar{\mathbf{X}}),\hat{\sigma}_k^2(\bar{\mathbf{X}}),\hat{\alpha}_k(\bar{\mathbf{X}}))}}{1+e^{\hat{\xi}_k(\hat{\mu}_k(\bar{\mathbf{X}}),\hat{\sigma}_k^2(\bar{\mathbf{X}}),\hat{\alpha}_k)(\bar{\mathbf{X}})}} & \text{if } \hat{\theta}_k < 1. \end{cases}$
$\mathrm{M}_{p_k} = \mathrm{median}_{\mathbf{Y}}[p_k(\mathbf{Y},\boldsymbol{\mu}_{\mathbf{X}})]$ (Population Median)	$\mathrm{m}_{p_k} = \mathrm{median}_i[p_{k,i}(\mathbf{Y}_i,\bar{\mathbf{X}})]$ (Sample Median)

Table 3.3: MOLs and their point estimators

We will not attempt to derive the conditional expectations and variances of the other point estimators nor will we derive the distribution of these point estimators for finite samples. These are nontrivial tasks and for interval estimation, other methods can be used. Instead we have examined the finite sample distribution of the point estimators through plots of simulated estimates as in Section 3.4. The estimates were simulated as follows:

1. *for* $r = 1:1000$

   (a) Generate n_s pseudo-seals, $\mathbf{Y}_1,\ldots,\mathbf{Y}_{ns}$ with diet π, using $\mathbf{X}_1,\ldots,\mathbf{X}_I$.

(b) Compute $\mu_{\mathbf{X}_k} \equiv \bar{\mathbf{X}}_k$, $k = 1, \ldots, I$.

(c) Compute $\mathrm{p}_i^r(\mathbf{Y}_i, \mu_{\mathbf{X}}), i = 1, \ldots, n_s$.

(d) Compute each of the point estimators in Table 3.3. For example, compute $\bar{p}_k^r(\mu_{\mathbf{X}})$.

Note that as before, we are treating the sample of prey FA signatures as the population.

Histograms of the point estimators in Table 3.3 along with the fit of the normal distribution are given in Figures 3.8-3.11. As some skewness in the distributions was expected to be present when n_s is small, the fit of the univariate skew-normal ($\mathcal{SN}$) distribution (as defined in Chapter 2) is also included for comparison with the normal distribution. We have chosen to illustrate only the distributions of the point estimators at $n_s = 5$, for Diet 4, and with the AIT distance measure. The distribution of the point estimators are similar for Diet 1 and the KL distance measure. Furthermore, by $n_s = 10$, all of the point estimators appeared to be roughly normally distributed (particularly for $\pi_k > 0$). Note that we have not shown the distribution of $\hat{\lambda}_k^s$ since usually $\hat{\lambda}_k^s$ is very similar to $\hat{\lambda}_k$.

From the Figures, even at $n_s = 5$, the distributions of $\bar{p}_k$, $\hat{\eta}_k$ and $\hat{\lambda}_k$ appear to be approximately normally distributed when $\pi_k > 0.15$. For $\pi_k \leq 0.15$, the estimates are well fit by the $\mathcal{SN}$ distribution. Note the similarity in the plots in Figures 3.8-3.10 and in particular the closeness between the distributions of $\bar{p}_k$ and $\hat{\eta}_k$. Recall from Section 2.4 that by the Delta method, $\mu_{p_k} \approx \eta_k$ and $\mu_{p_k} \approx \lambda_k$ so that the point estimators $\bar{p}_k$, $\hat{\eta}_k$ and $\hat{\lambda}_k$ should be similar. Also, for each population $\mathbf{V}_b$ (where $\mathbf{V}_b$ gives the non-zero components of $\mathbf{p}(\mathbf{Y}, \mu_{\mathbf{X}})$), if there is only one observation, then $\bar{p}_k = \hat{\eta}_k$. Consequently for n_s small, $\bar{p}_k$ and $\hat{\eta}_k$ are often the same. The median, m_{p_k}, appears to require a larger n_s for the normal approximation to suffice.

The tendency of the finite sample distributions of the point estimators in Table 3.3 towards normality is in agreement with Section 3.5 where it will be shown that the point estimators are all asymptotically normally distributed. Based on our investigation, a normal approximation to the finite sample distributions generally appears to be valid for $n_s \geq 5$ when $\pi_k > 0.15$, for $n_s \geq 10$ when $\pi_k > 0$ (for the median, m_{p_k}, a larger n_s may be required), and for $n_s \geq 25$ for all π_k. The implication is that CIs

based on a normal approximation should roughly be appropriate in these cases.

3.5 Asymptotic Properties of the Point Estimators

We now consider the behaviour of the point estimators in Table 3.3 when n_s and/or n_k are large. Based on the plots and discussions of the distributions of the finite sample estimators presented in Section 3.4, we expect the large sample distributions to be approximately normal. In this section we prove this to be the case when both n_s and n_k are large.

We begin with the simplified case of $\mu_{\mathbf{X}}$ being known and examine properties of the estimators as $n_s \to \infty$. In this ideal case, if $\mathbf{Y}_1, \ldots, \mathbf{Y}_{ns}$ is a random sample of the seal FA signatures, then $p_{k,1}(\mathbf{Y}_1, \mu_{\mathbf{X}}), \ldots, p_{k,n_s}(\mathbf{Y}_{ns}, \mu_{\mathbf{X}})$ are independent identically distributed (iid) random variables with common mean $\mu_{p_k} = \mathrm{E}_{\mathbf{Y}}[\mathrm{p}(\mathbf{Y}, \mu_{\mathbf{X}})]$ and variance $\sigma^2_{p_k} = \mathrm{VAR}_{\mathbf{Y}}[p_k(\mathbf{Y}, \mu_{\mathbf{X}})]$. Then Theorems 2.1-2.3 are applicable and the asymptotic distributions of the estimators are evident. For example, we have, by the Central Limit Theorem,

$$\frac{\bar{p}_{k,n_s}(\mu_{\mathbf{X}}) - \mu_{p_k}}{\frac{\sigma_{p_k}}{\sqrt{n_s}}} \to_d \mathcal{N}(0,1). \tag{3.8}$$

If $\mu_{p_k|\bar{\mathbf{X}}}$ and $\sigma^2_{p_k|\bar{\mathbf{X}}}$ are defined as in Section 3.4 then we have also that

$$\frac{\bar{p}_{k,n_s|\bar{\mathbf{X}}}(\bar{\mathbf{X}}) - \mu_{p_k|\bar{\mathbf{X}}}}{\frac{\sigma_{p_k|\bar{\mathbf{X}}}}{\sqrt{n_s}}} \to_d \mathcal{N}(0,1), \tag{3.9}$$

since $p_{k|\bar{\mathbf{X}}}(\mathbf{Y}_i, \bar{\mathbf{X}})$, $i = 1, \ldots, n_s$ are iid. Conditional on $\bar{\mathbf{X}}$, the asympotic distribution of the other point estimators may be similarly derived.

When $\mu_{\mathbf{X}}$ is unknown, as is the case in practice, the asymptotic distributions of the estimators (for both n_s and n_k large) are also relatively straightforward to derive, provided it can be shown that $p_{k,n}(\mathbf{Y}, \bar{\mathbf{X}}_n)$ converges in probability to $p_k(\mathbf{Y}, \mu_{\mathbf{X}})$ ($p_{k,n}(\mathbf{Y}, \bar{\mathbf{X}}_n) \to_p p_k(\mathbf{Y}, \mu_{\mathbf{X}})$), or equivalently that for every $\epsilon > 0$,

$$\lim_{n\to\infty} P[|p_{k,n}(\mathbf{Y}, \bar{\mathbf{X}}_n) - p_k(\mathbf{Y}, \mu_{\mathbf{X}})| \geq \epsilon] = 0. \tag{3.10}$$

To prove this, first note that by the weak law of large numbers (WLLN), we have that

$$\bar{\mathbf{X}}_{n_k} \to_p \mu_{\mathbf{X}_k} \ \forall k.$$

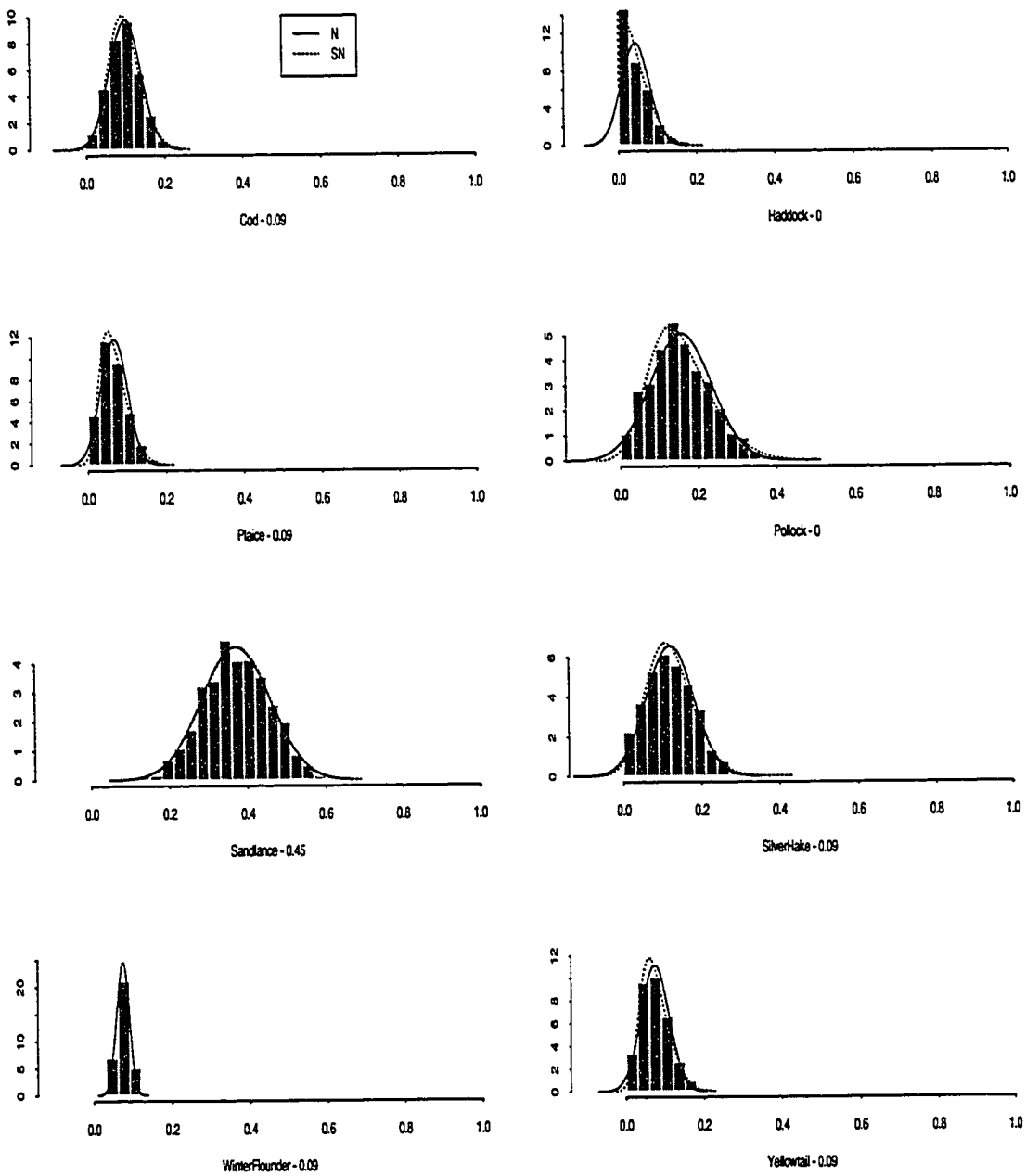

Figure 3.8: Diet 4, $n_s = 5$, AIT distance: Distribution of $\bar{p}_k$. (Based on 1000 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

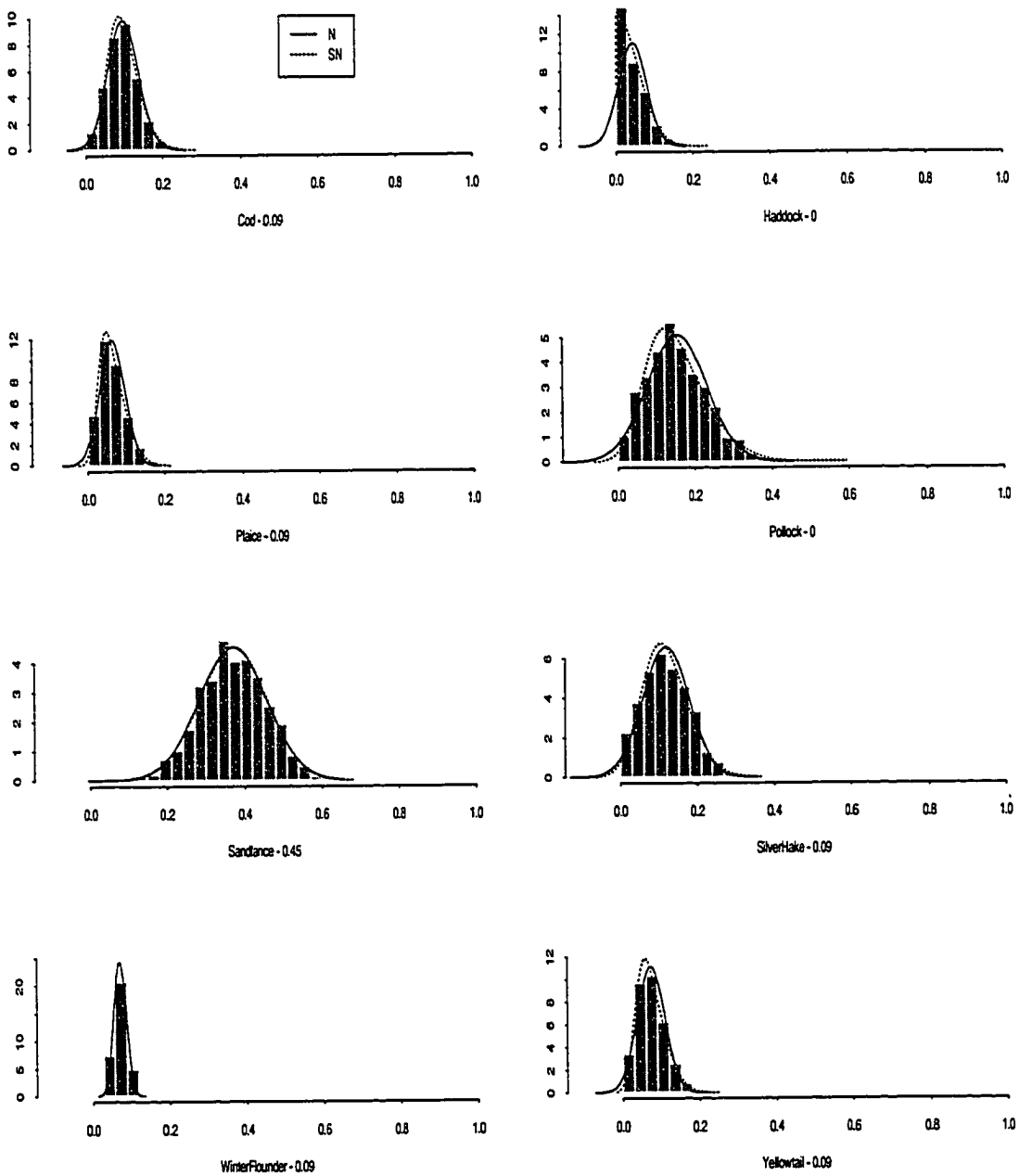

Figure 3.9: Diet 4, $n_s = 5$, AIT distance: Distribution of $\hat{\eta}_k$. (Based on 1000 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

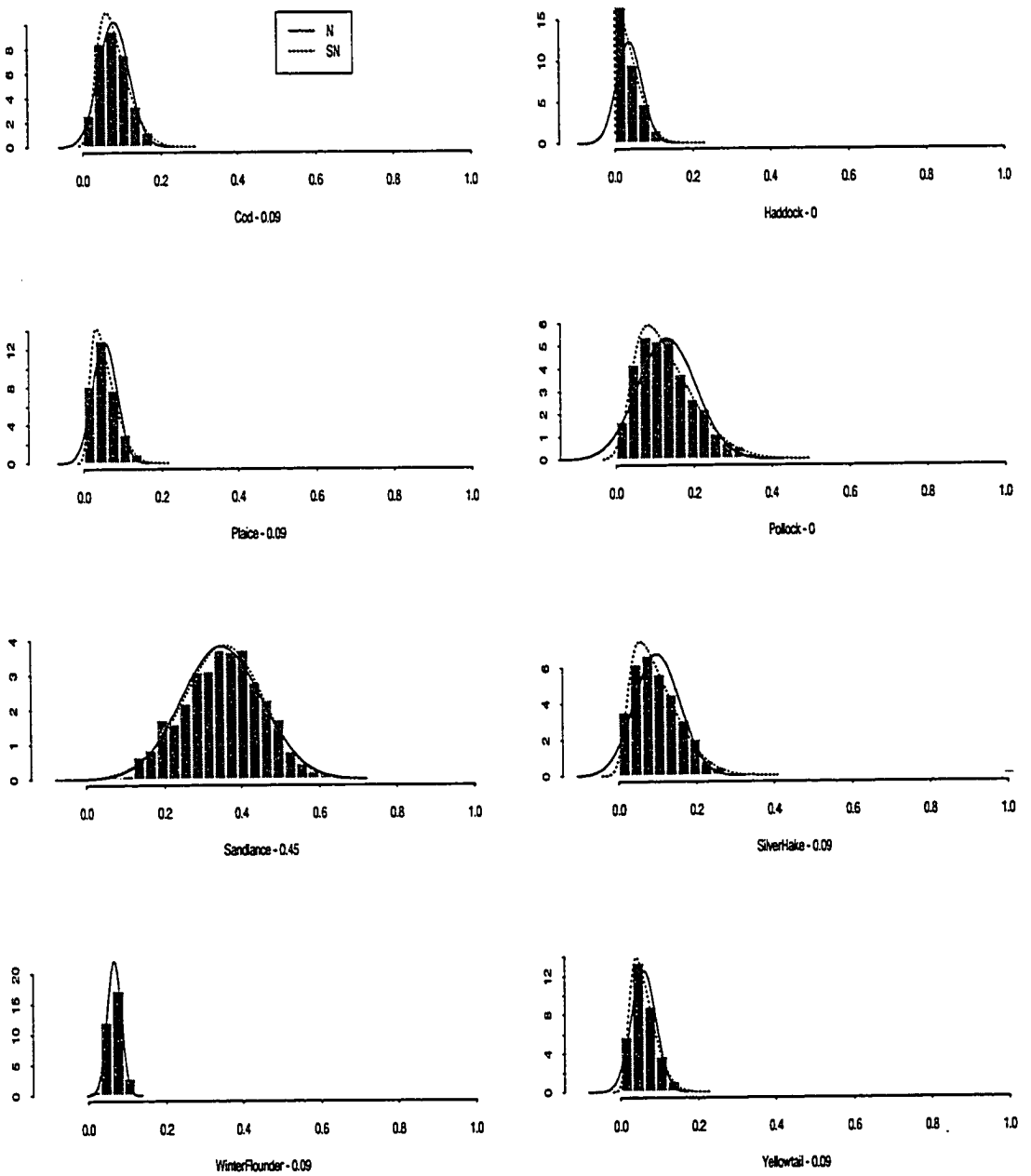

Figure 3.10: Diet 4, $n_s = 5$, AIT distance: Distribution of $\hat{\lambda}_k$. (Based on 1000 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

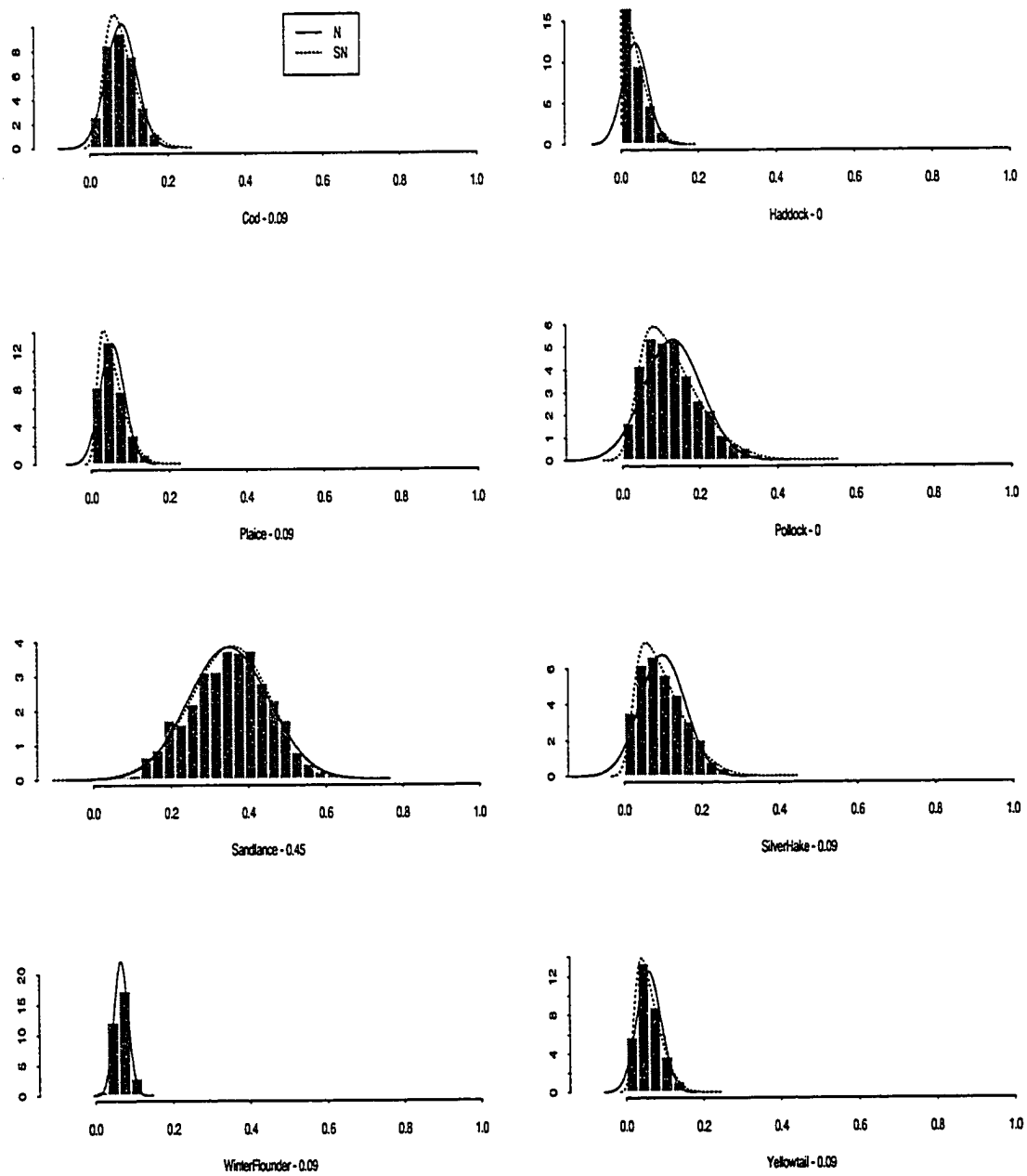

Figure 3.11: Diet 4, $n_s = 5$, AIT distance: Distribution of m_{p_k}. (Based on 1000 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

By the following theorem, it then suffices to show that $\mathbf{p}_n(\mathbf{Y}, \bar{\mathbf{X}}_n)$ is continuous from $\mathcal{S}^{n_{FA}(I+1)} \rightarrow \mathcal{S}^I$.

Theorem 3.1 *Let $\mathbf{X}_n$ be a sequence of random vectors of length D. If $\mathbf{X}_n \rightarrow_p \mathbf{X}$ and g is continuous from $\mathcal{R}^D$ to $\mathcal{R}^M$, then $g(\mathbf{X}_n) \rightarrow_p g(\mathbf{X})$. (Bickel and Doksum, 2001)*

Consequently, to show that $\mathbf{p}_n(\mathbf{Y}, \bar{\mathbf{X}}_n) \rightarrow_p \mathbf{p}(\mathbf{Y}, \mu_{\mathbf{X}})$ we will attempt the simpler task of proving that $\mathbf{p}_n(\mathbf{Y}, \bar{\mathbf{X}}_n)$ is a continuous function of $\bar{\mathbf{X}}_n$ or equivalently that

$$\lim_{n\rightarrow\infty} \mathbf{p}_n(\mathbf{Y}, \bar{\mathbf{X}}_n) = \mathbf{p}(\mathbf{Y}, \mu_{\mathbf{X}}) \ \forall (\mathbf{Y}, \mu_{\mathbf{X}}) \in \mathcal{S}^{n_{FA}(I+1)}. \tag{3.11}$$

First observe the following two statements

1. Since $\mathbf{p}_n(\mathbf{Y}, \bar{\mathbf{X}}_n)$ lies in a compact set (that is, since $0 \leq \mathbf{p}_n(\mathbf{Y}, \bar{\mathbf{X}}_n) \leq 1$), $\mathbf{p}_n(\mathbf{Y}, \bar{\mathbf{X}}_n)$ has a convergent subsequence. Without loss of generality, assume that the sequence $\mathbf{p}_n(\mathbf{Y}, \bar{\mathbf{X}}_n)$ is the convergent subsequence. Then let

$$\lim_{n\rightarrow\infty} \mathbf{p}_n(\mathbf{Y}, \bar{\mathbf{X}}_n) = \mathbf{a}. \tag{3.12}$$

2.

$$\begin{aligned} & \mathbf{p}_n(\mathbf{Y}, \bar{\mathbf{X}}_n) = \arg\min_{\mathbf{q}} \operatorname{dist}(\mathbf{Y}, \hat{\mathbf{Y}}(\mathbf{q}, \bar{\mathbf{X}}_n)) \\ \Leftrightarrow \ & \operatorname{dist}(\mathbf{Y}, \hat{\mathbf{Y}}(\mathbf{p}_n, \bar{\mathbf{X}}_n)) \leq \operatorname{dist}(\mathbf{Y}, \hat{\mathbf{Y}}(\mathbf{q}, \bar{\mathbf{X}}_n)) \ \forall \mathbf{q}, n, \end{aligned} \tag{3.13}$$

where dist(·) represents either the KL or AIT distance functions in Equations 3.1 and 3.2 respectively.

It is straightforward to show that the KL and AIT distance functions are continuous in $\bar{\mathbf{X}}_n$ and $\mathbf{q}$ so that

$$\begin{aligned} \lim_{n\rightarrow\infty} \operatorname{dist}(\mathbf{Y}, \hat{\mathbf{Y}}(\mathbf{q}, \bar{\mathbf{X}}_n)) &= \operatorname{dist}(\mathbf{Y}, \hat{\mathbf{Y}}(\mathbf{q}, \mu_{\mathbf{X}})) \ \forall \mathbf{q} \text{ and} \\ \lim_{n\rightarrow\infty} \operatorname{dist}(\mathbf{Y}, \hat{\mathbf{Y}}(\mathbf{p}_n, \bar{\mathbf{X}}_n)) &= \operatorname{dist}(\mathbf{Y}, \hat{\mathbf{Y}}(\mathbf{a}, \mu_{\mathbf{X}})). \end{aligned}$$

Now since Equation 3.13 is satisfied for all n, we may take the limit of both sides of the inequality to obtain that

$$\operatorname{dist}(\mathbf{Y}, \hat{\mathbf{Y}}(\mathbf{a}, \mu_{\mathbf{X}})) \leq \operatorname{dist}(\mathbf{Y}, \hat{\mathbf{Y}}(\mathbf{q}, \mu_{\mathbf{X}})) \ \forall \mathbf{q}.$$

Therefore,

$$\begin{aligned} \mathbf{a} &= \arg\min_{\mathbf{q}} \operatorname{dist}(\mathbf{Y}, \hat{\mathbf{Y}}(\mathbf{q}, \boldsymbol{\mu}_{\mathbf{X}})) \\ &= \mathbf{p}(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}}). \end{aligned}$$

Then Equation 3.11 is satisfied by replacing $\mathbf{a}$ with $\mathbf{p}(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})$ in Equation 3.12. We may conclude that $\mathbf{p}_n(\mathbf{Y}, \bar{\mathbf{X}}_n) \rightarrow_p \mathbf{p}(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})$.

To now show that the point estimators in Table 3.3 are asymptotically normal, given that $p_{k,n}(\mathbf{Y}, \bar{\mathbf{X}}_n) \rightarrow_p p_k(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})\ \forall k$, we will apply Slutsky's Theorem (Theorem 2.4).

Consider, for example, deriving the asymptotic distribution of $\bar{p}_{k,n_s,n}(\mathbf{Y}, \bar{\mathbf{X}}_n)$. Let

$$\begin{aligned} Z_{k,n_s,n}(\bar{\mathbf{X}}_n) &= \frac{\bar{p}_{k,n_s,n}(\bar{\mathbf{X}}_n) - \mu_{p_k}}{\frac{\sigma_{p_k}}{\sqrt{n_s}}}, \text{ and} \\ Z_{k,n_s}(\boldsymbol{\mu}_{\mathbf{X}}) &= \frac{\bar{p}_{k,n_s}(\boldsymbol{\mu}_{\mathbf{X}}) - \mu_{p_k}}{\frac{\sigma_{p_k}}{\sqrt{n_s}}} \end{aligned}$$

We may write

$$Z_{k,n_s,n}(\bar{\mathbf{X}}_n) = Z_{k,n_s}(\boldsymbol{\mu}_{\mathbf{X}}) + \left(Z_{k,n_s,n}(\bar{\mathbf{X}}_n) - Z_{k,n_s}(\boldsymbol{\mu}_{\mathbf{X}})\right). \tag{3.14}$$

By Equation 3.8,

$$Z_{k,n_s}(\boldsymbol{\mu}_{\mathbf{X}}) \rightarrow_d \mathcal{N}(0,1).$$

Also, for n_s fixed,

$$\begin{aligned} Z_{k,n_s,n}(\bar{\mathbf{X}}_n) &\rightarrow_p Z_{k,n_s}(\boldsymbol{\mu}_{\mathbf{X}}) \text{ or} \\ Z_{k,n_s,n}(\bar{\mathbf{X}}_n) - Z_{k,n_s}(\boldsymbol{\mu}_{\mathbf{X}}) &\rightarrow_p 0, \end{aligned}$$

by Theorem 3.1 since it was shown that $p_{k,n}(\mathbf{Y}, \bar{\mathbf{X}}_n) \rightarrow_p p_k(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})\ \forall k$ and $Z(\cdot)$ is a continuous function of $p_{k,n}(\mathbf{Y}, \bar{\mathbf{X}}_n)$. Now take $n_s \rightarrow \infty$ and apply Slutsky's Theorem to the right hand side of Equation 3.14. We have

$$Z_{k,n_s,n}(\bar{\mathbf{X}}_n) = \frac{\bar{p}_{k,n_s,n}(\bar{\mathbf{X}}_n) - \mu_{p_k}}{\frac{\sigma_{p_k}}{\sqrt{n_s}}} \rightarrow_d \mathcal{N}(0,1).$$

It is straightforward to show that

$$s^2_{p_k}(\bar{\mathbf{X}}) = \frac{1}{n_s - 1} \sum_{i=1}^{n_s} \left(p_{k,n}(\mathbf{Y}_i, \bar{\mathbf{X}}) - \bar{p}_{k,n_s,n}(\bar{\mathbf{X}})\right)^2 \rightarrow_p \sigma^2_{p_k}.$$

Then from the arguments in Subsection 2.5.2,

$$\frac{\bar{p}_{k,n_s,n}(\bar{\mathbf{X}}) - \mu_{p_k}}{\frac{s_{p_k}(\bar{\mathbf{X}})}{\sqrt{n_s}}} \rightarrow_d \mathcal{N}(0,1). \tag{3.15}$$

Using the asymptotic normality of $\hat{\eta}_{k,n_s}(\boldsymbol{\mu}_{\mathbf{X}})$, $\hat{\lambda}_{k,n_s}(\boldsymbol{\mu}_{\mathbf{X}})$, $\hat{\lambda}^s_{k,n_s}(\boldsymbol{\mu}_{\mathbf{X}})$ and $\mathrm{m}_{p_k,n_s}(\boldsymbol{\mu}_{\mathbf{X}})$ (by Theorems 2.2-2.3) as well as their continuity in $\bar{\mathbf{X}}$ (that is, when $\boldsymbol{\mu}_{\mathbf{X}}$ is replaced by $\bar{\mathbf{X}}$), we can derive similar results for these point estimators as functions of $\bar{\mathbf{X}}$. As discussed in Subsection 2.5.2, when n_s and n_k are large we will prefer the simpler result in Equation 3.15.

3.6 Another Diet Estimation Method: Maximum Likelihood Estimation

We have considered a maximum likelihood estimation (MLE) approach to the diet estimation problem but have not it to be useful. In this section we outline the algorithm and the various problems that we encountered.

Aitchison and Bacon-Shone (1999) offer a general method of estimating mixtures of compositions. Since the seal FA signature may be regarded as a mixture of the prey signatures, their method can, in theory, be applied to the diet estimation problem. We will refer to their method as the MLE algorithm.

In Aitchison and Bacon-Shone, the problem of interest is described as determining the distribution of the D-part composition, $\mathbf{Y}$, formed as a convex linear combination

$$\mathbf{Y} = \mathrm{cvx}(\mathbf{X}, \boldsymbol{\pi}) = \pi_1 \mathbf{X}_1 + \ldots + \pi_I \mathbf{X}_I,$$

where $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_I]$ is a set of I independent D-part compositions with known distributions and where $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_I]$ is a vector of nonnegative mixing proportions.

In our case $\mathbf{Y}$ is the seal FA signature, $\mathbf{X}_k$ is the FA signature of the kth species and D corresponds to the number of FAs, n_{FA}. The mixing proportions represent the true common diet of the seals in a given region, from which the sample of n_s seals was drawn. In accordance with Aitchison and Bacon-Shone's terminology, the seal FA signatures are the *target* compositions while the prey FA signatures are the *source* compositions.

Before presenting various models for $\mathbf{Y}$, a few comments concerning the notation to be used in this section is needed. Recall from Chapter 2 that if $\mathbf{X} \sim \mathcal{L}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

for some D-part composition $\mathbf{X}$, then $\log\left(\frac{\mathbf{X}_{-D}}{X_D}\right) \sim \mathcal{N}^d(\mu, \Sigma)$. Aitchison and Bacon-Shone (1999) have chosen to re-parameterize through a one-to-one transformation. They use instead the notation $\mathbf{X} \sim \mathcal{L}^D(\xi, \mathbf{T})$ where the components of ξ and $\mathbf{T}$ are given by

$$\begin{aligned} \xi_i &= \frac{e^{\mu_i}}{e^{\mu_1} + \cdots + e^{\mu_d} + 1}, \; i = 1, \ldots, d, \\ \xi_D &= \frac{1}{e^{\mu_1} + \cdots + e^{\mu_d} + 1}, \\ \tau_{ij} &= \sigma_i^2 + \sigma_j^2 - 2\sigma_{ij}, \; i, j = 1, \ldots D, \end{aligned}$$

and where $\sigma_{Dj} = \sigma_{iD} = \sigma_{DD} = 0$. For the remainder of this Section, we have adopted their notation with $D = n_{FA}$.

Aitchison and Bacon-Shone suggest using one of the following three approximate distributions to model $\mathbf{Y}$:

Fixed-Mixture Model

The distribution of $\mathbf{Y} = \text{cvx}(\mathbf{X}, \pi) = \pi_1\mathbf{X}_1 + \ldots + \pi_I\mathbf{X}_I$, where $\mathbf{X}_1, \ldots, \mathbf{X}_I$ are independently distributed as $\mathcal{L}^{n_{FA}}(\xi_1, \mathbf{T}_1), \ldots, \mathcal{L}^{n_{FA}}(\xi_I, \mathbf{T}_I)$, is approximately $\mathcal{L}^{n_{FA}}(\eta, \Theta)$, $\Theta = [\theta_{ij}]$ and

$$\eta = \sum_{k=1}^{I} \pi_k \xi_k, \; \theta_{ij} = -\frac{1}{2} \sum_{k=1}^{I} \sum_{b=1}^{n_{FA}} \sum_{l=1}^{n_{FA}} G_{kijb} G_{kijl} \tau_{kbl},$$

where

$$G_{kijb} = \rho_{ki}(\delta_{ib} - \xi_{kb}) - \rho_{kj}(\delta_{jb} - \xi_{kb}), \rho_{ki} = \frac{\pi_k \xi_{ki}}{\eta_i},$$

and

$$\delta_{ib} = \begin{cases} 1 & \text{if } b = i, \\ 0 & \text{if } b \neq i. \end{cases}$$

The Fixed-Mixture model has $I-1$ unknown parameters since ξ_k and $\mathbf{T}_k$, $k = 1, \ldots, I$ are assumed to be known.

Convolution Model

This model is similar to the Fixed-Mixture model but now it is assumed that π is distributed as $\mathcal{L}^{n_{FA}}(\alpha, \Omega)$. Then $\mathbf{Y}$ is approximately $\mathcal{L}^{n_{FA}}(\kappa, \Lambda)$ where

$$\kappa = \sum_{k=1}^{I} \alpha_k \xi_k, \ \lambda_{ij} = -\frac{1}{2} \sum_{k=1}^{I} \sum_{b=1}^{n_{FA}} \sum_{l=1}^{n_{FA}} H_{kijb} H_{kijl} \tau_{kbl} - \frac{1}{2} \sum_{a=1}^{I} \sum_{k=1}^{I} B_{aij} B_{kij} \omega_{ak},$$

and

$$H_{kijb} = \chi_{ki}(\delta_{ib} - \xi_{kb}) - \chi_{kj}(\delta_{jb} - \xi_{kb}), \ \chi_{ki} = \frac{\alpha_k \xi_{ki}}{\kappa_i}, \ B_{kij} = \chi_{ki} - \chi_{kj}.$$

The Fixed-Mixture model is the special case where $\Omega = \mathbf{0}$ and $\alpha = \pi$. This model contains $I - 1 + \frac{1}{2}I(I-1)$ unknown parameters.

Perturbation Model

In the perturbation model, the distribution of $\mathbf{Y} = \text{cvx}(\mathbf{X}, \pi) \circ \mathbf{U}$ is approximated. If $\mathbf{U} \sim \mathcal{L}^{n_{FA}}(\mathbf{e}, \Psi)$, where $\mathbf{e} = [\frac{1}{D}, \ldots, \frac{1}{D}]$ is the identity of the perturbation group, then the distribution of $\mathbf{Y}$ is approximately $\mathcal{L}(\eta, \Theta + \Psi)$ distributed (by Property 2.2) where η and Θ are given in the Fixed-Mixture model. Note that the Fixed-Mixture model is the special case where $\Psi = \mathbf{0}$. There are $I - 1 + \frac{1}{2}n_{FA}(n_{FA} - 1)$ unknown parameters in the Perturbation model.

Aitchison and Bacon-Shone explain that the more complicated Convolution and Perturbation models arise because often the target samples have total measures of variation similar to or greater than the source measures. Their simulation studies have shown, however, that the measure of total variation for the convex linear combination $\mathbf{Y}$ tends to be less.

As a first step to using Aitchison and Bacon-Shone's method to estimate the diet, π, a model must be chosen. While it is possible to carry out an approximate likelihood ratio test (LRT) as suggested in Aitchison and Bacon-Shone (1999), for the diet estimation problem, n_s will often be small (and much less than n_{FA}) and the test may not be valid. In the real-life example used in Aitchison and Bacon-Shone (1999), three separate target samples were analyzed and the LRT yielded a different model for all three. Since time will be a factor in our simulation studies, we will choose the Fixed-Mixture model as there will be considerably fewer parameters to estimate.

For example, if we use $I = 8$ (I will actually be much larger in practice but is the number of species used in our simulations) and $n_{FA} = 40$, the number of parameters that require estimating for each of the models are given in Table 3.4.

Model	No. Parameters
Fixed-Mixture	7
Convolution	35
Perturbation	787

Table 3.4: For each model, the number of unknown parameters for $I = 8$ and $n_{FA} = 40$.

Point estimation of π is by ML estimation. For the Fixed-Mixture model, the likelihood function is

$$L(\pi) = \prod_{i=1}^{n_s} \frac{1}{(2\pi)^{-d/2}|\Sigma(\pi)|^{-\frac{1}{2}}(y_{1i}\cdots y_{n_{FA}i})} \times e^{-\frac{1}{2}\left[\log\left(\frac{y_{i,-n_{FA}}}{y_{i,n_{FA}}}\right)-\mu(\pi)\right]'\Sigma^{-1}(\pi)\left[\log\left(\frac{y_{i,-n_{FA}}}{y_{i,n_{FA}}}\right)-\mu(\pi)\right]}.$$

where $\mu(\pi) = \log\left(\frac{\eta_{-n_{FA}}}{\eta_{n_{FA}}}\right)$ and $\sigma_{ij} = \frac{1}{2}(\theta_{in_{FA}} + \theta_{jn_{FA}} - \theta_{ij})$, $i, j = 1, \ldots, n_{FA}$. Recall that η and Θ are functions of π. Also, when calibration coefficients are needed, we assume that the sample of seals, $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_s}$, have been calibrated before $L(\pi)$ is evaluated.

It should be mentioned that Aitchison and Bacon-Shone (1999) actually recommend an initial investigation into whether an additive logistic skew-normal distribution ($\mathcal{LS}$) approximation is worthwhile. In their examples, however, they found that $\mathcal{LS}$ modeling did not yield significantly better fits than the $\mathcal{L}$ models.

When the MLE algorithm (with Fixed-Mixture model) was applied to the diet estimation problem, a few problems arose. Firstly, even for n_s small, the MLEs took several hours to calculate with S-PLUS's *nlminb* function. This was due mainly to the slowness in computing Θ with $n_{FA} = 40$. A second problem, also related to the dimension of the FA signatures, was that *nlminb* did not always find the maximum of $\log L(\pi)$ when calibration factors were used. This phenomenon is illustrated in Tables 3.5 and 3.6 where different starting values yield different estimates of the diet when the algorithm is applied to 1000 pseudo-seals. Note that we have used the notation

$\hat{\pi}$ to denote *nlminb*'s estimate of π, but with $n_s = 1000$, $\hat{\pi}$ may be considered to be a parameter. If $\hat{\pi}$ is actually the MLE and the underlying model appropriate, then with $n_s = 1000$, it should be the case that $\hat{\pi} \approx \pi$, the true diet. With $n_s = 1000$, we may therefore compare $\hat{\pi}$ to any of the parameters discussed in Section 3.3 such as $\mu_{\mathbf{p}} = \mathrm{E}_{\mathbf{Y}}[\mathrm{p}(\mathbf{Y}, \mu_{\mathbf{X}})]$.

For Diet 1 (Table 3.5), although $\log L(\hat{\pi})$ is the largest when equal proportions are used as the starting values, the results are not sensible. For Diet 4 (Table 3.6), the starting values $\mu_{\mathbf{p}}$ and Diet 4 give the same $\hat{\pi}$ and $\log L(\hat{\pi})$ is largest for these starting values. With these starting values and except for Sandlance ($\pi_k = 0.45$), $\hat{\pi}$ is a poor parameterization of the diet. To verify that this difficulty in finding the maximum of $\log L(\pi)$ was actually related to the dimension of the FA signatures, the optimization was carried out for three randomly selected FAs using the three sets of starting values. For $n_{FA} = 3$, all three sets of starting values gave the same results suggesting that the maximum was found in this simplified case.

Without calibration factors (see Tables 3.7 and 3.8) the only problem was the slowness of the algorithm as $\hat{\pi}$ was the same for all three sets of starting values and appears to be a reasonable parameterization of the diet. Note that while the components of $\mu_{\mathbf{p}}$ (computed without calibration) are usually slightly closer to π than are the components of $\hat{\pi}$, the Convolution or Perturbation models, or the use of the $\mathcal{SL}$ distribution may have given improved results.

Although without calibration factors the MLE algorithm gave adequate results, calibration effects will be present in the diet estimation problem. The poor results obtained with the calibration factors combined with the challenge of the high dimension of the FA signatures have led us to conclude that the MLE algorithm is not particularly useful for our application.

			COD	HAD	PLC	POL	SAND	SH	WF	YT
			0.3	0.3	0	0.15	0	0.15	0	0
Start	log L($\hat{\pi}$)	μ_p	0.283	0.245	0.023	0.224	0.045	0.134	0.017	0.029
Diet 1	93472	$\hat{\pi}$	0	1	0	0	0	0	0	0
$[\frac{1}{7},\ldots,\frac{1}{7}]$	105447	$\hat{\pi}$	0.481	0	0.333	0	0	0	0	0.186
μ_p	95888	$\hat{\pi}$	0.122	0.878	0	0	0	0	0	0

Table 3.5: Diet 1, Calibration: MLE algorithm estimates ($\hat{\pi}$) and $\log L(\hat{\pi})$ for various starting values in *nlminb*. (Based on 1000 pseudo-seals with $\epsilon = 10\%$ and $n^p = 30$.)

			COD	HAD	PLC	POL	SAND	SH	WF	YT
			0.09	0	0.09	0	0.45	0.09	0.09	0.09
Start	log L(π)	μ_p	0.104	0.043	0.069	0.155	0.369	0.117	0.074	0.069
Diet 4	104315	$\hat{\pi}$	0.363	0	0	0	0.489	0	0	0.147
$[\frac{1}{7},\ldots,\frac{1}{7}]$	95325	$\hat{\pi}$	0.329	0	0.339	0	0.146	0	0	0.186
μ_p	104315	$\hat{\pi}$	0.363	0	0	0	0.489	0	0	0.147

Table 3.6: Diet 4, Calibration: MLE algorithm estimates ($\hat{\pi}$) and $\log L(\hat{\pi})$ for various starting values in *nlminb*. (Based on 1000 pseudo-seals with $\epsilon = 10\%$ and $n^p = 30$).

Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
Diet 1	0.3	0.3	0	0.15	0	0.15	0	0
μ_p	0.277	0.307	0.025	0.153	0.045	0.151	0.018	0.024
$\hat{\pi}$	0.279	0.194	0.026	0.196	0.079	0.151	0.065	0.011

Table 3.7: Diet 1, No calibration: MLE algorithm estimates ($\hat{\pi}$). (Based on 1000 pseudo-seals with $\epsilon = 10\%$ and $n^p = 30$.)

Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
Diet 4	0.09	0	0.09	0	0.45	0.09	0.09	0.09
μ_p	0.081	0.058	0.061	0.073	0.442	0.117	0.086	0.083
$\hat{\pi}$	0.127	0.095	0.047	0.127	0.354	0.077	0.101	0.071

Table 3.8: Diet 4, No calibration: MLE algorithm estimates ($\hat{\pi}$). (Based on 1000 pseudo-seals with $\epsilon = 10\%$ and $n^p = 30$.)

Chapter 4

Diet Interval Estimation

In Chapter 3 the DMA (Distance Minimization Algorithm) for estimating the diet of a single seal was introduced. It was then determined that certain MOLs (Measures of Location) of the diet estimates were sufficiently close to the true diet making it possible to parameterize the diet estimation problem. In this chapter, we consider interval estimation for the true diet of a seal (or common diet of a group of seals). A useful approach is to derive an interval for one of the MOLs and then to adjust the interval by the (estimated) difference between the MOL and the true diet.

The core of this chapter consists of the CI methods and a simulation study designed to compare the coverage probabilities and interval lengths of the various methods. Our recommended CI method is then applied to some real-life data.

4.1 Resampling Techniques

Most of our CIs methods rely on bootstrap methods in which we resample from our original data (the seal and prey FA signatures) to generate replicate samples. Appendix C describes a few resampling techniques including Davison and Hinkley's (1997) nonparametric and parametric bootstrap procedures, applied to the seal and prey FA signatures. Note that the ability to generate pseudo-seals allows for the seal FA signatures to be resampled in a nontraditional manner. For example, in the "pseudo-seal method" (detailed in Appendix C), we generate n_s pseudo-seals by letting the ith pseudo-seal have diet $\mathbf{p}_i(\mathbf{Y}_i, \bar{\mathbf{X}})$.

In the CI bootstrap algorithms outlined in the following subsections, we allow for the use of any of the resampling procedures given in Appendix C by simply indicating to "generate" a sample of seals or prey. In our simulation study (discussed in Section 4.4), we chose to use a nonparametric bootstrap of the prey and to generate pseudo-seals by the "pseudo-seal method".

4.2 Bias Adjustment

All of the CI methods to be presented require an adjustment to account for the difference between the parameter being estimated and the true diet, π, the parameter ultimately of interest. Recall from the discussions in Section 3.3 that there are various parameters that are close to the true diet. The difference between these parameters and the true diet was termed the bias even though it is only for the parameter $\mu_{\mathbf{p}} = \mathrm{E}_{\mathbf{Y}}[\mathbf{p}(\mathbf{Y}, \mu_{\mathbf{X}})]$ that the traditional definition of bias applies. Clearly CIs for these parameters are not necessarily CIs for π. Most of our CI methods rely on bootstrapping and a consequence is that even our nonparametric intervals (which are bootstrap based) require an adjustment. The reason is that bootstrap techniques (both parametric and nonparametric) generally assume that the point estimator can be expressed as $t(\hat{F})$, where $\hat{F}$ is the empirical distribution function (edf), and that the parameter of interest is $t(F)$, where F is the cumulative distribution function (cdf). Consider our simplest point estimator, $\bar{\mathbf{p}}(\bar{\mathbf{X}})$. We may write

$$\bar{\mathbf{p}}(\bar{\mathbf{X}}) = \frac{1}{n_s}\sum_{i=1}^{n_s}\mathbf{p}(\mathbf{Y}_i, \bar{\mathbf{X}}) = \int\left[\mathbf{p}\left(\mathbf{Y}, \int \mathbf{X}_1 d\hat{F}(\mathbf{X}_1), \ldots, \int \mathbf{X}_I d\hat{F}\mathbf{X}_I\right)\right] d\hat{F}_{\mathbf{Y}}.$$

Then the actual parameter that is being estimated is $\mu_{\mathbf{p}} = \mathrm{E}_{\mathbf{Y}}[\mathbf{p}(\mathbf{Y}, \mu_{\mathbf{X}})]$ and not π, the true diet. An algorithm is therefore needed to estimate the difference between $t(F)$ and π. The CIs are then shifted by the estimated bias.

Our algorithm for estimating this bias involves generating pseudo-seals from each of the n_s estimates of diet, computing (for each pseudo-seal) the parameter that our interval method is estimating ($t(F)$), and examining (for each pseudo-seal) the difference between our estimate and the parameter. The n_s estimates of bias are then summarized. The algorithm, in detail, is as follows:

Bias Estimation Algorithm

1. Choose the parameter for which a bias adjustment is required, for example, $\mu_{\mathbf{p}} = \mathrm{E}_{\mathbf{Y}}[\mathbf{p}(\mathbf{Y}, \mu_{\mathbf{X}})]$.

2. Compute n_s diet estimates: $\mathbf{p}_1(\mathbf{Y}_1, \bar{\mathbf{X}}), \ldots, \mathbf{p}_{n_s}(\mathbf{Y}_{n_s}, \bar{\mathbf{X}})$.

3. Compute a point estimate, say $\bar{\mathbf{p}}(\bar{\mathbf{X}})$, using diet estimates in 2.

4. Generate R_{ps} pseudo-seals from each of the diet estimates in 2.: $\mathbf{Y}_1^{*i}, \ldots, \mathbf{Y}_{R_{ps}}^{*i}$, $i = 1, \ldots, n_s$.

5. For the ith seal, compute the R_{ps} diet estimates: $\mathbf{p}_1^*(\mathbf{Y}_1^{*i}, \bar{\mathbf{X}}), \ldots, \mathbf{p}_{R_{ps}}^*(\mathbf{Y}_{R_{ps}}^{*i}, \bar{\mathbf{X}})$.

6. For the ith seal compute $\boldsymbol{\mu}_{\mathbf{p}^*}^i = \frac{1}{R_{ps}} \sum_{r=1}^{R_{ps}} \mathbf{p}_r^*(\mathbf{Y}_r^{*i}, \bar{\mathbf{X}})$.

7. For the ith seal, compute

$$\hat{b}_i = \boldsymbol{\mu}_{\mathbf{p}^*}^i - \mathbf{p}_i(\mathbf{Y}_i, \bar{\mathbf{X}}).$$

8. The bias estimate is then an average of $\hat{b}_i$, $i = 1, \ldots, n_s$.

Note that in Step 8 we chose to use the median of the $\hat{b}_i$ as our measure of the average bias. Also, it should be mentioned that we examined variations of this algorithm such as generating R_{ps} pseudo-seals from a point estimator of diet instead of from the individual diet estimates. While the performance of the bias algorithms varied with the parameter being estimated and the CI method, the above algorithm (using the median of the $\hat{b}_i$) was chosen as it appeared to perform adequately across the various parameters and CI methods.

Finally, it should be mentioned that except for the parameter, $\mathrm{E}_{\mathbf{Y}}[\mathbf{p}(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})]$, the sum of the parameter components over the species is not necessarily one. (This was discussed in Section 2.4.) Our CIs will be for the non-normalized parameter. In our bias adjustment algorithm, we do not normalize our point estimates and so this source of bias is essentially incorporated into our bias estimate. To assess our bias adjustment algorithm, in Section 4.4 we present our results with and without the bias adjustment. The results without the bias adjustment, however, have actually been adjusted to account for this normalization related bias. That is, for intervals that are not bias corrected, we divided the confidence limits by $\sum_{k=1}^{I} \hat{\pi}_k$, where $\hat{\pi}_k$ may represent any of our estimates.

4.3 CI Methods

4.3.1 Overview

We have divided our CI methods into four subsections: Large Sample Intervals, Parametric Intervals, Semi-Parametric Intervals and Nonparametric Intervals. The intent of this subsection is to motivate the CI methods and to provide some insight into their potential advantages and disadvantages.

The large sample intervals include a variety of intervals, and, as the name suggests, are expected to perform well when n_s (sample size of seals) and/or the n_k (sample size of prey from species k) are large. They include intervals that use the asymptotic normality of the point estimators, $\bar{p}_k$, $\hat{\eta}_k$, $\hat{\lambda}_k$ and m_k derived in Section 3.5 when both n_s and n_k are large as well as others that may be appropriate when n_s only is large. Of the latter proposed methods, one method assumes that the point estimators are approximately normally distributed, as in the asymptotic intervals, but attempts to incorporate the variability due to the prey using bootstrap methods. Because the large sample intervals are perhaps the simplest to implement and some do not require bootstrapping, we are interested in comparing their performance with some of the more complex methods and determining whether, for larger values of n_s, these simple intervals may suffice.

The parametric and semi-parametric methods were developed for the case when both n_s and the n_k are small or of moderate size. They are essentially modified versions of the finite sample intervals derived in Subsection 2.5.2 for compositional data modeled by one of the discussed parametric distributions. For the diet estimation application, the need for adjustments to the intervals in Subsection 2.5.2 is two-fold. Firstly, recall that these intervals involved certain nuisance parameters. For n_s small, simply substituting the MLEs in for these parameters is not a satisfactory approach. A second issue, specific to the diet estimation application, is the lack of independence in the diet estimates. Consequently, the intervals in Subsection 2.5.2 may only be applied to the diet estimates conditional on $\bar{\mathbf{X}}$ and, for n_k small, should be altered to reflect the variability in the prey. We have used bootstrap methods to manage these issues.

An obvious drawback of the parametric intervals is that their success depends on how well the diet estimates are modeled by the specific parametric distributions. If the parametric model is appropriate then the CIs should be most efficient. Although the semi-parametric CI method is more complex than the parametric methods, it does not rely on an underlying parametric density for the diet estimates. A further advantage of the semi-parametric intervals is that they make use of the populations, $\mathbf{V}_j, j = 1, \ldots, B$ (recall that $\mathbf{V}_j$ contains the indices of the non-zero components of the estimate) and therefore use some information from all of the components of the diet estimates into the univariate intervals. A disadvantage, is that unlike some of the large sample bootstrap and parametric intervals, when $n_s = 1$, the semi-parametric intervals are not computable.

Some nonparametric methods were implemented as well, including percentile based methods. In addition to the obvious advantage of not requiring an assumed underlying parametric model, these intervals are plausible for $n_s \geq 1$ and are less complex than the semi-parametric and parametric intervals.

4.3.2 Large Sample Intervals

Our discussion on large sample intervals is split into two cases based on whether n_k is considered to be small or large. In both cases, however, n_s is assumed to be "large", where "large" will be determined by our simulation study in Section 4.4.

Case 1: n_s and n_k, $k = 1, \ldots,$I, are large

When both n_s and the n_k are assumed to be large, CIs for the MOLs given in Table 3.3 follow almost immediately from the asymptotic results given in Section 3.5. We have the following asymptotic intervals for μ_{p_k}, η_k, λ_k, and λ_k^s respectively:

$$\bar{p}_k(\bar{\mathbf{X}}) \pm z_{\frac{\alpha}{2}} \frac{s_{p_k}(\bar{\mathbf{X}})}{\sqrt{n_s}}, \tag{4.1}$$

$$\hat{\eta}_k(\bar{\mathbf{X}}) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{I_{\hat{\eta}_k}(\hat{\theta}_{\mathbf{v}_j}, \hat{\mu}_k^{\mathbf{v}_j}(\bar{\mathbf{X}}), \hat{\sigma}_k^{2\mathbf{v}_j}(\bar{\mathbf{X}}))}{n_s}}, \tag{4.2}$$

$$\hat{\lambda}_k(\bar{\mathbf{X}}) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{I_{\hat{\lambda}_k}(\hat{\theta}_k, \hat{\mu}_k(\bar{\mathbf{X}}), \hat{\sigma}_k^2(\bar{\mathbf{X}}))}{n_s}}, \tag{4.3}$$

$$\hat{\lambda}_k^s(\bar{\mathbf{X}}) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{I_{\hat{\lambda}_k^s}(\hat{\theta}_k, \hat{\mu}_k(\bar{\mathbf{X}}), \hat{\sigma}_k^2(\bar{\mathbf{X}}), \hat{\alpha}_k(\bar{\mathbf{X}}))}{n_s}}, \tag{4.4}$$

where $I_{\hat{\eta}_k}$, $I_{\hat{\lambda}_k}$, and $I_{\hat{\lambda}_k^s}$ denote the corresponding Fisher information. Since all four of the above intervals are valid for n_s and n_k large, we will prefer to use the simplest intervals, namely those given by Equation 4.1, provided we can accurately estimate the difference between μ_{p_k} and π_k. We have therefore not derived the Fisher information for the other estimators.

Note that for n_s only moderately large using quantiles of the t distribution or the finite sample intervals presented in Subsection 2.5.2 (derived from our parametric models for compositional data), may yield improved results over a normal approximation. (Application of the finite sample interval methods of Subsection 2.5.2 to the diet estimates is discussed in detail in Subsections 4.3.3 and 4.3.4.) We have therefore computed these intervals as well and consider them to be alternative large sample (Case 1) interval methods.

Observe also that not all of the sample sizes of prey in the prey base used in our simulation study would likely be considered large relative to the dimension of the prey FA signatures (see Appendix A for the sample sizes). We will still, however, compute the above intervals to compare their coverage probability and length with the more computationally intensive methods described in the remainder of this chapter.

Case 2: n_s is large

As mentioned above, often the chosen prey base will not be considered large enough to assume that $\boldsymbol{\mu}_{\mathbf{X}} \approx \bar{\mathbf{X}}$. When n_s only is large, Equations 4.1-4.4 give approximate CIs for $\mu_{p_k|\bar{\mathbf{X}}}$, $\eta_{k|\bar{\mathbf{X}}}$, and $\lambda_{k|\bar{\mathbf{X}}}$ respectively. To obtain CIs for the unconditional parameters, our approach involves adjusting the variability of the point estimators using the following bootstrap procedure:

Large Sample Bootstrap Algorithm

1. Compute n_s diet estimates: $\mathbf{p}_1(\mathbf{Y}_1, \bar{\mathbf{X}}), \ldots, \mathbf{p}_{n_s}(\mathbf{Y}_{n_s}, \bar{\mathbf{X}})$.

2. Compute a point estimate, say $\bar{\mathbf{p}}(\bar{\mathbf{X}})$, using diet estimates in 1.

3. *for* $r = 1, \ldots, R$

   (a) Generate n_s seals: $\mathbf{Y}_1^{*r}, \ldots, \mathbf{Y}_{n_s}^{*r}$. (See Appendix C.)

   (b) Generate sample of prey: $\mathbf{X}_1^{*r}, \ldots, \mathbf{X}_I^{*r}$. (See Appendix C.)

   (c) Compute n_s diet estimates: $\mathbf{p}_1^*(\mathbf{Y}_1^{*r}, \bar{\mathbf{X}}^{*r}), \ldots, \mathbf{p}_{n_s}^*(\mathbf{Y}_{n_s}^{*r}, \bar{\mathbf{X}}^{*r})$.

   (d) Compute the point estimate, $\bar{\mathbf{p}}^{*r}(\bar{\mathbf{X}}^{*r})$, using diet estimates in (c).

4. Compute

$$\sigma_{\bar{p}_k}^{2\text{boot}} = \frac{1}{R}\sum_{r=1}^{R}\left(\bar{p}_k^{*r}(\bar{\mathbf{X}}^{*r}) - \overline{\bar{p}^*}_k(\bar{\mathbf{X}}^{*r})\right)^2,$$

   where $\overline{\bar{p}^*}_k(\bar{\mathbf{X}}^{*r}) = \frac{1}{R}\sum_{r=1}^{R}\bar{p}_k^{*r}(\bar{\mathbf{X}}^{*r})$.

Observe that if $n_s = 1$, the above bootstrap algorithm yields an estimate of the variance in the diet estimates themselves, $\text{VAR}_{\mathbf{Y},\bar{\mathbf{X}}}[\mathbf{p}(\mathbf{Y}, \bar{\mathbf{X}})]$. Although similar to the bootstrap estimate of variance discussed in Iverson *et al* (2004) (and in Subsection 3.2), our proposed estimate includes the variability due to the seal FAs and not just the prey. Incorporating the variability due to the seals when $n_s = 1$ is only possible because we can generate pseudo seals (see Method 1 in Appendix C) and do not have to rely on traditional parametric or nonparametric bootstrap methods.

An approximate $100(1-\alpha)\%$ CI for μ_k is then

$$\bar{p}_k(\bar{\mathbf{X}}) \pm z_{\frac{\alpha}{2}}\sigma_{\bar{p}_k}^{\text{boot}}, \tag{4.5}$$

and similarly for η_k, λ_k and M_k. We have also examined the effect of modifying the intervals slightly by the use of a variance stabilizing transformation, namely the binomial distribution variance stabilizing transformation:

$$h(p) = \frac{2}{\pi}\arcsin(\sqrt{p}).$$

To compute these intervals, the above transformation is applied to $\bar{p}(\bar{\mathbf{X}})$ in Step 2 and to $\bar{p}^{*r}(\bar{\mathbf{X}})$ in Step (d). An approximate $100(1-\alpha)\%$ CI for $\frac{2}{\pi}\arcsin(\sqrt{\mu_{p_k}})$, is then given by

$$\bar{p}_k^T(\bar{\mathbf{X}}) \pm z_{\frac{\alpha}{2}}\sigma_{\bar{p}_k^T(\bar{\mathbf{X}})}^{\text{boot}} \tag{4.6}$$

where $\bar{p}_k^T(\bar{\mathbf{X}}) = \frac{2}{\pi}\arcsin(\sqrt{\bar{p}_k(\bar{\mathbf{X}})})$. An approximate CI for μ_{p_k} then follows by applying the inverse of the transformation to the confidence limits.

Realize that although the $\bar{p}_k^{*T}$ are themselves compositions, we have not attempted transformations involving logarithms (such as those proposed by Aitchison (1986) and discussed in Chapter 2) due to the zeros that will often be present. CI methods that make use of the discussions in Chapter 2 are given in Subsections 4.3.3 and 4.3.4.

4.3.3 Parametric Intervals

In Section 3.4 it was shown that the diet estimates, $p_k(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})$, could be modeled reasonably well by the $Mix\mathcal{M}$, $SMix\mathcal{M}$ and $SMixS\mathcal{M}$ distributions. In Subsection 2.5.2 finite sample intervals for compositional data having these distributions were discussed. Recall from the Overview (Subsection 4.3.1) however, that these intervals are potentially problematic for the diet estimation application when n_s and n_k are small. We have accordingly developed more suitable, modified finite sample intervals for diet estimates modeled by each of these mixture distributions.

Before presenting the parametric interval methods, a comment concerning the absence of independence in the diet estimates is needed. We will actually model the estimates conditional on $\bar{\mathbf{X}}$ so that the marginals of $\mathbf{p}_{1|\bar{\mathbf{X}}}(\mathbf{Y}_1, \bar{\mathbf{X}}), \ldots, \mathbf{p}_{n_s|\bar{\mathbf{X}}}(\mathbf{Y}_{n_s}, \bar{\mathbf{X}})$ may be considered a random sample from one of the above mentioned mixture distributions. Unless n_k is large, a bootstrap of the prey will be incorporated into the interval methods to yield unconditional CIs, as in Case 2 of the previous subsection.

Suppose first that $p_{k,1}(\mathbf{Y}_1, \boldsymbol{\mu}_{\mathbf{X}}), \ldots, p_{k,n_s}(\mathbf{Y}_{n_s}, \boldsymbol{\mu}_{\mathbf{X}}) \sim Mix\mathcal{M}(\theta_{\mathbf{v}_j}, \mu_k^{\mathbf{v}_j}, \sigma_k^2)$. Then Subsection 2.5.2 described a method of obtaining a CI for η_k if the $\theta_{\mathbf{v}_j}$ are known. (If n_k is small, then $\bar{\mathbf{X}}$ is used in place of $\boldsymbol{\mu}_{\mathbf{X}}$ and the CI is actually for $\eta_{k|\bar{\mathbf{X}}}$.) Recall that this method involved pooling P-values computed for each of the populations into a single, overall P-value. One way uses the $\theta_{\mathbf{v}_j}$ as weights. In practice, particularly for n_s small, our method must allow for the uncertainty in the $\theta_{\mathbf{v}_j}$ since the populations may be poorly estimated by ML estimation. While we may use bootstrapping methods to obtain an improved estimate of the populations, the issue in doing so is that we will likely estimate populations for which we do not have any observations, making the test statistic in Equation 2.25 impractical. We have developed a semi-parametric

CI method, to be presented in Subsection 4.3.4, that utilizes the bootstrap estimate of the populations and a modified test statistic.

Now suppose that $p_{k,1}(\mathbf{Y}_1, \mu_{\mathbf{X}}), \ldots, p_{k,n_s}(\mathbf{Y}_{n_s}, \mu_{\mathbf{X}}) \sim SMixM(\theta_k, \mu_k, \sigma_k^2)$. Then from the discussions in Subsection 2.5.2, if θ_k is known and $\theta_k < 1$, a $100(1-\alpha)\%$ CI for λ_k (or for $\lambda_{k|\bar{\mathbf{X}}}$, if the n_k are small) is given by

$$\left[(1-\theta_k)\frac{e^{\hat{\mu}_k - t_{[(n'_{sk}-1),1-\alpha/2]}\frac{s_k}{\sqrt{n'_{sk}}}}}{1+e^{\hat{\mu}_k - t_{[(n'_{sk}-1),1-\alpha/2]}\frac{s_k}{\sqrt{n'_{sk}}}}}, (1-\theta_k)\frac{e^{\hat{\mu}_k + t_{[(n'_{sk}-1),1-\alpha/2]}\frac{s_k}{\sqrt{n'_{sk}}}}}{1+e^{\hat{\mu}_k + t_{[(n'_{sk}-1),1-\alpha/2]}\frac{s_k}{\sqrt{n'_{sk}}}}}\right], \tag{4.7}$$

where n'_{sk} denotes the number of non-zero estimates from species k in the sample. (If $\theta_k = 1$, a CI is unnecessary as $\lambda_k = 0$.) Recall that a problem with this CI is that when θ_k is large and n_s small, often all diet estimates for species k will be zero and we will not have an estimate of $\hat{\mu}_k$ and s_k to use in Equation 4.7. If θ_k is estimated by $\hat{\theta}_k = \frac{n_s - n'_{sk}}{n_s}$, then when all estimates are zero, $\hat{\theta}_k = 1$ which implies that $\lambda_k = 0$. Again for n_s small, presumably this method would not be effective. We have therefore developed an alternative method of obtaining CIs for estimates modeled by the $SMixM$ distribution. This method uses a bootstrap estimate of θ_k, which we surmise will be more useful than the MLE when n_s is small, and incorporates the variability of the bootstrap estimate into the intervals. We will refer to this method as the PAR interval method.

The PAR intervals will be determined by inverting the hypothesis test

$$\begin{aligned} \mathrm{H}_0 &: \lambda_k = \lambda_{k0} \\ \mathrm{H}_1 &: \lambda_k \neq \lambda_{k0}, \end{aligned} \tag{4.8}$$

using, as our test statistic, $T_k = |\hat{\lambda}_k(\bar{\mathbf{X}}) - \lambda_{k0}|$. As the finite sample distribution of T_k is unknown, we will compute a bootstrap P-value, say $p^{\text{boot}}(\lambda_{k0})$ for the test in Equation 4.8. The $100(1-\alpha)\%$ CI is then given by the set

$$\{\lambda_{k0} : p^{\text{boot}}(\lambda_{k0}) \geq \alpha\}.$$

The bootstrap P-value algorithm consists of an initial nonparametric bootstrap in which θ_k and σ_k^2 are estimated, followed by a parametric bootstrap in which estimates are generated from the null distribution. We now examine this bootstrap algorithm in more detail, beginning with the parametric bootstrap.

To simplify the discussion, consider first carrying out the hypothesis test in Equation 4.8 with both nuisance parameters, θ_k and σ_k^2, known. Realize that in a few cases, λ_{k0} will be such that a P-value is not needed. In particular, suppose that $\theta_k = 1$ then

$$\theta_k = 1 \Leftrightarrow \lambda_k = 0. \tag{4.9}$$

Further, note that

$$\begin{aligned} & \lambda_k = (1-\theta_k)\frac{e^{\mu_k}}{1+e^{\mu_k}} \text{ and } 0 < \frac{e^{\mu_k}}{1+e^{\mu_k}} < 1 \\ \Rightarrow \quad & \lambda_k < (1-\theta_k). \end{aligned} \tag{4.10}$$

As a result of Equations 4.9 and 4.10, a P-value is not needed if $\theta_k = 1$ or if $\lambda_{k0} \geq (1-\theta_k)$. It will be convenient, however, for $p^{\text{boot}}(\lambda_{k0})$ to be defined for all λ_{k0}. If, for example, $\lambda_{k0} \neq 0$ but $\theta_k = 1$ (or if $\lambda_{k0} = 0$ but $\theta_k < 1$), then we would like our algorithm to reject H_0 and we will accordingly set $p^{\text{boot}}(\lambda_{k0}) = 0$. If $\lambda_{k0} = 0$ and $\theta_k = 1$ then $p^{\text{boot}}(0)$ will be generated from $U[0,1]$. Similarly, if $\lambda_{k0} \geq (1-\theta_k)$ then our algorithm will set $p^{\text{boot}}(\lambda_{k0}) = 0$. For all other cases, $p^{\text{boot}}(\lambda_{k0})$ can be computed using a typical parametric bootstrap. That is, we will generate estimates under the null distribution, namely $SMix\mathcal{M}(\mu_{k0}, \sigma_k^2, \theta_k)$ where

$$\mu_{k0} = \log\left(\frac{\frac{\lambda_{k0}}{1-\theta_k}}{1-\frac{\lambda_{k0}}{1-\theta_k}}\right),$$

since $\lambda_{k0} = (1-\theta_k)\frac{e^{\mu_{k0}}}{1+e^{\mu_{k0}}}$. We then compare our observed test statistic to the test statistics computed with the bootstrap estimates.

In addition to the obvious issue of θ_k and σ_k^2 being unknown in practice, when the n_k are small, there is also an issue concerning the implied independence in the estimates generated under the null. To validate the independence, we will actually compute a P-value conditional on $\bar{\mathbf{X}}$. Through a nonparametric bootstrap of the prey we will compute various conditional P-values and use the average as $p^{\text{boot}}(\lambda_{k0})$. Regarding θ_k and σ_k being unknown in practice, we will simply estimate these parameters by generating many pseudo-seals, estimating their diets, and computing $\hat{\theta}_{k|\bar{\mathbf{X}}}$ and $s_{k|\bar{\mathbf{X}}}$ for this bootstrap sample. We will repeat the entire procedure a number of times by generating various samples of seals to account for the variability in the nuisance parameters due to the seals.

For species k, the following algorithm will be used to compute $p^{\text{boot}}(\lambda_{k0})$ for the hypothesis test in Equation 4.8:

PAR Algorithm

1. Compute n_s diet estimates: $p_{k,1}(\mathbf{Y}_1, \bar{\mathbf{X}}), \ldots, p_{k,n_s}(\mathbf{Y}_{n_s}, \bar{\mathbf{X}})$.

2. Compute $\hat{\lambda}_k$ and $T_k = |\hat{\lambda}_k - \lambda_{k0}|$.

3. *for* $r_s = 1{:}R_s$

   (a) Generate n_s seals: $\mathbf{Y}_1^{*r_s}, \ldots, \mathbf{Y}_{n_s}^{*r_s}$. (See Appendix C.)

      i. *for* $r_p = 1{:}R_p$

      ii. Generate sample of prey: $\mathbf{X}_1^{*r_p}, \ldots, \mathbf{X}_I^{*r_p}$. (See Appendix C.)

Initial Bootstrap

      iii. Compute n_s diet estimates: $p^*_{k,1}(\mathbf{Y}_1^{*r_s}, \bar{\mathbf{X}}^{*r_p}), \ldots, p^*_{k,n_s}(\mathbf{Y}_{n_s}^{*r_s}, \bar{\mathbf{X}}^{*r_p})$.

      iv. Generate $\frac{R_{ps}}{n_s}$ pseudo-seals from each of the diet estimates in iii: $\mathbf{Y}_1^{**(r_s,r_p)}, \ldots, \mathbf{Y}_{R_{ps}}^{**(r_s,r_p)}$.

      v. Compute R_{ps} diet estimates: $p^{**}_{k,1}(\mathbf{Y}_1^{**(r_s,r_p)}, \bar{\mathbf{X}}^{*r_p}), \ldots, p^{**}_{k,R_{ps}}(\mathbf{Y}_{R_{ps}}^{**(r_s,r_p)}, \bar{\mathbf{X}}^{*r_p})$.

      vi. Compute

$$\theta_{k|\bar{\mathbf{X}}^{*r_p}}^{r_s,\text{boot}} = \frac{R_{ps} - R'_{ps}}{R_{ps}}$$

$$s_{k|\bar{\mathbf{X}}^{*r_p}}^{r_s,\text{boot}} = \frac{1}{R'_{ps}} \sum_{r_{ps}=1}^{R'_{ps}} \left(\log \left(\frac{p^{**\prime}_{k,r_{ps}}(\mathbf{Y}_{r_{ps}}^{**(r_s,r_p)}, \bar{\mathbf{X}}^{*r_p})}{1 - p^{**\prime}_{k,r_{ps}}(\mathbf{Y}_{r_{ps}}^{**(r_s,r_p)}, \bar{\mathbf{X}}^{*r_p})} \right) - \mu_{k|\bar{\mathbf{X}}^{*r_p}}^{r_s,\text{boot}} \right),$$

      where $\mu_{k|\bar{\mathbf{X}}^{*r_p}}^{r_s,\text{boot}} = \frac{1}{R'_{ps}} \sum_{r_{ps}=1}^{R'_{ps}} \log \left(\frac{p^{**\prime}_{k,r_{ps}}(\mathbf{Y}_{r_{ps}}^{**(r_s,r_p)}, \bar{\mathbf{X}}^{*r_p})}{1 - p^{**\prime}_{k,r_{ps}}(\mathbf{Y}_{r_{ps}}^{**(r_s,r_p)}, \bar{\mathbf{X}}^{*r_p})} \right)$ and the prime notation corresponds to non-zero estimates.

Parametric Bootstrap

      vii. *if* $\lambda_{k0} = 0$

A. *if* $\theta_{k|\bar{\mathbf{X}}^{*r_p}}^{r_s,\text{boot}} = 1$, generate $p_{r_s}^{\text{boot}}(\lambda_{k0|\bar{\mathbf{X}}^{*r_p}})$ from $U[0,1]$. Go to Step i.

B. *else* set $p_{r_s}^{\text{boot}}(\lambda_{k0|\bar{\mathbf{X}}^{*r_p}}) = 0$. Go to Step i.

viii. *if* $\lambda_{k0} \geq (1 - \theta_{k|\bar{\mathbf{X}}^{*r_p}}^{r_s,\text{boot}})$, set $p_{r_s}^{\text{boot}}(\lambda_{k0|\bar{\mathbf{X}}^{*r_p}}) = 0$. Go to Step i.

ix. *for* r = 1:R

A. Generate n_s estimates from the null distribution, $SMixM(\theta_{k|\bar{\mathbf{X}}^{*r_p}}^{r_s,\text{boot}}, \mu_{k0}, s_{k|\bar{\mathbf{X}}^{*r_p}}^{r_s,\text{boot}})$, where $\mu_{k0} = \log\left(\frac{\frac{\lambda_{k0}}{1-\theta_{k|\bar{\mathbf{X}}^{*r_p}}^{r_s,\text{boot}}}}{1-\frac{\lambda_{k0}}{1-\theta_{k|\bar{\mathbf{X}}^{*r_p}}^{r_s,\text{boot}}}}\right)$.

B. Compute λ_k^{*r} from the estimates in A.

C. Compute $T_k^{*r} = |\lambda_k^{*r} - \lambda_{k0}|$.

x. Compute conditional bootstrap P-value

$$p_{r_s}^{\text{boot}}(\lambda_{k0|\bar{\mathbf{X}}^{*r_p}}) = \frac{\#\{T_k^{*r} \geq T_k\}}{R}.$$

(b) Compute unconditional bootstrap P-value

$$p_{r_s}^{\text{boot}}(\lambda_{k0}) = \frac{1}{R_p}\sum_{r_p=1}^{R_p} p_{r_s}^{\text{boot}}(\lambda_{k0|\bar{\mathbf{X}}^{*r_p}}).$$

4. Compute overall P-value

$$p^{\text{boot}}(\lambda_{k0}) = \frac{1}{R_s}\sum_{r_s=1}^{R_s} p_{r_s}^{\text{boot}}(\lambda_{k0}).$$

Note that the size of R_s, R_p, R_{ps}, and R is discussed further in Section 4.4, and was chosen based on time constraints. This will similarly be the case for the bootstrap parameters in the CI methods to be discussed.

As with all of the CI methods, to then solve $\{\lambda_{i0} : p^{\text{boot}}(\lambda_{i0}) \geq \alpha\}$ for the $100(1-\alpha)\%$ CI, we will use the bisection root-finding technique.

Recall from Section 3.4 that the $SMixS\mathcal{M}$ distribution appeared to best fit the diet estimates. We now indicate the modifications to the PAR procedure needed to obtain CIs based on this distribution. In Section 2.4, we defined the natural MOL for observations modeled by the $SMixS\mathcal{M}(\theta_k, \mu_k, \sigma_k, \alpha_k)$ distribution to be

$$\lambda_k^s = \begin{cases} 0 & \text{if } \theta_k = 1 \\ (1-\theta_k)\frac{e^{\xi_k(\mu_k,\sigma_i^2,\alpha_k)}}{1+e^{\xi_k(\mu_k,\sigma_k^2,\alpha_k)}} & \text{if } \theta_k < 1, \end{cases}$$

where $\xi_k(\mu_k, \sigma_k^2, \alpha_k) = \sigma_k \left(\frac{2}{\pi}\right)^{\frac{1}{2}} \frac{\alpha_k}{(1+\alpha_k^2)^{\frac{1}{2}}} + \mu_k$.

The extension of the PAR intervals to CIs for λ_k^s is straightforward. The test statistic in Step 2. is now $T_k = |\hat{\lambda}_k^s - \lambda_{k0}^s|$ and in Step vi. of the initial bootstrap, simply let $s_{k|\bar{\mathbf{X}}^{*r_p}}^{r_s,\text{boot}}$, be the MLE of $\sigma_{k|\bar{\mathbf{X}}}$ computed using the R_{ps} estimates in Step v. From these estimates, compute also the MLE of α_k and call this $\alpha_{k|\bar{\mathbf{X}}^{*r_p}}^{r_s,\text{boot}}$. To carry out the parametric bootstrap, modify Step A. to generate observations from the $SMixS\mathcal{M}$ distribution using the function *rsn* in the *sn* library with location parameter,

$$\mu_{k0} = \log\left(\frac{\frac{\lambda_{k0}^s}{1-\theta_{k|\bar{\mathbf{X}}}}}{1-\frac{\lambda_{k0}^s}{1-\theta_{k|\bar{\mathbf{X}}}}}\right) - \sigma_{k|\bar{\mathbf{X}}}\left(\frac{2}{\pi}\right)^{\frac{1}{2}} \frac{\alpha_{k|\bar{\mathbf{X}}}}{(1+\alpha_{k|\bar{\mathbf{X}}}^2)^{\frac{1}{2}}},$$

and nuisance parameters, $\theta_{k|\bar{\mathbf{X}}}$, $\sigma_{k|\bar{\mathbf{X}}}$, and $\alpha_{k|\bar{\mathbf{X}}}$, determined by the initial bootstrap. We will refer to this interval method as the Skew-Parametric (SKEW-PAR) method.

4.3.4 Semi-Parametric Intervals

Recall that the main disadvantage of the intervals derived for estimates modeled by the $Mix\mathcal{M}$ distribution is the poor estimate of the populations that would no doubt result when n_s is small. We have therefore developed a modified, bootstrap CI method. While we will still proceed by computing, for each population, a P-value for the test of

$$\begin{aligned} \mathrm{H}_0 &: \ \eta_k(\mathbf{v}_j) = \eta_{k0} \\ \mathrm{H}_1 &: \ \eta_k(\mathbf{v}_j) \neq \eta_{k0}, \end{aligned} \tag{4.11}$$

(and consider η_k to be our parameter of interest), we will not assume that the diet estimates are modeled by a specific parametric density. We have accordingly termed the intervals of this subsection "semi-parametric".

As with the PAR intervals, the semi-parametric (SEMI-PAR) intervals involve an initial bootstrap to estimate the populations. A major difficulty in the development of the SEMI-PAR intervals was the potential for there to be no observations in our sample from one or more of the populations determined from our bootstrap sample. As a result, for a given population, any test statistic that requires observations from this population (such as the test statistic in Equation 2.25) is unusable. We will instead use a test statistic based on the estimates $\mathbf{q}^{\mathbf{v}_b}(\mathbf{Y}, \bar{\mathbf{X}})$, where $\mathbf{q}^{\mathbf{v}_b}(\mathbf{Y}, \bar{\mathbf{X}})$ is the

estimate of diet obtained when the optimization is restricted to the species in group $\mathbf{v}_b$. We now consider the details of the SEMI-PAR interval method.

As discussed in Subsection 2.5.2, if the $\theta_{\mathbf{v}_b}$ were known and $\eta_{k0} = 0$, a P-value for the overall test

$$\begin{aligned} \mathrm{H}_0 &: \eta_k = \eta_{k0} \\ \mathrm{H}_1 &: \eta_k \neq \eta_{k0}, \end{aligned} \tag{4.12}$$

would be unnecessary since

$$\begin{aligned} \eta_k = 0 &\Leftrightarrow \nexists b \text{ such that } k \in \mathbf{v}_b \\ &\Leftrightarrow \sum_{\{\mathbf{v}_b : k \notin \mathbf{v}_b\}} \theta_{\mathbf{v}_b} = 1. \end{aligned}$$

Furthermore, within any population, say $\mathbf{v}_b$, if $k \notin \mathbf{v}_b$, then $\eta_k(\mathbf{v}_b) = 0$ and a P-value need not be computed for the test in Equation 4.11. For these special cases, we will simply assign zero or a generated $U[0,1]$ random variable to the P-value (as in the PAR interval method) depending on whether or not H_0 should be rejected.

Now suppose that $k \in \mathbf{v}_b$. Since

$$\eta_k(\mathbf{v}_b) = \frac{e^{\mu_k^{\mathbf{v}_b}}}{1 + e^{\mu_k^{\mathbf{v}_b}}},$$

we have

$$\eta_k(\mathbf{v}_b) = \eta_{k0} \Leftrightarrow \mu_{k0}^{\mathbf{v}_b} = \log\left(\frac{\eta_{k0}}{1 - \eta_{k0}}\right).$$

Consider estimating $\mu_{k|\bar{\mathbf{X}}}^{\mathbf{v}_b}$ by

$$\hat{\mu}_k^{\mathbf{q}^{\mathbf{v}_b}} = \frac{1}{n_s} \sum_{i=1}^{n_s} \log\left(\frac{q_k^{\mathbf{v}_b}(\mathbf{Y}_i, \bar{\mathbf{X}})}{1 - q_k^{\mathbf{v}_b}(\mathbf{Y}_i, \bar{\mathbf{X}})}\right). \tag{4.13}$$

Although we do not know the exact distribution of $\hat{\mu}_k^{\mathbf{q}^{\mathbf{v}_b}}$, we will assume that

$$Z_{k|\bar{\mathbf{X}}}^{\mathbf{v}_b} = \frac{\hat{\mu}_k^{\mathbf{q}^{\mathbf{v}_b}} - \mu_{k|\bar{\mathbf{X}}}^{\mathbf{v}_b}}{\mathrm{VAR}_{\mathbf{Y}}\left[\hat{\mu}_k^{\mathbf{q}^{\mathbf{v}_b}} | \bar{\mathbf{X}}\right]}, \tag{4.14}$$

is approximately pivotal, not depending on any unknown parameters. Let $\frac{s_k^{2,\mathbf{q}^{\mathbf{v}_b}}}{n_s}$ be our estimate of $\mathrm{VAR}_{\mathbf{Y}}\left[\hat{\mu}_{k|\bar{\mathbf{X}}}^{\mathbf{q}^{\mathbf{v}_b}}\right]$, where

$$s_k^{2,\mathbf{q}^{\mathbf{v}_b}} = \frac{1}{n_s - 1} \sum_{i=1}^{n_s} \left(\log\left(\frac{q_k^{\mathbf{v}_b}(\mathbf{Y}_i, \bar{\mathbf{X}})}{1 - q_k^{\mathbf{v}_b}(\mathbf{Y}_i, \bar{\mathbf{X}})}\right) - \hat{\mu}_k^{\mathbf{q}^{\mathbf{v}_b}}\right).$$

For 4.14 pivotal, we may carry out a studentized bootstrap as described in Davison and Hinkley (1997). More specifically, our observed studentized test statistic is

$$Z_{k0}^{\mathbf{v}_b} = \frac{\hat{\mu}_k^{\mathbf{q}^{\mathbf{v}_b}} - \log\left(\frac{\eta_{k0}}{1-\eta_{k0}}\right)}{\frac{s^{\mathbf{q}^{\mathbf{v}_b}}}{\sqrt{n_s}}},$$

and to approximate the distribution of this test statistic we will compute

$$Z_k^{\mathbf{v}_b,*r} = \frac{\hat{\mu}_k^{\mathbf{q}^{\mathbf{v}_b,*r}} - \hat{\mu}_k^{\mathbf{q}^{\mathbf{v}_b}}}{\frac{s^{\mathbf{q}^{\mathbf{v}_b,*r}}}{\sqrt{n_s}}}, \; r = 1, \ldots, R.$$

For an unconditional studentized bootstrap P-value, we will bootstrap the prey in addition to the seals.

In summary, the SEMI-PAR interval method consists of an initial bootstrap to estimate the $\theta_{\mathbf{v}_b}$, followed by a studentized bootstrap that yields a P-value, say $p^{\text{boot}}(\eta_{k0}(\mathbf{v}_b))$, for each population. The overall P-value is then either

$$p^{\text{boot}}(\eta_{k0}) = \sum_{\{b:k\in\mathbf{v}_b\}} \theta_{\mathbf{v}_b}^{\text{boot}} p^{\text{boot}}(\eta_{k0}(\mathbf{v}_b)),$$

or

$$p^{\text{boot}}(\eta_{k0}) = P\left[T^{\text{boot}} > T_{\text{obs}}^{\text{boot}}\right],$$

where

$$T^{\text{boot}} = -2 \sum_{\{b:k\in\mathbf{v}_b\}} \log\left[p^{\text{boot}}(\eta_{k0}(\mathbf{v}_b))\right] \dot{\sim} \chi^2_{2\#\{b:k\in\mathbf{v}_b\}}.$$

The CI is determined by

$$\{\eta_{k0} : p^{\text{boot}}(\eta_{k0}) \geq \alpha\}.$$

It should be mentioned that the initial bootstrap is carried out in such a way that $\theta_{\mathbf{v}_b}^{\text{boot}}$ is not conditional on $\bar{\mathbf{X}}$. That is, for each generated pseudo-seal the corresponding diet estimate will use resampled prey. Also, the entire algorithm will be repeated several times (by generating multiple samples of seals and prey) to incorporate the variability in $\theta_{\mathbf{v}_b}^{\text{boot}}$. Finally, although the SEMI-PAR intervals do not require n_s to be large, $n_s \neq 1$ due to the variance estimate used in the Z test statistic.

The SEMI-PAR algorithm for computing $p^{\text{boot}}(\eta_{k0})$ for the hypothesis test in Equation 4.12 is as follows:

SEMI-PAR Algorithm

1. *for* $r_o = 1 : R_o$

   (a) Generate n_s seals: $\mathbf{Y}_1^{*r_o}, \ldots, \mathbf{Y}_{n_s}^{*r_o}$. (See Appendix C.)

   (b) Generate sample of prey: $\mathbf{X}_1^{*r_o}, \ldots, \mathbf{X}_I^{*r_o}$. (See Appendix C.)

   (c) Compute n_s diet estimates: $p_{k,1}^*(\mathbf{Y}_1^{*r_o}, \bar{\mathbf{X}}^{*r_o}), \ldots, p_{k,n_s}^*(\mathbf{Y}_{n_s}^{*r_o}, \bar{\mathbf{X}}^{*r_o})$.

Initial Bootstrap

   (d) Generate $\frac{R_{ps}}{n_s}$ pseudo-seals from each of the diet estimates in (c) : $\mathbf{Y}_1^{**r_o}, \ldots, \mathbf{Y}_{R_{ps}}^{**r_o}$.

   (e) Generate R_{ps} samples of prey using prey in (b): $\mathbf{X}_1^{*r_o,r_{ps}}, \ldots, \mathbf{X}_I^{*r_o,r_{ps}}$, $r_{ps} = 1, \ldots, R_{ps}$.

   (f) Compute R_{ps} diet estimates: $p_{k,1}^{**}(\mathbf{Y}_1^{**r_o}, \bar{\mathbf{X}}^{*r_o,1}), \ldots, p_{k,R_{ps}}^{**}(\mathbf{Y}_{R_{ps}}^{**r_o}, \bar{\mathbf{X}}^{*r_o,R_{ps}})$.

   (g) Compute

$$\theta_{\mathbf{v}_b}^{\text{boot},r_o} = \frac{R_{ps}^{\mathbf{v}_b}}{R_{ps}},$$

   where $R_{ps}^{\mathbf{v}_b}$ denotes the number of estimates in (f) belonging to population $\mathbf{v}_b$.

Studentized Bootstrap

   (h) *if* $\eta_{k0} = 0$

      i. *if* $\sum_{\{\mathbf{v}_b : k \notin \mathbf{v}_b\}} \theta_{\mathbf{v}_b}^{\text{boot},r_o} = 1$, generate $p^{\text{boot}}(\eta_{k0})$ from $U[0,1]$. Go to Step 1.

      ii. *else* set $p^{\text{boot}}(\eta_{k0}) = 0$. Go to Step 1.

   (i) *for* $b = 1 : n_B$

      i. *if* $k \notin \mathbf{v}_b$, set $p^{\text{boot},r_o}(\eta_{k0}(\mathbf{v}_b)) = 0$. Go to Step (i).

      ii. Compute n_s "restricted" diet estimates: $\mathbf{q}_1^{\mathbf{v}_b}(\mathbf{Y}_1, \bar{\mathbf{X}}), \ldots, \mathbf{q}_{n_s}^{\mathbf{v}_b}(\mathbf{Y}_{n_s}, \bar{\mathbf{X}})$. (Note that if for some species, say s, $s \in \mathbf{v}_b$ but becomes zero in the restricted optimization, we modify the estimate to be a small quantity and re-normalize.)

iii. Compute

$$Z_{k0}^{\mathbf{v}_b} = \frac{\hat{\mu}_k^{\mathbf{q}^{\mathbf{v}_b}} - \log\left(\frac{\eta_{k0}}{1-\eta_{k0}}\right)}{\frac{s^{\mathbf{q}^{\mathbf{v}_b}}}{\sqrt{n_s}}}.$$

iv. *for* $r = 1 : R$

A. Generate sample of seals (using *original* seals): $\mathbf{Y}_1^{*r}, \ldots, \mathbf{Y}_{n_s}^{*r}$

B. Generate sample of prey (using *original* prey): $\mathbf{X}_1^{*r}, \ldots, \mathbf{X}_I^{*r}$.

C. Compute "restricted" diet estimates:
$q_{k,1}^{*\mathbf{v}_b}(\mathbf{Y}_1^{*r}, \bar{\mathbf{X}}^{*r}), \ldots, q_{k,n_s}^{*\mathbf{v}_b}(\mathbf{Y}_{n_s}^{*r}, \bar{\mathbf{X}})$.

D. Compute

$$Z_k^{\mathbf{v}_b,*r} = \frac{\hat{\mu}_k^{\mathbf{q}^{\mathbf{v}_b,*r}} - \hat{\mu}_k^{\mathbf{q}^{\mathbf{v}_b}}}{\frac{s^{\mathbf{q}^{\mathbf{v}_b,*r}}}{\sqrt{n_s}}}.$$

v. Compute

$$p^{\text{boot},r_o}(\eta_{k0}(\mathbf{v}_b)) = \frac{\#\left\{(Z_k^{\mathbf{v}_b,*r})^2 \geq (Z_{k0}^{\mathbf{v}_b})^2\right\}}{R}.$$

(j) Compute

$$p^{\text{boot},r_o}(\eta_{k0}) = \sum_{b=1}^{n_B} \theta_{\mathbf{v}_b}^{\text{boot},r_o} p^{\text{boot},r_o}(\eta_{k0}(\mathbf{v}_b)).$$

or

$$p^{\text{boot},r_o}(\eta_{k0}) = P\left[T^{\text{boot},r_o} > T_{\text{obs}}^{\text{boot},r_o}\right],$$

where

$$T^{\text{boot},r_o} = -2 \sum_{\{b:k\in\mathbf{v}_b\}} \log\left[p^{\text{boot},r_o}(\eta_{k0}(\mathbf{v}_b))\right] \dot{\sim} \chi^2_{2\#\{b:k\in\mathbf{v}_b\}}.$$

2. Compute

$$p^{\text{boot}}(\eta_{k0}) = \frac{1}{R_o} \sum_{r_0=1}^{R_o} p^{\text{boot},r_o}(\eta_{k0}).$$

4.3.5 Nonparametric Intervals

In this subsection we consider two nonparametric methods of obtaining confidence intervals through bootstrapping. We first consider methods that use percentiles of the bootstrap distribution of the point estimators of diet. Our final CI method involves generating diet estimates (nonparametrically) under a null hypothesis and inverting

the hypothesis test as we have done in previous methods.

Percentile Methods

In Subsection 4.3.2, CIs based on the point estimators of diet ($\bar{p}_k(\bar{\mathbf{X}})$, for example) being approximately normally distributed were presented. Recall that we also considered applying the normal approximation to these estimates after an arcsin transformation. The percentile methods discussed in DiCiccio and Efron (1996) and in Davison and Hinkley (1997) assume that a transformation to normality is possible but do not require that the transformation be found. It is important to note that although the percentile methods allow for a bias correction factor, this bias essentially adjusts for the difference between $\mathrm{E}_{\mathbf{Y},\bar{\mathbf{X}}}\left[\bar{\mathbf{p}}(\bar{\mathbf{X}})\right]$ and $\boldsymbol{\mu}_{\mathbf{p}} = \mathrm{E}_{\mathbf{Y}}[\mathbf{p}(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})]$, and similarly for the other point estimators. The bias between $\boldsymbol{\mu}_{\mathbf{p}}$ and $\boldsymbol{\pi}$ must still be estimated using our bias adjustment method given in Section 4.2

For the remainder of this section we will use $\bar{p}_k(\bar{\mathbf{X}})$ as our example point estimator and describe the percentile method approach of obtaining a CI for $\boldsymbol{\mu}_{\mathbf{p}}$. To simplify the notation, let $\bar{p}_k = \bar{p}_k(\bar{\mathbf{X}})$. Recall, however, that for n_s and n_k, $k = 1, \ldots, I$ large, $\bar{p}_k$ is approximately normally distributed and a CI interval for μ_{p_k} is given in Equation 4.1.

Underlying DiCiccio and Efron's (1996) CI method is the assumption of a monotone increasing transformation, say $\phi_k = t(\mu_{p_k})$, such that $\hat{\phi}_k = t(\bar{p}_k)$ is normally distributed for all μ_{p_k} with a possible bias and nonconstant variance. That is, DiCiccio and Efron (1996) assume that,

$$\hat{\phi}_k \sim \mathcal{N}(\phi_k - z_0\sigma_{\phi_k}, \sigma^2_{\phi_k}), \tag{4.15}$$

where $\sigma_{\phi_k} = 1 + a\phi_k$. They then derive confidence limits for ϕ_k and ultimately for μ_{p_k} by transforming these limits back to the original scale using the bootstrap distribution of $\bar{p}_k$. (The parameters z_0 and a need to be estimated and are discussed shortly.)

More specifically, consider carrying out Steps (a)-(d) in the large sample bootstrap algorithm (Subsection 4.3.2) to obtain $\bar{p}^{*r}$, $r = 1, \ldots, R$. Approximate the cdf of $\bar{p}_k$ by

$$\hat{G}(c) = \frac{\#\{\bar{p}^{*r} < c\}}{R}.$$

If z_0 and a were known, an α confidence limit for μ_{p_k}, based on Equation 4.15, is

$$\bar{p}^*_{k[(R+1)\tilde{\alpha}]},$$

where

$$\tilde{\alpha} = \hat{G}^{-1}\left[\Phi\left(z_0 + \frac{z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)}\right)\right],$$

Φ is the standard normal cdf and $\bar{p}^*_{k[(R+1)\tilde{\alpha}]}$ is the $(R+1)\tilde{\alpha}$ largest $\bar{p}^{*r}_k$. (See Davison and Hinkley (1997) for details.) If we let $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$ correspond to $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ respectively then a $(1-\alpha)100\%$ CI for μ_{p_k} is given by

$$\left[\bar{p}^*_{k[(R+1)\tilde{\alpha}_1]}, \bar{p}^*_{k[(R+1)\tilde{\alpha}_2]}\right]. \tag{4.16}$$

(If R is not such that $(R+1)\tilde{\alpha}_1$ and $(R+1)\tilde{\alpha}_2$ are integers, then we may use *floor*$((R+1)\tilde{\alpha}_1)$ and *ceiling*$((R+1)\tilde{\alpha}_2)$ or interpolation.) DiCiccio and Efron (1996) call a the acceleration and accordingly have termed the intervals given by Equation 4.16 the "bias-corrected and accelerated" (BC_a) intervals. If z_0 and a are assumed to be zero, we obtain Davison and Hinkley's (1997) *basic* percentile CIs. The $(1-\alpha)100\%$ basic percentile (PERC) interval for μ_{p_k} is given by

$$\left[\bar{p}^*_{k[(R+1)\frac{\alpha}{2}]}, \bar{p}^*_{k[(R+1)(1-\frac{\alpha}{2})]}\right]. \tag{4.17}$$

Otherwise, z_0 and a must be estimated.

Davison and Hinkley (1997) derived the following estimate for z_0

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\bar{p}^{*r} \leq \bar{p}_k\}}{R+1}\right). \tag{4.18}$$

The estimate of a may be parametric, if the $\bar{p}^{*r}_k$ were generated through a parametric bootstrap, or nonparametric as will be the case in our simulations. If $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_s}$ and $\mathbf{X}_1, \ldots, \mathbf{X}_I$ are treated as $I+1$ independent samples, then Davison and Hinkley's (1997) recommended estimate of a is

$$\hat{a} = \frac{1}{6}\frac{\sum_{m=1}^{I+1} n_m^{-3} \sum_{i=1}^{n_m} l_{mi}^3}{\left(\sum_{m=1}^{I+1} n_m^{-2} \sum_{i=1}^{n_s} l_{mi}^2\right)^{\frac{3}{2}}},$$

where l_{mi}, $i = 1, \ldots, n_m$, $m = 1, \ldots, (I+1)$ are the empirical influence values. To approximate the empirical influence values the jackknife method given in Davison and Hinkley (1997) may be used. The jackknife estimate of l_{mi} is given by

$$l_{\text{jack},im} = (n_m - 1)\left(\bar{p}_k - \bar{p}_{k,-mi}\right),$$

where $\bar{p}_{k,-mi}$ is the point estimator computed without the ith observation in the mth sample.

DiCiccio and Efron (1996) recommend that approximately $R = 2000$ bootstrap replications be used to compute the BC_a intervals. Since for each r ($r = 1, \ldots, 2000$), n_s optimizations are required, this procedure is extremely computationally intensive. An alternative approach is to use the ABC method (discussed also in DiCiccio and Efron (1996) and in Davison and Hinkley (1997)) in which the BC_a endpoints are approximated analytically. Essentially the ABC intervals require estimating the parameters a and z_0, as well as an additional parameter, and this can be accomplished without bootstrapping. The S-PLUS function *abc.ci* in the *boot* library may be used to compute approximate nonparametric CIs. Note that the function requires that the statistic be in weighted form. If our statistic is $\bar{\mathbf{p}}$ then we may write it as

$$\bar{\mathbf{p}} = \sum_{i=1}^{n_s} \frac{1}{n_s} \mathbf{p}\left(\mathbf{Y}_i, \sum_{i=1}^{n_1} \frac{1}{n_1}\mathbf{X}_1, \ldots, \sum_{i=1}^{n_I} \frac{1}{n_I}\mathbf{X}_I\right).$$

For our problem $n = n_s + \sum_{k=1}^{I} n_k$ will be large (with the reduced prey base, $n \approx 500$) and the number of optimizations will still be extensive.

We have, consequently, not found the BC_a and ABC methods, applied in this manner, to be practical for our problem. That is, computing BC_a and/or ABC intervals by re-sampling the data and then computing the diet estimates is a computational burden. If, instead, the BC_a and/or ABC CIs were computed using a nonparametric bootstrap of the diet estimates themselves, CIs could be obtained relatively quickly, specifically for the ABC intervals. In this case the estimate of z_0 is still given by Equation 4.18 though the $\bar{p}_k^{*r}$ are computed from the resampled diet estimates and not the large sample bootstrap algorithm specified in Subsection 4.3.2. Also, for the estimator $\bar{p}_k$, the estimate of a simplifies as follows

$$\begin{aligned}
\hat{a} &= \frac{1}{6}\frac{\sum_{i=1}^{n_s} l_i^3}{\left(\sum_{i=1}^{n_s} l_i^2\right)^{\frac{3}{2}}} \\
&= \frac{1}{6}\frac{\sum_{i=1}^{n_s}\left[p_k(\mathbf{Y}, \bar{\mathbf{X}}) - \bar{p}_k\right]^3}{\left\{\sum_{i=1}^{n_s}\left[p_k(\mathbf{Y}, \bar{\mathbf{X}}) - \bar{p}_k\right]^2\right\}^{\frac{3}{2}}}.
\end{aligned}$$

For the other point estimators, a jackknife approximation to the empirical influence

values may be used. Note that if we simply let $a = 0$ and $z_0 = 0$ we obtain alternative basic percentile method intervals that are based on the re-sampling of the diet estimates. To distinguish these percentile intervals from those given in Equation 4.17 (which require the seal and prey FA signatures to be re-sampled) we will denote the intervals in Equation 4.17 by PERC (Correct), since they incorporate the variability due to the prey and are more correct in this sense.

Nonparametric Bootstrap Intervals

We now discuss a second nonparametric approach to obtaining a CI for π_k which we will refer to as the NONPAR interval method.

As in both the PAR and SEMI-PAR methods, the nonparametric (NONPAR) CIs of this subsection are found by inverting a hypothesis test for which bootstrap P-values are computed. Our NONPAR test of interest is

$$\begin{aligned} \mathrm{H}_0 &: \ \pi_k = \pi_{k0} \\ \mathrm{H}_1 &: \ \pi_k \neq \pi_{k0}, \end{aligned} \tag{4.19}$$

and we obtain a bootstrap P-value, say $p^{\text{boot}}(\pi_{k0})$, by generating pseudo-seals under the null in Equation 4.19. A difficulty is that π_j, $j \neq k$, are essentially nuisance parameters. Let $\mathbf{p}^{-k}(\mathbf{Y}, \bar{\mathbf{X}})$ denote the diet estimate without species k (that is, the optimization is restricted to species j, $j \neq k$), then we will estimate the nuisance parameter π_j, $j \neq k$ by $(1 - \pi_{k0})p_j^{-k}(\mathbf{Y}, \bar{\mathbf{X}})$, where $p_j^{-k}(\mathbf{Y}, \bar{\mathbf{X}})$ is the component of $\mathbf{p}^{-k}(\mathbf{Y}, \bar{\mathbf{X}})$ corresponding to the jth species. Our test statistic will be $T = |\hat{\pi}_k - \pi_{k0}|$, where $\hat{\pi}_k$ could be any of the point estimators in Table 3.3. The complete algorithm for computing $p^{\text{boot}}(\pi_{k0})$ is as follows:

NONPAR Algorithm

1. Compute n_s diet estimates: $p_{k,1}(\mathbf{Y}_1, \bar{\mathbf{X}}), \ldots, p_{k,n_s}(\mathbf{Y}_{n_s}, \bar{\mathbf{X}})$.

2. Compute $\hat{\pi}_k$ and $T_k = |\hat{\pi}_k - \pi_{k0}|$.

3. Compute $p_i^{-k}(\mathbf{Y}_i, \bar{\mathbf{X}})$, $i = 1, \ldots, n_s$.

4. *for* $r = 1 : R$

   (a) Generate n_s seals, $\mathbf{Y}_1^{*r}, \ldots, \mathbf{Y}_{n_s}^{*r}$, where the ith seal has null diet

$$\Big[(1-\pi_{k0})p_{1,i}^{-k}(\mathbf{Y}_i, \bar{\mathbf{X}}), \ldots, (1-\pi_{k0})p_{(k-1),i}^{-k}(\mathbf{Y}_i, \bar{\mathbf{X}}), \\ \pi_{k0}, p_{(k+1),i}^{-k}(\mathbf{Y}_i, \bar{\mathbf{X}}), \ldots, p_{I,i}^{-k}(\mathbf{Y}_i, \bar{\mathbf{X}})\Big].$$

   (b) Generate sample of prey: $\mathbf{X}_1^{*r}, \ldots, \mathbf{X}_I^{*r}$.

   (c) Compute n_s diet estimates: $\mathbf{p}_1(\mathbf{Y}_1^{*r}, \bar{\mathbf{X}}_k^{*r}), \ldots, \mathbf{p}_{n_s}(\mathbf{Y}_{n_s}^{*r}, \bar{\mathbf{X}}_k^{*r})$.

   (d) Compute $T_k^{*r} = |\hat{\pi}_k^{*r} - \pi_{k0}|$.

5. Compute

$$p^{\text{boot}}(\pi_{k0}) = \frac{\#\{T_k^{*r} \geq T_k\}}{R}.$$

4.4 Simulation Study

To assess the CI methods of the previous section, a simulation study was carried out in which coverage probabilities and CI lengths were computed for the various methods. Before presenting our results, we first address issues relating to the implementation of the study.

4.4.1 Implementation

In our simulation study, M (to be discussed) samples of pseudo-seals were generated with a specified true diet and, for each sample, a CI was computed using the methods of Section 4.3. The coverage probability for the kth species, calculated as

$$\frac{\#\text{ of CI containing } \pi_k}{M},$$

and the average length of the M CIs were used to compare the various methods. We chose to compute Bonferroni style simultaneous CIs and used $\alpha = \frac{0.10}{I}$ so that our overall target coverage probability was 0.90.

In carrying out a thorough simulation study, we were fairly limited by the computational intensiveness and slowness of the CI methods. Consequently, we could not examine all of the combinations of sample size, distance measure and diet of interest.

Furthermore, we were compelled to use a smaller number of bootstrap samples than would have been preferred. For many methods, computing a single CI could take several hours even when a small number of bootstrap samples was used. We therefore opted to first carry out a small preliminary study through which it was anticipated that we would be able to eliminate some of the methods. Further simulations were then carried out in accordance with the results of this preliminary study.

In our preliminary study, we chose to examine the CI methods at one sample size ($n_s = 10$), with one distance measure (AIT distance) and one diet (Diet 4). The M samples of n_s pseudo-seals were generated as in Appendix B. In this preliminary study we used only $M = 20$ but M was later increased to 100 for our chosen methods. Table 4.1 contains the re-sampling related parameters that were used throughout the simulations. The size of the parameters was chosen to be such that the simulations could be carried out in a reasonable amount of time. In practice, where only a few CIs might be needed, it is recommended that the size of the parameters be set as large as possible.

Even with the smaller than desired parameter sizes specified in Table 4.1, the SEMI-PAR, SKEW-PAR and NONPAR methods were found to be exceptionally slow. To assist in speeding up the SEMI-PAR method, groups containing only one observation were assigned a P-value that was the weighted average of the P-values from groups containing more than one observation. To speed up the SKEW-PAR method, we replaced $\hat{\lambda}_k^s$ (obtained using functions in the *sn* library) by $\hat{\lambda}_k$ since $\hat{\lambda}_k$ was much faster to obtain and the difference between the two estimates is usually minimal. Due to the extreme slowness of the NONPAR method, it was found to be an impractical method when run in S-PLUS as it could take a couple of days to obtain one CI at $n_s = 10$, with $R = 50$. In addition to using a reduced size of R of only 50, we attempted to speed up the method by using improved starting values in the optimization required to obtain the diet estimates. That is, instead of using equal proportions as the starting diet vector in our bootstrap samples, we made use of the diet estimates computed from our original sample. Still the method took a couple of days to obtain one interval. We therefore had Wade Blanchard (Dalhousie University) convert the S-PLUS code to Fortran and it ran much faster. Although

Method	Parameters
Large Sample	
Case 1	—
Case 2	$R = 300$
PAR	$R_s = 10$
	$R_p = 10$
	$R_{ps} = 50$
	$R = 50$
SKEW PAR	$R_s = 10$
	$R_p = 10$
	$R_{ps} = 50$
	$R = 50$
SEMI-PAR	$R_o = 5$
	$R_{ps} = 50$
	$R = 50$
Nonparametric	
Percentile	
PERC (Correct)	$R = 300$
PERC, BCA	$R = 2000$
NONPAR	$R = 100$
Bias	$R_{ps} = 100$

Table 4.1: Re-sampling parameters used in the various CI methods and in the bias estimation algorithm.

not formally timed, with $n_s = 10$ and $R = 100$, obtaining one CI took roughly a few hours to run in Fortran versus a few days in S-PLUS.

4.4.2 Preliminary Results

Our preliminary results are based on a sample size of 10, the AIT distance measure and true diet, Diet 4. We used $M = 20$ and would therefore expect our true coverage probability to be within $\pm 1.96\sqrt{\frac{(0.90)(0.10)}{20}} = 0.13$ (with 95% confidence) of our computed coverage probability. The results are given in Tables 4.2-4.6. A graphical summary of the results is given in Figure 4.1.

Tables 4.2 and 4.3 contain the large sample interval results discussed in Subsection 4.3.2. The intervals in Table 4.2 correspond to Case 1 and were computed without the

use of re-sampling techniques (except to obtain the bias estimate). Type "Normal" refers to CIs computed using Equation 4.1 while the type "t" intervals are similar but quantiles of the t distribution were used instead. Intervals labeled type "$SMix\mathcal{M}$" were computed using Equation 4.7. Although intervals based on the $Mix\mathcal{M}$ distribution were also computed (see the P-value method discussed Subsections 2.5.2 and 4.3.3), they were found to be problematic at $n_s = 10$ because often all groups contained only one observation and the test statistic relies on an estimate of variance from at least one group. Recall Figures 3.4 - 3.7 where histograms displaying the distribution of the diet estimates are given. Often there did not appear to be a large difference between the fits of the $Mix\mathcal{M}$ and $SMixM$ distributions and we have chosen to simply present the results of CIs based on the $SMix\mathcal{M}$ distribution.

From Table 4.2 and Figure 4.1 it appears from our preliminary results that a sample size of 10 might be too small to guarantee a coverage probability of 0.90 for all species using a normal approximation. The coverage probabilities presented for the t and $SMix\mathcal{M}$ intervals are better and generally good when a bias adjustment (denoted by BA) is used. The average lengths of the intervals are relatively long however, specifically for Sandlance, compared to the other methods to be discussed. The longer intervals are perhaps reflective of less accurate estimates of the variance of the diet estimates compared to the bootstrapped based methods. Note that in terms of bringing the coverage probability closer to 0.90, the bias correction generally yields equivalent coverage probabilities (for example 1.00 to 0.80) or an improved coverage probability. An exception is with SilverHake where the coverage probability using the t intervals decreased from 1.00 to 0.75.

For Case 2 of the large sample intervals (Table 4.3), the lengths of the intervals, for Sandlance in particular, are shorter but the coverage probabilities are a little low. The latter is especially true when the mean, $\bar{p}_k$, is our estimate. For the sample median, m_k, the coverage probabilities are more reasonable except for SilverHake. Using an arcsin transformation does not appear to be helpful for either estimator.

Overall, the Case 1 large sample intervals are simple to compute and appear to have some potential for $n_s \geq 10$. Consequently, we will examine these intervals further in Subsection 4.4.3. The Case 2 intervals, however, will not be investigated further

as the nonparametric methods, to be discussed, tend to give better results.

Type	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
	Diet 4	0.09	0	0.09	0	0.45	0.09	0.09	0.09
Normal ($\bar{p}_k$)	No BA	0.95 (0.18)	1.00 (0.09)	1.00 (0.16)	0.45 (0.25)	0.80 (0.30)	1.00 (0.17)	0.70 (0.07)	0.90 (0.13)
	BA	0.85 (0.17)	1.00 (0.06)	0.70 (0.16)	0.80 (0.17)	0.90 (0.30)	0.60 (0.13)	0.85 (0.07)	0.90 (0.13)
t ($\bar{p}_k$)	No BA	1.00 (0.22)	1.00 (0.10)	1.00 (0.19)	0.80 (0.29)	0.95 (0.37)	1.00 (0.20)	0.85 (0.09)	0.95 (0.16)
	BA	0.95 (0.20)	1.00 (0.07)	1.00 (0.19)	0.90 (0.21)	0.90 (0.37)	0.75 (0.16)	0.90 (0.09)	0.90 (0.15)
$SMixM$ ($\hat{\lambda}_k$)	No BA	0.95 (0.21)	0.80 (0.20)	1.00 (0.21)	0.30 (0.29)	0.90 (0.39)	1.00 (0.28)	0.90 (0.10)	1.00 (0.18)
	BA	0.85 (0.20)	0.90 (0.19)	0.80 (0.21)	0.70 (0.25)	0.90 (0.39)	0.95 (0.26)	0.90 (0.10)	0.90 (0.18)

Table 4.2: Diet 4, AIT, $n_s = 10$: Coverage probabilities (average lengths) of CIs computed using large sample methods (Case 1). (Based on 20 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients).

In Table 4.4, results from the percentile methods discussed in Subsection 4.3.5 are presented. Recall that we can obtain percentile intervals by re-sampling the seals and the prey or by re-sampling the diet estimates. The former are considered to be more correct as they incorporate the variability due to the prey. These intervals are denoted by PERC (Correct) in the table. When the diet estimates are re-sampled, we can also obtain the BC_a intervals and these are given in addition to the basic percentile intervals.

The coverage probabilities for the percentile intervals obtained using the median, m_k, are decent when the bias adjustment is used, and are usually larger than the coverage probabilities based on the mean. A comparison of the performance of the PERC (Correct) method based on the mean versus the median is more easily seen in Figure 4.1. While the lengths of the median based intervals are noticeably wider than those obtained using the mean, the lengths are still reasonable and we will prefer the median based intervals. Except for SilverHake with the BCa method, all three median percentile intervals give similar coverage probabilities when the bias adjustment is applied. The bias adjustment does not appear to be dramatically changing the coverage probabilities though it does improve the coverage obtained

Type	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
	Diet 4	0.09	0	0.09	0	0.45	0.09	0.09	0.09
Normal ($\bar{p}_k$)	No BA	0.95 (0.18)	1.00 (0.12)	1.00 (0.16)	0.15 (0.18)	0.50 (0.18)	1.00 (0.12)	0.75 (0.072)	0.90 (0.10)
	BA	0.90 (0.17)	1.00 (0.08)	0.90 (0.16)	0.80 (0.12)	0.75 (0.18)	0.35 (0.09)	0.80 (0.07)	0.80 (0.10)
Arcsin ($\bar{p}_k$)	No BA	0.90 (0.18)	0.80 (0.11)	1.00 (0.16)	0.05 (0.17)	0.55 (0.18)	1.00 (0.12)	0.85 (0.08)	0.95 (0.10)
	BA	0.90 (0.18)	0.90 (0.08)	0.90 (0.16)	0.70 (0.13)	0.75 (0.18)	0.50 (0.10)	0.85 (0.08)	0.80 (0.10)
Normal (m_k)	No BA	1.00 (0.30)	1.00 (0.13)	1.00 (0.23)	0.80 (0.25)	0.95 (0.33)	0.70 (0.13)	0.90 (0.12)	0.90 (0.15)
	BA	0.95 (0.24)	1.00 (0.11)	0.90 (0.19)	0.85 (0.18)	0.80 (0.26)	0.50 (0.11)	0.85 (0.09)	0.80 (0.13)
Arcsin (m_k)	No BA	1.00 (0.37)	1.00 (0.14)	0.95 (0.26)	0.70 (0.25)	0.95 (0.34)	0.65 (0.16)	0.95 (0.14)	0.90 (0.18)
	BA	0.95 (0.29)	0.80 (0.11)	0.80 (0.21)	0.75 (0.19)	0.80 (0.27)	0.55 (0.13)	0.85 (0.11)	0.85 (0.14)

Table 4.3: Diet 4, AIT, $n_s = 10$: Coverage probabilities (average lengths) of CIs computed using large sample methods (Case 2). (Based on 20 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients).

with Pollock and the PERC (Correct) method. The lengths of the intervals are generally largest for the PERC method. With the exception of Sandlance, the lengths of the PERC (Correct) and BCa intervals are similar. We again surmise that the shorter intervals with the PERC (Correct) intervals are due to a better estimate of the variability in the diet estimates. Based on our results, we have chosen to further investigate the PERC (Correct) method with the median.

The CIs given in Table 4.5 were discussed in Subsections 4.3.3 and 4.3.4. These intervals were developed from the parametric models proposed in Chapter 2 for dealing with compositional data with zeros allowed. Note that two versions of the SEMI-PAR method are shown since we computed our overall P-value in two different ways as discussed in Subsection 4.3.4. The intervals denoted by "χ^2 approx" use the χ^2 approximation while the other SEMI-PAR intervals use the weighted P-Value. Perhaps what is most noticeable from the table is the high coverage probabilities associated with the SEMI-PAR (weighted P-Value) method but the accompanying extremely large interval widths. Although for larger sample sizes the lengths of the SEMI-PAR

Type	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
	Diet 4	0.09	0	0.09	0	0.45	0.09	0.09	0.09
PERC (Corr) ($\bar{p}_k$)	No BA	0.95 (0.18)	0.20 (0.14)	1.00 (0.16)	0.05 (0.18)	0.10 (0.17)	1.00 (0.12)	0.65 (0.07)	0.95 (0.09)
	BA	0.95 (0.18)	0.90 (0.12)	0.95 (0.15)	0.65 (0.14)	0.60 (0.17)	0.40 (0.10)	0.90 (0.07)	0.85 (0.09)
PERC ($\bar{p}_k$)	No BA	0.90 (0.17)	0.75 (0.09)	1.00 (0.15)	0.40 (0.24)	0.85 (0.28)	1.00 (0.17)	0.70 (0.07)	0.90 (0.12)
	BA	0.85 (0.16)	0.95 (0.06)	0.70 (0.15)	0.80 (0.17)	0.85 (0.28)	0.70 (0.13)	0.80 (0.07)	0.90 (0.12)
BC_a ($\bar{p}_k$)	No BA	0.90 (0.17)	0.65 (0.11)	1.00 (0.16)	0.15 (0.26)	0.80 (0.29)	1.00 (0.20)	0.70 (0.07)	0.90 (0.13)
	BA	0.90 (0.17)	0.90 (0.08)	0.65 (0.16)	0.80 (0.20)	0.85 (0.29)	0.85 (0.17)	0.85 (0.07)	0.90 (0.13)
PERC (Corr) (m_k)	No BA	1.00 (0.30)	1.00 (0.20)	1.00 (0.24)	0.75 (0.31)	0.85 (0.32)	1.00 (0.17)	0.90 (0.12)	0.95 (0.17)
	BA	1.00 (0.23)	0.80 (0.16)	1.00 (0.19)	0.85 (0.23)	0.90 (0.25)	1.00 (0.13)	0.90 (0.09)	0.95 (0.13)
PERC (m_k)	No BA	1.00 (0.29)	1.00 (0.12)	0.95 (0.25)	0.90 (0.41)	0.95 (0.50)	0.90 (0.26)	0.90 (0.12)	0.90 (0.20)
	BA	0.95 (0.22)	0.80 (0.09)	0.85 (0.20)	0.95 (0.31)	0.90 (0.40)	0.85 (0.21)	0.90 (0.10)	0.90 (0.16)
BC_a (m_k)	No BA	1.00 (0.27)	1.00 (0.05)	0.90 (0.23)	0.90 (0.28)	0.95 (0.48)	0.65 (0.17)	0.90 (0.12)	0.850 (0.18)
	BA	0.95 (0.21)	0.80 (0.04)	0.85 (0.18)	0.95 (0.21)	0.90 (0.38)	0.65 (0.14)	0.90 (0.10)	0.85 (0.14)

Table 4.4: Diet 4, AIT, $n_s = 10$: Coverage probabilities (average lengths) of CIs computed using percentile methods. (Based on 20 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients).

intervals should decrease, for $n_s \leq 10$ these intervals do not appear to be useful and we have not carried out additional simulations for this method. Note, however, that for this method the individual CIs (not shown) were much shorter and of reasonable lengths compared to the simultaneous Bonferroni intervals that are shown. As previously mentioned the Bonferroni CIs are conservative and perhaps alternative simultaneous CIs could be investigated for this method.

For the SEMI-PAR (χ^2 approximation) method, the interval widths are exceptionally small and the coverage probabilities are fairly low. There is, in particular, a problem with the coverage probability for SilverHake. An addditional difficulty that we encountered with this method was that sometimes our root-finding technique timed out while searching for starting values to use in the bisection algorithm. This occurred once for Haddock, once for Plaice and twice for YellowTail and CIs could not be found in these cases. Because of these problems we did not carry out further simulations with this method.

The PAR and SKEW-PAR methods yield very similar coverage probabilities but the SKEW-PAR intervals are, on the average, shorter and are consequently preferred. Observe, however, that the SKEW-PAR intervals are generally similar to or longer than the PERC (Correct) intervals. (See Figure 4.1.) Also, although the bias adjustment greatly improved the coverage probability for Haddock and Pollock (whose true contributions are zero), for Pollock the coverage probability is still low (0.50). Because 10% noise was used, the true proportion of Pollock in the diet may be closer to 0.05 and a coverage probability based on 0.05 rather than 0 might be higher. Recall, however, from Figures 3.1 and 3.2 that the bias associated with Pollock was large when the AIT distance measure was used, even after taking the noise into account . We are likely underestimating the large bias for Pollock.

It should be noted that a problem with the SKEW-PAR method arose when 20 CIs were attempted to be constructed with Diet 1. For one of the samples, a CI for WinterFlounder ($\pi_k = 0$) could not be found as the bisection algorithm had difficulty finding the desired root. When the P-values were plotted for this sample, they appeared to be very unstable compared to the plots for the other species. The bootstrap parameters were increased to $R_s = 15$ (from 10) , $R_p = 15$ (from 10),

$R_{ps} = 75$ (from 50), and $R = 100$ (from 50) but this did not help. Note that in practice we could use the plot to at least obtain rough CI limits but that this is impractical when attempting to carry out a simulation study.

Type	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
	Diet 4	0.09	0	0.09	0	0.45	0.09	0.09	0.09
PAR	No BA	0.95 (0.36)	0.25 (0.31)	1.00 (0.35)	0 (0.38)	0.95 (0.46)	1.00 (0.35)	1.00 (0.27)	1.00 (0.29)
	BA	0.90 (0.30)	0.90 (0.25)	0.75 (0.29)	0.55 (0.29)	0.85 (0.38)	0.95 (0.28)	0.95 (0.22)	0.90 (0.24)
SKEW-PAR	No BA	0.85 (0.27)	0.20 (0.22)	0.95 (0.25)	0 (0.28)	0.95 (0.37)	0.95 (0.26)	1.00 (0.17)	1.00 (0.19)
	BA	0.85 (0.22)	0.90 (0.17)	0.75 (0.21)	0.50 (0.21)	0.85 (0.30)	0.95 (0.20)	0.90 (0.14)	0.80 (0.15)
SEMI-PAR	No BA	1.00 (0.80)	1.00 (0.89)	1.00 (0.71)	0.90 (0.92)	1.00 (0.44)	1.00 (0.78)	1.00 (0.22)	1.00 (0.84)
	BA	1.00 (0.77)	0.95 (0.85)	1.00 (0.69)	1.00 (0.84)	0.95 (0.44)	1.00 (0.73)	1.00 (0.22)	1.00 (0.83)
SEMI-PAR (χ^2 approx)	No BA	1.00 (0.15)	0.95* (0.05)	0.84* (0.11)	0.55 (0.13)	0.55 (0.21)	0.10 (0.06)	0.60 (0.07)	0.67* (0.12)
	BA	0.90 (0.14)	0.95* (0.03)	0.74* (0.10)	0.85 (0.07)	0.80 (0.20)	0.05 (0.03)	0.70 (0.06)	0.72* (0.12)

Table 4.5: Diet 4, AIT, $n_s = 10$: Coverage probabilities (average lengths) of CIs computed with PAR, SKEW-PAR and SEMI-PAR methods. * Denotes that one or more CIs could not be found. (Based on 20 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients).

Because of the longer interval lengths and the issue of not always being able to find the CI, we prefer the less complex PERC (Correct) method intervals and have not carried out further simulations with the SKEW-PAR method.

Finally, the results of the NONPAR method are given in Table 4.6. Although the coverage probabilities are exceptionally high when our point estimator is the median, the interval lengths appear to be too long. When the mean is used, the coverage probabilities are sometimes a little low but overall they are satisfactory, and the lengths are comparable with the PERC (Correct) method when the median is used. (See Figure 4.1.) While our bias correction generally improves or has little effect on the coverage probabilities, SilverHake is the exception and its coverage probability, after correction, is fairly low. From the preliminary results, the NONPAR method

(using the mean) appears to be a potentially useful method but further results are needed. These results, along with the results for the PERC (Correct) and Large Sample (Case 1) methods, are given in Subsection 4.4.3.

PE	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
	Diet 4	0.09	0	0.09	0	0.45	0.09	0.09	0.09
$\bar{p}_k$	No BA	0.90 (0.22)	1.00 (0.13)	1.00 (0.19)	0.65 (0.20)	0.70 (0.22)	1.00 (0.15)	0.90 (0.08)	1.00 (0.13)
	BA	0.90 (0.20)	1.00 (0.10)	0.80 (0.19)	0.80 (0.13)	0.80 (0.22)	0.70 (0.13)	0.85 (0.08)	0.85 (0.13)
m_k	No BA	1.00 (0.52)	1.00 (0.44)	1.00 (0.51)	1.00 (0.42)	1.00 (0.60)	1.00 (0.35)	1.00 (0.24)	1.00 (0.28)
	BA	1.00 (0.40)	0.80 (0.35)	1.00 (0.41)	1.00 (0.32)	1.00 (0.48)	1.00 (0.28)	1.00 (0.19)	1.00 (0.23)

Table 4.6: Diet 4, AIT, $n_s = 10$: Coverage probabilities (average lengths) of CIs computed using NONPAR method. PE denotes "Point Estimator". (Based on 20 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients).

4.4.3 Results

We now present additional results for our selected CI methods. We begin with results from the PERC (Correct) method, followed by the the NONPAR method and finally the large sample (Case 1) methods. For all of the results to be presented, the coverage probabilities and average lengths are based on $M = 100$ samples of pseudo-seals (generated as in the previous subsection and in Appendix B). The true coverage probability should be within $\pm 1.96\sqrt{\frac{(0.90)(0.10)}{100}} \approx 0.06$) (with 95% confidence) of our computed coverage probability.

Tables 4.7-4.10 and Figures 4.2 -4.3 contain the coverage probabilities and average lengths of CIs computed using the PERC (Correct) method. (For the remainder of the chapter we will usually drop "Correct" when denoting these intervals.) Results are shown for both diets (Diet 1 and Diet 4), both distance measures (AIT and KL) and sample sizes $n_s = 1, 5, 10$, and 25. While we would have liked to have examined sample sizes larger than 25 as well, the slowness of the method is related to the size of the sample and even at $n_s = 25$ the program ran fairly slowly.

Notice that two bias adjustment results are given namely BA 1 and BA 2. The

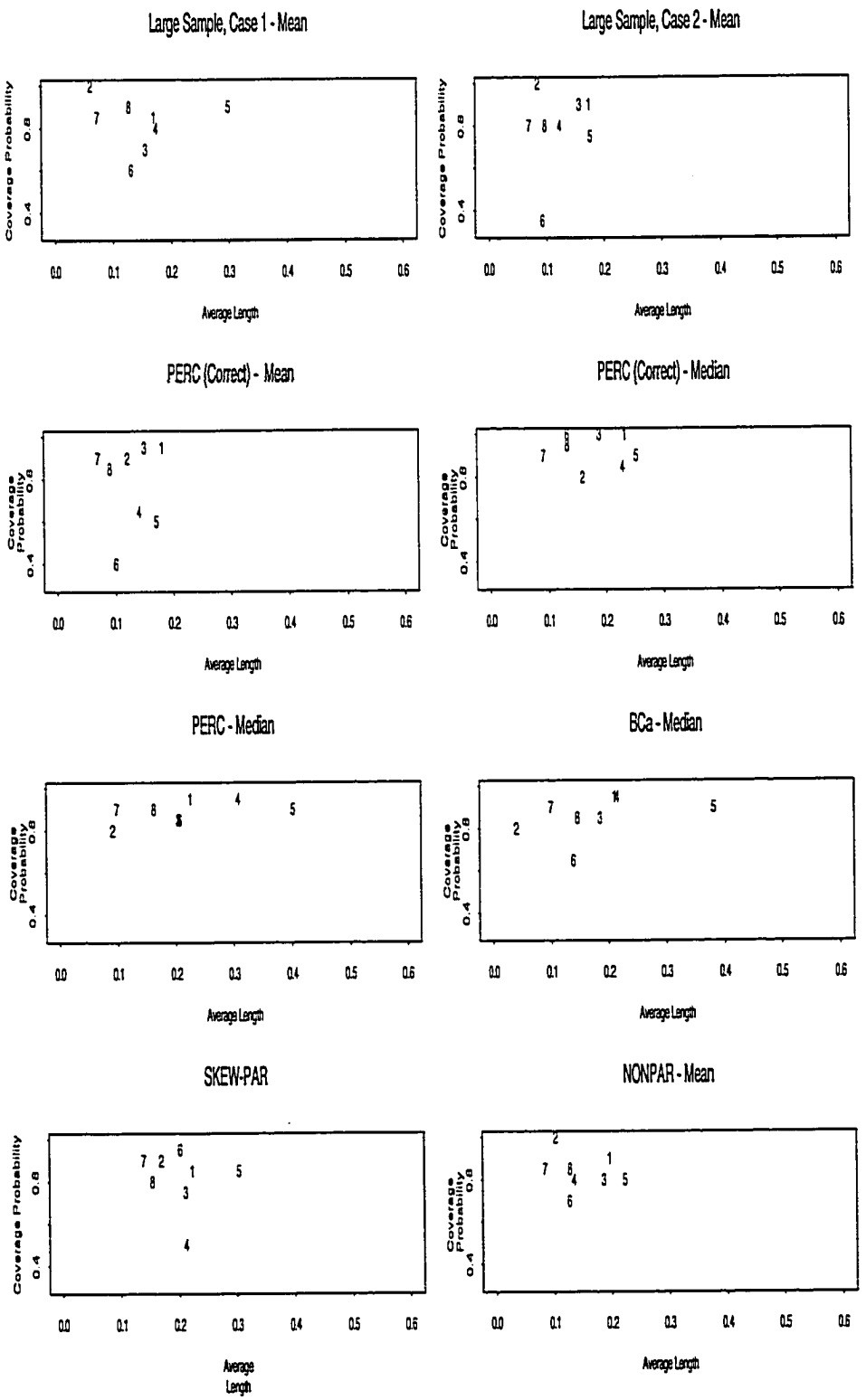

Figure 4.1: Diet 4, AIT, $n_s = 10$: Plots of coverage probabilities versus average lengths for selected (bias adjusted) preliminary results. 1 ≡ Cod, 2 ≡ Haddock, 3 ≡ Plaice, 4 ≡ Pollock, 5 ≡ Sandlance, 6 ≡ SilverHake, 7 ≡ WinterFlounder, 8 ≡ YellowTail.

bias adjustment labeled BA 1 is simply the bias correction that we have discussed and used thus far. It was noticed, however, that often when $\pi_k = 0$, this bias correction worsened the coverage probability. The reason is that with the median as our point estimator, when $\pi_k = 0$, the CIs (without the bias correction) often (correctly) have zero as the lower limit. If the estimate of bias is even slightly in the wrong direction, the shifted interval no longer includes the true diet of zero. This issue appears to be more crucial for the smaller sample sizes (roughly $n_s \leq 10$) though even at $n_s = 25$, Plaice (Diet 1) is problematic with BA 1. (The problem with Plaice could be that its true contribution is more than zero due to the 10% noise that was used.) When $\pi_k \neq 0$, either the bias correction estimate is more often in the correct direction or shifting in the wrong direction does not usually cause the resulting interval to no longer include π_k. To help with this bias correction problem, we decided to only shift the lower limit of an interval if the lower limit was not zero. The results using this bias adjustment are denoted by BA 2. Note that the upper limits were shifted as before and there is consequently the potential for the interval lengths to become wider. From our results, when the average interval length did increase, the increase was only slight.

Consider first Tables 4.7 and 4.8 where the results are based on the AIT distance measure. (See also the graphical display in Figure 4.2.) At $n_s = 1$ the coverage probabilities are fairly good (with BA 2) though the coverage probability for Sandlance in Diet 4 is slightly low. Note that except for Sandlance (Diet 4) a bias correction does not appear to be needed when $n_s = 1$. This is no doubt due to the relatively long lengths of the CIs at $n_s = 1$. Observe that the intervals for Haddock in Diet 1 ($\pi_k = 0.30$) are the widest but the intervals are also wide for Cod in Diet 1 ($\pi_k = 0.30$) and Sandlance in Diet 4 ($\pi_k = 0.45$). In general, it appears to be the case that longer CIs are associated with larger true diet proportions.

At $n_s = 5$ the lengths of the CIs are noticeably shorter than at $n_s = 1$ and the coverage probabilities (with BA 2) are at least as high. Note that the coverage probability for Sandlance (Diet 4) increased from 0.72 to 0.91. By $n_s = 10$ the lengths of the intervals are more reasonable and the coverage probabilities are at least 0.85 for all species. At $n_s = 25$, except for SilverHake and Pollock (both diets), the coverage

probabilities generally stayed the same or increased and are all roughly 0.90 or higher. Observe that without a bias correction, as n_s increased the coverage probability for Pollock decreased. This is not surprising since, as previously mentioned, the bias for Pollock (with the AIT distance) was found to be large compared to the other species. (See Figures 3.1 and 3.2.) As n_s increased, we obtained tighter intervals for the median rather than for the true contribution of Pollock and these are apparently very different. For both diets, the bias correction greatly improved the coverage probabilities for Pollock but, for Diet 4, appeared to underestimate the true bias. Again, the problem could be related to the 10% noise that was used. For SilverHake, the reason for the decrease in the coverage probability as n_s increases (with a bias adjustment) is not clear. Note that in Diet 4 in particular, the bias estimate appears to be the problem since, at $n_s = 25$, the coverage probability without adjustment is 0.99 and 0.79 afterwards. As will be seen, Pollock and SilverHake are problematic in all of our methods.

When the KL distance was used instead (see Tables 4.9-4.10 and Figure 4.3), the major differences were the very low coverage probabilities for Sandlance in Diet 1 (with and without a bias correction) and for SilverHake in Diet 4 (with a bias adjustment). The reason for the coverage problem with Sandlance is not at all obvious since the bias was found to be zero for the median and KL distance in Section 3.3 (see Figure 3.1). While from Figure 3.2, for SilverHake (with the median and KL distance), there appears to be a large bias, the coverage before the bias adjustment is surprisingly greater than 0.90 for all n_s.

Due to time constraints we have not carried out further simulations with the KL distance measure. For the PERC method used with the median and our reduced prey base of eight species, the AIT distance appears to be the more favourable distance measure but this may not always be the case. In Chapter 5 where a goodness of fit statistic is investigated, a larger prey base is used and the KL distance measure performed much better than the AIT distance measure.

The NONPAR method results are given in Tables 4.11 and 4.12 and in Figure 4.4. We chose to only carry out simulations for $n_s = 1$ and $n_s = 10$ because this method was slow to run (even in Fortran) and at these sample sizes the results were generally

	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
n_s	Diet 1	0.30	0.30	0	0.15	0	0.15	0	0
1	No BA	0.9 (0.47)	0.98 (0.53)	1.00 (0.25)	0.86 (0.46)	0.95 (0.22)	0.99 (0.33)	1.00 (0.11)	0.99 (0.15)
	BA 1	0.87 (0.46)	0.93 (0.53)	0.6 (0.24)	0.87 (0.43)	0.68 (0.22)	0.97 (0.32)	0.62 (0.11)	0.63 (0.15)
	BA 2	0.87 (0.48)	0.93 (0.60)	1.00 (0.26)	0.87 (0.43)	0.94 (0.24)	0.99 (0.38)	1.00 (0.12)	0.99 (0.17)
5	No BA	1.00 (0.43)	0.95 (0.44)	1.00 (0.18)	0.94 (0.40)	0.99 (0.18)	0.95 (0.26)	1.00 (0.08)	1.00 (0.12)
	BA 1	0.99 (0.36)	0.95 (0.38)	0.63 (0.15)	0.97 (0.32)	0.65 (0.16)	0.94 (0.22)	0.85 (0.07)	0.72 (0.10)
	BA 2	0.99 (0.37)	0.95 (0.44)	1.00 (0.16)	0.97 (0.32)	0.99 (0.17)	0.94 (0.27)	1.00 (0.07)	1.00 (0.11)
10	No BA	0.99 (0.32)	0.98 (0.37)	1.00 (0.14)	0.83 (0.31)	1.00 (0.14)	0.92 (0.20)	1.00 (0.06)	1.00 (0.092)
	BA 1	0.98 (0.27)	1.00 (0.32)	0.62 (0.12)	0.93 (0.26)	0.68 (0.12)	0.86 (0.17)	0.84 (0.05)	0.73 (0.08)
	BA 2	0.98 (0.27)	1.00 (0.34)	1.00 (0.12)	0.93 (0.26)	1.00 (0.13)	0.86 (0.21)	1.00 (0.05)	1.00 (0.08)
25	No BA	0.99 (0.26)	0.98 (0.31)	1.00 (0.12)	0.62 (0.25)	0.99 (0.12)	0.83 (0.17)	1.00 (0.05)	1.00 0.075
	BA 1	1.00 (0.21)	1.00 (0.26)	0.67 (0.10)	0.89 (0.20)	0.92 (0.10)	0.84 (0.14)	0.94 (0.04)	0.94 0.061
	BA 2	1.00 (0.21)	1.00 (0.26)	1.00 (0.10)	0.89 (0.20)	1.00 (0.10)	0.84 (0.18)	1.00 (0.04)	1.00 0.062

Table 4.7: Diet 1, AIT: Coverage probabilities (average lengths) of CIs computed using PERC method and m_k. (Based on 100 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients).

	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
n_s	Diet 4	0.09	0	0.09	0	0.45	0.09	0.09	0.09
1	No BA	0.98 (0.41)	1.00 (0.35)	0.99 (0.34)	0.80 (0.43)	0.64 (0.42)	1.00 (0.28)	0.98 (0.16)	0.97 (0.22)
	BA 1	0.96 (0.38)	0.72 (0.34)	0.94 (0.33)	0.68 (0.40)	0.72 (0.42)	0.79 (0.27)	0.98 (0.16)	0.86 (0.22)
	BA 2	0.96 (0.39)	1.00 (0.37)	0.95 (0.36)	0.81 (0.40)	0.72 (0.45)	1.00 (0.32)	0.98 (0.18)	0.91 (0.25)
5	No BA	1.00 (0.37)	1.00 (0.26)	1.00 (0.27)	0.94 (0.39)	0.86 (0.43)	0.96 (0.22)	0.99 (0.15)	1.00 (0.20)
	BA 1	1.00 (0.28)	0.83 (0.21)	0.97 (0.21)	0.90 (0.29)	0.91 (0.34)	0.89 (0.18)	0.98 (0.12)	0.97 (0.16)
	BA 2	1.00 (0.29)	1.00 (0.22)	0.97 (0.23)	0.95 (0.29)	0.91 (0.35)	0.95 (0.20)	0.98 (0.13)	0.97 (0.19)
10	No BA	0.99 (0.31)	1.00 (0.20)	1.00 (0.23)	0.77 (0.32)	0.84 (0.32)	0.97 (0.17)	0.94 (0.12)	0.99 (0.17)
	BA 1	0.99 (0.23)	0.82 (0.16)	0.99 (0.18)	0.85 (0.23)	0.85 (0.25)	0.86 (0.13)	0.92 (0.09)	0.96 (0.13)
	BA 2	0.99 (0.24)	1.00 (0.16)	0.99 (0.19)	0.85 (0.23)	0.85 (0.25)	0.87 (0.15)	0.92 (0.09)	0.96 (0.14)
25	No BA	1.00 (0.28)	1.00 (0.17)	1.00 (0.19)	0.66 (0.28)	0.89 (0.26)	0.99 (0.15)	1.00 (0.09)	1.00 (0.14)
	BA 1	1.00 (0.20)	0.99 (0.13)	1.00 (0.14)	0.82 (0.19)	0.89 (0.19)	0.79 (0.11)	0.94 (0.07)	0.94 (0.10)
	BA 2	1.00 (0.20)	1.00 (0.13)	1.00 (0.15)	0.82 (0.19)	0.89 (0.19)	0.79 (0.12)	0.94 (0.07)	0.94 (0.11)

Table 4.8: Diet 4, AIT: Coverage probabilities (average lengths) of CIs computed using PERC method and m_k. (Based on 100 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients).

	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
n_s	Diet 1	0.30	0.30	0	0.15	0	0.15	0	0
1	No BA	0.84	0.95	1.00	0.99	0.82	0.96	1.00	0.97
		(0.42)	(0.58)	(0.25)	(0.32)	(0.22)	(0.28)	(0.11)	(0.18)
	BA 1	0.88	0.85	0.82	0.87	0.59	0.84	0.77	0.56
		(0.42)	(0.58)	(0.25)	(0.31)	(0.21)	(0.28)	(0.10)	(0.17)
	BA 2	0.88	0.85	1.00	0.87	0.80	0.87	1.00	0.95
		(0.461)	(0.60)	(0.26)	(0.34)	(0.22)	(0.31)	(0.11)	(0.19
5	No BA	0.97	0.98	1.00	0.99	0.84	0.97	1.00	1.00
		(0.40)	(0.54)	(0.19)	(0.29)	(0.22)	(0.28)	(0.09)	(0.14)
	BA 1	0.93	0.97	0.96	0.91	0.47	0.82	0.79	0.70
		(0.33)	(0.45)	(0.15)	(0.24)	(0.18)	(0.23)	(0.07)	(0.11)
	BA 2	0.93	0.97	1.00	0.91	0.83	0.82	1.00	1.00
		(0.36)	(0.46)	(0.15)	(0.26)	(0.20)	(0.26)	(0.07)	(0.12)
10	No BA	0.94	1.00	1.00	0.98	0.62	0.97	1.00	0.99
		(0.32)	(0.42)	(0.14)	(0.23)	(0.16)	(0.22)	(0.07)	(0.11)
	BA 1	0.94	0.97	0.98	0.85	0.35	0.84	0.95	0.61
		(0.26)	(0.34)	(0.12)	(0.19)	(0.13)	(0.18)	(0.05)	(0.09)
	BA 2	0.94	0.97	1.00	0.85	0.54	0.84	1.00	0.99
		(0.26)	(0.34)	(0.12)	(0.19)	(0.14)	(0.20)	(0.05)	(0.10)
25	No BA	0.93	0.99	1.00	1.00	0.27	1.00	1.00	1.00
		(0.26)	(0.34)	(0.12)	(0.20)	(0.13)	(0.19)	(0.06)	(0.10)
	BA 1	0.94	0.98	1.00	0.94	0.24	0.85	1.00	0.79
		(0.20)	(0.26)	(0.09)	(0.15)	(0.10)	(0.15)	(0.04)	(0.07)
	BA 2	0.94	0.98	1.00	0.94	0.28	0.85	1.00	1.00
		(0.20)	(0.26)	(0.09)	(0.15)	(0.10)	(0.16)	(0.04)	(0.08)

Table 4.9: Diet 1, KL: Coverage probabilities (average lengths) of CIs computed using PERC method and m_k. (Based on 100 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients).

	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
n_s	Diet 4	0.09	0	0.09	0	0.45	0.09	0.09	0.09
1	No BA	1.00 (0.33)	0.94 (0.42)	1.00 (0.40)	0.88 (0.30)	0.53 (0.34)	0.93 (0.22)	0.98 (0.17)	0.99 (0.24)
	BA 1	0.92 (0.32)	0.54 (0.41)	0.79 (0.40)	0.40 (0.29)	0.56 (0.34)	0.75 (0.22)	0.96 (0.17)	0.85 (0.23)
	BA 2	0.95 (0.34)	0.94 (0.44)	0.98 (0.45)	0.87 (0.32)	0.56 (0.34)	0.91 (0.25)	0.96 (0.19)	0.96 (0.27)
5	No BA	1.00 (0.29)	1.00 (0.37)	1.00 (0.36)	0.98 (0.26)	0.86 (0.40)	0.95 (0.18)	0.99 (0.16)	1.00 (0.21)
	BA 1	0.99 (0.23)	0.64 (0.28)	0.90 (0.28)	0.63 (0.20)	0.84 (0.31)	0.81 (0.14)	0.97 (0.12)	0.98 (0.17)
	BA 2	0.99 (0.24)	1.00 (0.30)	0.98 (0.31)	0.98 (0.22)	0.84 (0.31)	0.88 (0.16)	0.97 (0.14)	0.99 (0.19)
10	No BA	1.00 (0.25)	1.00 (0.31)	1.00 (0.32)	0.92 (0.21)	0.78 (0.29)	0.92 (0.14)	0.96 (0.13)	1.00 (0.19)
	BA 1	0.98 (0.18)	0.70 (0.23)	0.92 (0.24)	0.57 (0.15)	0.79 (0.21)	0.61 (0.10)	0.90 (0.10)	0.99 (0.14)
	BA 2	0.98 (0.20)	1.00 (0.24)	0.93 (0.27)	0.92 (0.16)	0.79 (0.21)	0.62 (0.12)	0.90 (0.10)	0.99 (0.15)
25	No BA	1.00 (0.23)	1.00 (0.28)	1.00 (0.29)	0.89 (0.21)	0.77 (0.26)	0.91 (0.13)	0.99 (0.11)	1.00 (0.17)
	BA 1	1.00 (0.16)	0.82 (0.19)	0.97 (0.20)	0.73 (0.14)	0.85 (0.17)	0.35 (0.09)	0.90 (0.08)	0.97 (0.12)
	BA 2	1.00 (0.17)	1.00 (0.19)	0.97 (0.22)	0.88 (0.14)	0.85 (0.17)	0.35 (0.09)	0.90 (0.08)	0.97 (0.12)

Table 4.10: Diet 4, KL: Coverage probabilities (average lengths) of CIs computed using PERC method and m_k. (Based on 100 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients).

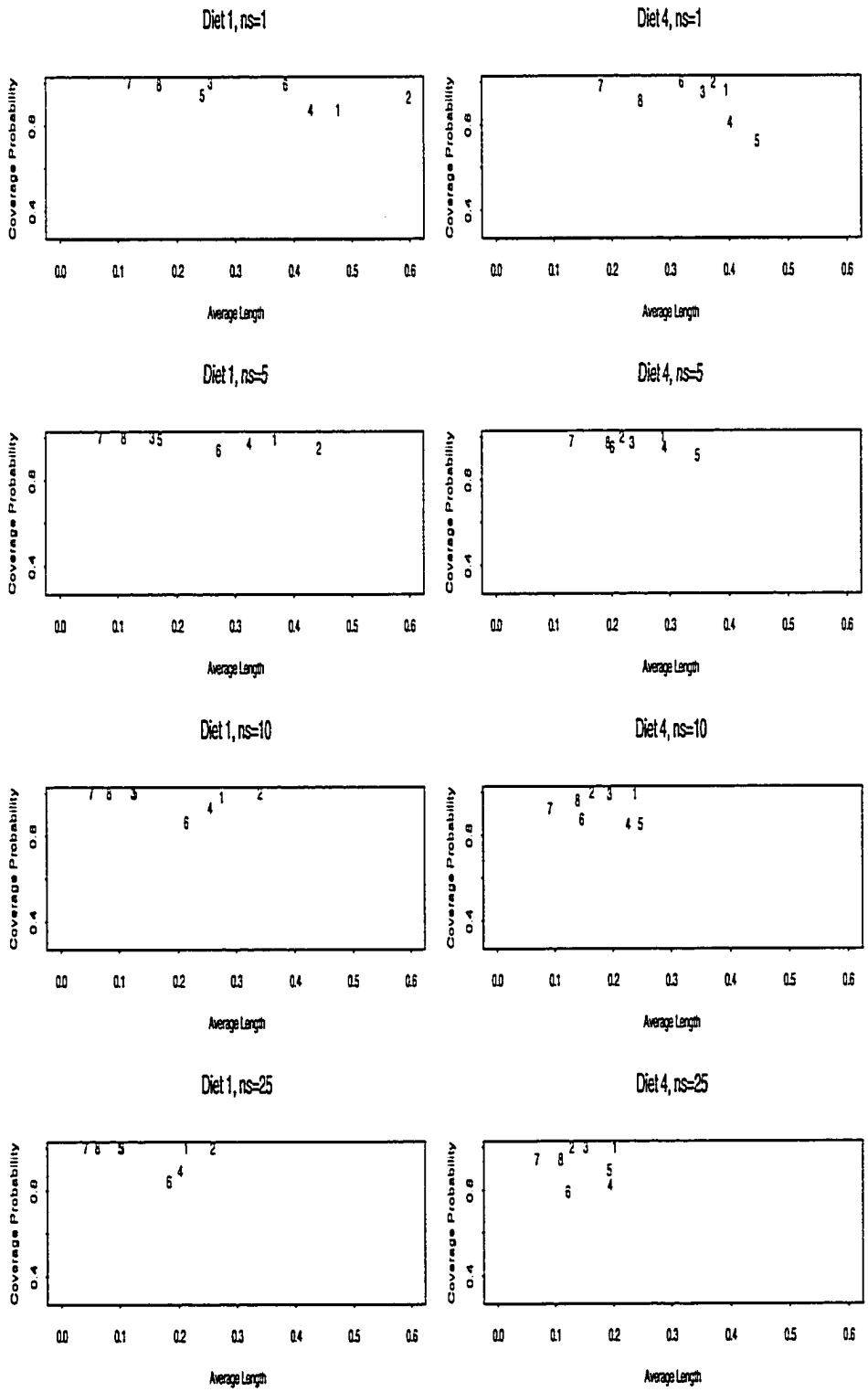

Figure 4.2: PERC (Correct), Median, AIT: Plots of coverage probabilities versus average lengths for (BA 2) results. 1 ≡ Cod, 2 ≡ Haddock, 3 ≡ Plaice, 4 ≡ Pollock, 5 ≡ Sandlance, 6 ≡ SilverHake, 7 ≡ WinterFlounder, 8 ≡ YellowTail.

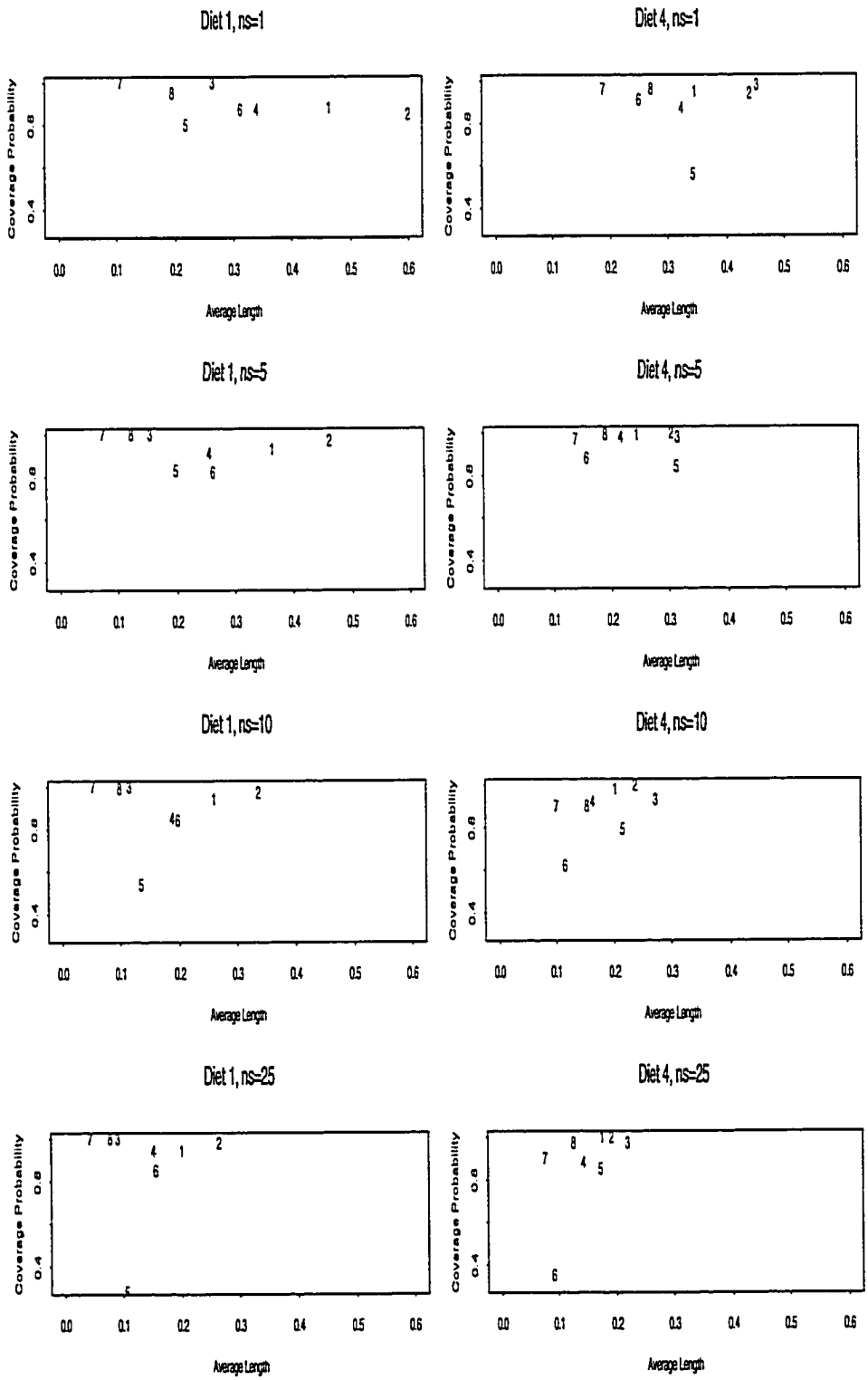

Figure 4.3: PERC (Correct), Median, KL: Plots of coverage probabilities versus average lengths for (BA 2) results. 1 ≡ Cod, 2 ≡ Haddock, 3 ≡ Plaice, 4 ≡ Pollock, 5 ≡ Sandlance, 6 ≡ SilverHake, 7 ≡ WinterFlounder, 8 ≡ YellowTail.

not as good as with the PERC method, particularly in terms of coverage. While for Diet 1 the results are satisfactory with either bias adjustment, for Diet 4 the coverage probabilities are low for Pollock, Sandlance, SilverHake and YellowTail when the bias correction is used.

	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
n_s	Diet 1	0.30	0.30	0	0.15	0	0.15	0	0
1	No BA	0.83	0.8	0.99	0.84	0.89	0.89	0.98	0.97
		(0.48)	(0.48)	(0.31)	(0.42)	(0.25)	(0.49)	(0.13)	(0.20)
	BA 1	0.73	0.70	0.83	0.62	0.76	0.74	0.82	0.71
		(0.46)	(0.47)	(0.29)	(0.36)	(0.22)	(0.47)	(0.12)	(0.18)
	BA 2	0.73	0.70	0.98	0.62	0.89	0.74	0.98	0.97
		(0.47)	(0.50)	(0.29)	(0.36)	(0.22)	(0.48)	(0.12)	(0.18)
10	No BA	0.99	0.99	1.00	0.69	0.89	0.94	0.99	0.96
		(0.24)	(0.30)	(0.11)	(0.21)	(0.11)	(0.20)	(0.05)	(0.08)
	BA 1	0.91	0.92	0.99	0.79	0.90	0.79	1.00	0.94
		(0.24)	(0.30)	(0.08)	(0.20)	(0.08)	(0.19)	(0.04)	(0.06)
	BA 2	0.91	0.92	1.00	0.79	0.90	0.79	1.00	0.96
		(0.24)	(0.31)	(0.08)	(0.20)	(0.08)	(0.19)	(0.04)	(0.06)

Table 4.11: Diet 1, AIT: Coverage probabilities (average lengths) of CIs computed using NONPAR method and $\bar{p}_k$. (Based on 100 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients).

Finally, Tables 4.13 and 4.14 contain the large sample (Case 1) results which use the sample mean as the point estimator. These results are also displayed in Figure 4.5. Recall that these intervals require no bootstrapping and are very simple to compute. They do not, however, take into account the variability due to the prey and cannot be used when $n_s = 1$. We are primarily interested in determining whether using the more computationally intensive PERC method is worthwhile for $n_s \geq 10$. Note that we have only showed the BA 1 adjusted results as the BA 2 results were almost identical.

At $n_s = 10$, the t intervals, without a bias correction, are giving good coverage probabilities and generally performing better than the Normal intervals. When the bias correction is used, there is an improvement in the coverage probability for Pollock (Diet 4) but a decrease in the coverage probability for SilverHake. The lengths are comparable to the PERC method intervals but are longer for Sandlance (Diet 4). The

	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
n_s	Diet 4	0.09	0	0.09	0	0.45	0.09	0.09	0.09
1	No BA	0.94 (0.41)	0.97 (0.38)	0.95 (0.39)	0.79 (0.38)	0.71 (0.39)	0.84 (0.33)	0.95 (0.18)	0.92 (0.23)
	BA 1	0.89 (0.36)	0.77 (0.34)	0.90 (0.37)	0.76 (0.31)	0.65 (0.39)	0.79 (0.30)	0.75 (0.18)	0.77 (0.21)
	BA 2	0.89 (0.37)	0.97 (0.36)	0.90 (0.38)	0.79 (0.31)	0.65 (0.39)	0.79 (0.30)	0.75 (0.18)	0.77 (0.23)
10	No BA	0.92 (0.22)	1.00 (0.14)	0.97 (0.18)	0.63 (0.20)	0.75 (0.23)	0.93 (0.15)	0.90 (0.08)	0.94 (0.13)
	BA 1	0.92 (0.19)	0.97 (0.12)	0.87 (0.18)	0.77 (0.13)	0.78 (0.23)	0.62 (0.12)	0.90 (0.08)	0.76 (0.12)
	BA 2	0.92 (0.19)	1.00 (0.12)	0.87 (0.18)	0.77 (0.13)	0.78 (0.23)	0.62 (0.12)	0.90 (0.08)	0.76 (0.13)

Table 4.12: Diet 4, AIT: Coverage probabilities (average lengths) of CIs computed using NONPAR method and $\bar{p}_k$. (Based on 100 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients).

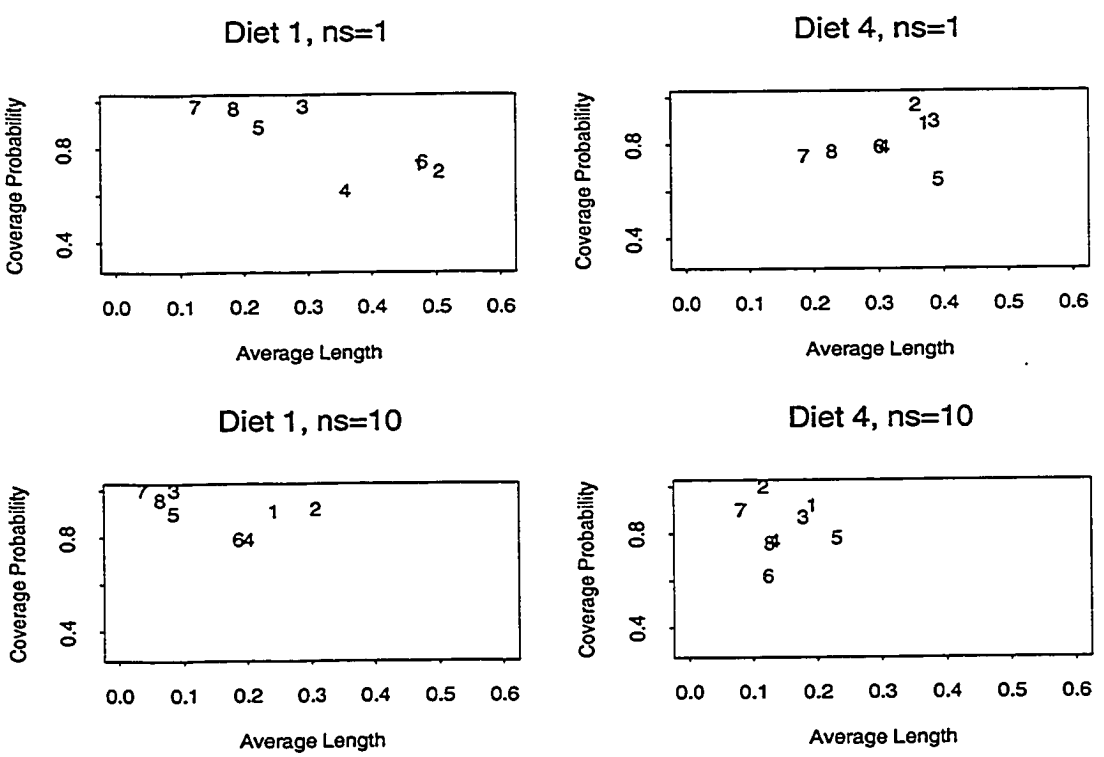

Figure 4.4: NONPAR, Mean, AIT: Plots of coverage probabilities versus average lengths for (BA 2) results. 1 ≡ Cod, 2 ≡ Haddock, 3 ≡ Plaice, 4 ≡ Pollock, 5 ≡ Sandlance, 6 ≡ SilverHake, 7 ≡ WinterFlounder, 8 ≡ YellowTail,

$SMixM$ intervals appear to give somewhat decent coverage probabilities (at $n_s = 10$) provided the bias adjustment is used. For Plaice in Diet 1 and for Pollock in Diet 4, the coverage probability with the $SMixM$ method is still a little low, however, even after the adjustment.

As n_s increases, none of the large sample methods yield coverage probabilities that are consistently good for all species. The trouble appears to be with Pollock and SilverHake (both Diets) and with YellowTail in Diet 4. With the $SMixM$ method, Plaice (Diet 1) is also problematic. Note that with the Normal and t intervals (which are similar for larger n_s), the problem with Pollock is only evident at $n_s = 50$. As before, for Pollock, the bias correction often greatly improves the coverage probability but not enough. For SilverHake, the bias correction tends to worsen the coverage probability.

4.4.4 Recommendations

Overall the recommended method is the PERC method in which both the seals and prey are re-sampled and the median is the point estimator. This method is simple to implement but can be time consuming to run when n_s is large. For this method we recommend using the BA 2 adjustment. With the AIT distance, for $n_s = 1$, the coverage probabilities were generally good with or without the adjustment but the intervals could be a bit long particularly when $\pi_k \geq 0.15$. For $5 \leq n_s \leq 10$ and with the AIT distance, the method worked very well with the BA 2 adjustment. Although for $n_s = 25$ the coverage probabilities and lengths were fairly good (with the AIT distance), there is some concern that the coverage probabilities for Pollock and SilverHake (with either BA correction) may worsen as n_s increases. Although for Pollock the problem could be a poor estimate of the bias, in Diet 4 it could also be due to the 10% noise that was used. For SilverHake, the problem appeared to be related to the bias estimate.

Except for Pollock and SilverHake (and YellowTail with Diet 4), a normal approximation without bootstrapping the prey worked fairly well for $n_s \geq 25$. If time is an issue, for $n_s \geq 25$, these intervals may be adequate.

It should be mentioned that the parametric CIs may have some potential if they

	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
Type	Diet 1	0.30	0.30	0	0.15	0	0.15	0	0
Normal $n_s = 10$	No BA	0.95 (0.26)	0.73 (0.25)	0.89 (0.07)	0.91 (0.29)	0.95 (0.12)	0.83 (0.19)	0.96 (0.03)	0.93 (0.06)
	BA 1	0.89 (0.26)	0.87 (0.25)	0.97 (0.049)	0.93 (0.28)	0.97 (0.09)	0.76 (0.17)	0.97 (0.02)	0.95 (0.05)
Normal $n_s = 25$	No BA	0.99 (0.17)	0.60 (0.16)	0.33 (0.06)	0.62 (0.19)	0.62 (0.09)	0.84 (0.12)	0.8 (0.02)	0.77 (0.05)
	BA 1	0.92 (0.17)	0.88 (0.16)	0.88 (0.04)	0.83 (0.19)	0.98 (0.06)	0.65 (0.12)	0.97 (0.01)	0.99 (0.03)
Normal $n_s = 50$	No BA	0.97 (0.12)	0.26 (0.12)	0 (0.04)	0.12 (0.13)	0.13 (0.06)	0.76 (0.09)	0.28 (0.02)	0.21 (0.03)
	BA 1	0.90 (0.12)	0.91 (0.12)	0.95 (0.03)	0.71 (0.13)	1.00 (0.04)	0.67 (0.09)	1.00 (0.01)	0.97 (0.03)
t $n_s = 10$	No BA	0.98 (0.32)	0.88 (0.31)	0.99 (0.08)	0.96 (0.36)	0.99 (0.14)	0.93 (0.22)	1.00 (0.04)	1.00 (0.07)
	BA 1	0.95 (0.32)	0.93 (0.31)	0.99 (0.06)	0.95 (0.33)	0.99 (0.11)	0.82 (0.21)	1.00 (0.03)	0.96 (0.06)
t $n_s = 25$	No BA	1.00 (0.18)	0.62 (0.18)	0.37 (0.06)	0.67 (0.20)	0.75 (0.09)	0.88 (0.14)	0.86 (0.03)	0.85 (0.05)
	BA 1	0.95 (0.18)	0.92 (0.18)	0.89 (0.04)	0.87 (0.20)	0.98 (0.06)	0.69 (0.13)	0.99 (0.02)	0.99 (0.03)
t $n_s = 50$	No BA	0.98 (0.12)	0.30 (0.12)	0 (0.04)	0.18 (0.14)	0.16 (0.06)	0.78 (0.09)	0.33 (0.02)	0.28 (0.04)
	BA 1	0.91 (0.12)	0.91 (0.12)	0.96 (0.03)	0.73 (0.14)	1.00 (0.04)	0.68 (0.09)	1.00 (0.01)	0.98 (0.03)
$SMixM$ $n_s = 10$	No BA	0.99 (0.34)	0.85 (0.30)	0.57 (0.14)	0.93 (0.33)	0.75 (0.23)	0.99 (0.25)	0.86 (0.10)	0.79 (0.16)
	BA 1	0.93 (0.34)	0.94 (0.30)	0.75 (0.13)	0.92 (0.32)	0.82 (0.22)	0.94 (0.24)	0.90 (0.10)	0.83 (0.15)
$SMixM$ $n_s = 25$	No BA	0.97 (0.20)	0.49 (0.18)	0.14 (0.05)	0.85 (0.20)	0.28 (0.08)	0.78 (0.12)	0.67 (0.02)	0.59 (0.05)
	BA 1	0.90 (0.20)	0.92 (0.18)	0.71 (0.04)	0.89 (0.20)	0.91 (0.07)	0.71 (0.12)	0.92 (0.02)	0.91 (0.04)
$SMixM$ $n_s = 50$	No BA	0.92 (0.14)	0.09 (0.12)	0.03 (0.03)	0.6 (0.14)	0.13 (0.05)	0.32 (0.08)	0.43 (0.01)	0.23 (0.03)
	BA 1	0.90 (0.14)	0.92 (0.12)	0.72 (0.02)	0.72 (0.14)	0.97 (0.04)	0.69 (0.08)	0.98 (0.01)	0.91 (0.02)

Table 4.13: Diet 1, AIT: Coverage probabilities (average lengths) of CIs computed using large sample methods (Case 1). (Based on 100 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients).

	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
Type	Diet 1	0.09	0	0.09	0	0.45	0.09	0.09	0.09
Normal $n_s = 10$	No BA	0.96 (0.19)	0.97 (0.10)	0.95 (0.15)	0.63 (0.25)	0.81 (0.30)	0.92 (0.16)	0.70 (0.07)	0.88 (0.13)
	BA 1	0.91 (0.18)	0.96 (0.07)	0.83 (0.15)	0.83 (0.17)	0.86 (0.30)	0.62 (0.13)	0.85 (0.07)	0.80 (0.13)
Normal $n_s = 25$	No BA	0.87 (0.13)	0.65 (0.08)	0.96 (0.09)	0 (0.18)	0.55 (0.20)	0.96 (0.12)	0.51 (0.05)	0.83 (0.08)
	BA 1	0.90 (0.13)	1.00 (0.05)	0.82 (0.09)	0.80 (0.13)	0.96 (0.20)	0.53 (0.10)	0.86 (0.05)	0.71 (0.08)
Normal $n_s = 50$	No BA	0.72 (0.09)	0.08 (0.06)	0.97 (0.07)	0 (0.13)	0.27 (0.14)	0.94 (0.08)	0.18 (0.03)	0.73 (0.06)
	BA 1	0.91 (0.09)	0.97 (0.04)	0.87 (0.07)	0.66 (0.11)	0.96 (0.14)	0.35 (0.08)	0.92 (0.03)	0.58 (0.06)
t $n_s = 10$	No BA	1.00 (0.23)	0.99 (0.11)	0.95 (0.18)	0.80 (0.29)	0.94 (0.37)	0.94 (0.19)	0.85 (0.09)	0.93 (0.16)
	BA 1	0.99 (0.21)	0.98 (0.08)	0.93 (0.18)	0.96 (0.21)	0.94 (0.37)	0.75 (0.15)	0.93 (0.09)	0.85 (0.15)
t $n_s = 25$	No BA	0.93 (0.14)	0.75 (0.08)	0.96 (0.10)	0.02 (0.19)	0.61 (0.21)	0.96 (0.13)	0.56 (0.05)	0.88 (0.09)
	BA 1	0.93 (0.14)	1.00 (0.06)	0.86 (0.10)	0.85 (0.14)	0.98 (0.21)	0.57 (0.10)	0.89 (0.05)	0.74 (0.09)
t $n_s = 50$	No BA	0.76 (0.09)	0.08 (0.06)	0.98 (0.07)	0 (0.13)	0.30 (0.14)	0.96 (0.09)	0.25 (0.03)	0.79 (0.06)
	BA 1	0.93 (0.09)	0.98 (0.05)	0.87 (0.07)	0.66 (0.11)	0.97 (0.14)	0.38 (0.08)	0.93 (0.03)	0.59 (0.06)
$SMixM$ $n_s = 10$	No BA	0.99 (0.24)	0.78 (0.21)	0.99 (0.19)	0.23 (0.31)	0.93 (0.38)	0.98 (0.25)	0.86 (0.10)	0.95 (0.18)
	BA 1	0.89 (0.23)	0.91 (0.19)	0.83 (0.19)	0.78 (0.27)	0.89 (0.38)	0.88 (0.23)	0.94 (0.10)	0.86 (0.18)
$SMixM$ $n_s = 25$	No BA	0.94 (0.13)	0.27 (0.08)	0.91 (0.09)	0 (0.16)	0.57 (0.23)	0.93 (0.11)	0.34 (0.05)	0.70 (0.08)
	BA 1	0.85 (0.13)	0.85 (0.07)	0.8 (0.09)	0.67 (0.14)	0.91 (0.23)	0.6 (0.10)	0.92 (0.05)	0.75 (0.08)
$SMixM$ $n_s = 50$	No BA	0.98 (0.08)	0.06 (0.05)	0.77 (0.06)	0 (0.11)	0.09 (0.16)	0.73 (0.07)	0 (0.04)	0.21 (0.05)
	BA 1	0.87 (0.08)	0.84 (0.04)	0.79 (0.06)	0.49 (0.10)	0.94 (0.16)	0.30 (0.07)	0.90 (0.04)	0.54 (0.05)

Table 4.14: Diet 4, AIT: Coverage probabilities (average lengths) of CIs computed using large sample methods (Case 1). (Based on 100 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients).

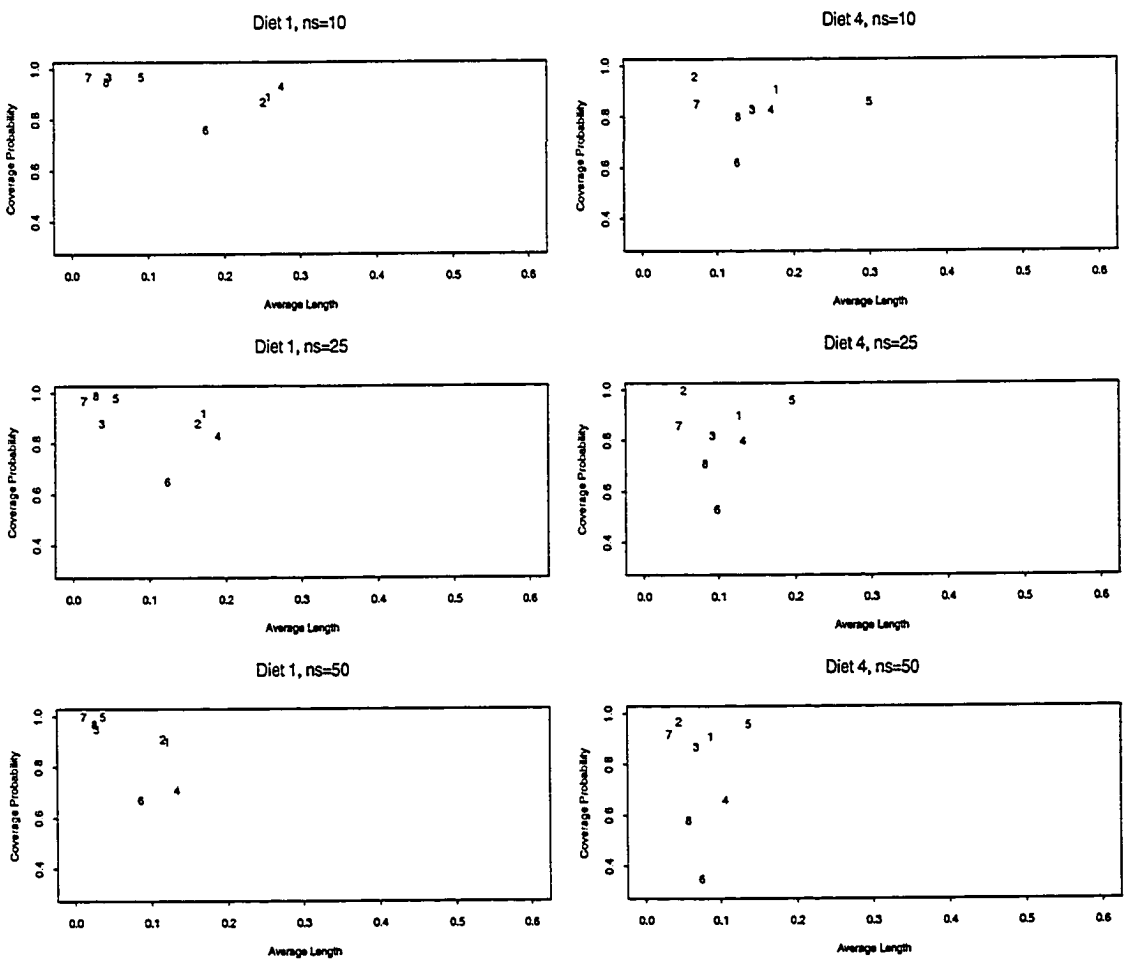

Figure 4.5: Large Sample (Case 1), Normal, AIT: Plots of coverage probabilities versus average lengths for (BA 1) results. 1 ≡ Cod, 2 ≡ Haddock, 3 ≡ Plaice, 4 ≡ Pollock, 5 ≡ Sandlance, 6 ≡ SilverHake, 7 ≡ WinterFlounder, 8 ≡ YellowTail.

were faster to compute. Since we used S-PLUS we were forced to use small bootstrap parameters and this may have affected the results.

As a final comment, realize that we used a fairly small number ($M = 100$) of samples of pseudo-seals and this would have to be increased if more clear-cut conclusions about the coverage probabilities and average lengths of the intervals from the various methods are desired.

4.5 Other Issues Relating to Interval Estimation

In this section, two issues specific to the diet estimation problem and which affect interval estimation are discussed. We first address the issue of a potential difference between real seals (that is, seals in the wild) and our pseudo-seals, and subsequently the issue of incorporating the fat content of the prey into our estimates.

4.5.1 Real Seals Versus Pseudo-Seals

As briefly discussed in Section 3.3, the diet estimates obtained using real seals may be more variable than those obtained using pseudo-seals. Recall that one reason for this difference is that seals sampled in the wild may not have identical diets. In our simulations we generated samples of pseudo-seals from a common diet and attempted to estimate this diet, thus eliminating this additional source of variability. If it is reasonable to assume that seals in a specific region of interest have long run diets that are approximately the same then this source of variability should not be a major concern. Otherwise, to obtain approximate coverage probabilities when our methods are used to estimate the true average diet of the seals, we could generate samples of pseudo-seals in such a way that each seal in the sample has a slightly different diet, where the magnitude of this difference would be determined in consultation with a biologist.

Another previously discussed but less straightforward reason for a potentially greater variability in the diet estimates in practice arises because our pseudo-seals simulate the situation where seals are random sampling from the prey. In reality, however, seals tend to cluster sample, as discussed in Section 3.3. A sample of seal FA signatures from the wild and corresponding diet estimates might then be more variable

than those of pseudo-seals and their estimates. While one way to manage this issue is to modify how the pseudo-seals are being generated, a somewhat simpler approach is to adjust the variability in the diet estimates in a sensible manner. One way of carrying out the latter approach is to make use of the various alternative ways of summarizing the prey (that is, other than by the sample means, $\bar{\mathbf{X}}_k$, $k = 1, \ldots, I$) that were discussed in Section 3.2. For example, for our sample of pseudo-seals, we could use the "Random Sampling Method" (RS method) and estimate the diet of the ith seal using a randomly selected prey FA signature from each prey species. Presumably this would generate diet estimates that are more variable than those based on the sample means of the prey. We would expect the "Multivariate Quantile" (MQ) and the "AIT/KL Quantile" (KLQ or AITQ) methods to yield estimates slightly less variable than the RS method but more variable than the MEAN method.

We carried out a small simulation study to examine the effect on the confidence intervals (particularly on their lengths) when these alternative methods of summarizing the prey were used to estimate the diet. We used the percentile (PERC) method with the median aggregate point estimator (our recommended CI method) but now estimated individual diets using a quantile instead of the mean. The algorithm used is similar to the large sample bootstrap algorithm (Subsection 4.3.2) and is given below in detail.

Large Sample Adjusted Bootstrap Algorithm

1. Choose a set of n_{quant} quantiles, say $S = \{s_1, \ldots, s_{n_{\text{quant}}}\}$. (For example, $S = \{0.25, 0.50, 0.75\}$.)

2. For the ith seal and kth species randomly select one of the elements of S and compute the corresponding quantile. Let $\mathbf{Q}_{ik}$ denote this quantile. Repeat for each species and let $\mathbf{Q_i} = (\mathbf{Q}_{i1}, \ldots, \mathbf{Q}_{iI})$, $i = 1, \ldots, n_s$.

3. Compute the n_s estimates of diet: $\mathbf{p}_1(\mathbf{Y}_1, \mathbf{Q}_1), \ldots, \mathbf{p}_{n_s}(\mathbf{Y}_{n_s}, \mathbf{Q}_{n_s})$, $i = 1, \ldots, n_s$.

4. Compute the median of the n_s diet estimates for each species and let $\mathbf{m} = (m_1, \ldots, m_I)$.

5. *for* $r = 1, \ldots, R$

   (a) Generate n_s seals: $\mathbf{Y}_1^{*r}, \ldots, \mathbf{Y}_{n_s}^{*r}$. (See Appendix C.)

   (b) Generate sample of prey: $\mathbf{X}_1^{*r}, \ldots, \mathbf{X}_I^{*r}$. (See Appendix C.)

   (c) Carry out Step 2. (using prey in Step (b)) to obtain $\mathbf{Q}_i^{*r}$, $i = 1, \ldots, n_s$.

   (d) Compute n_s diet estimates: $\mathbf{p}_1^*(\mathbf{Y}_1^{*r}, \mathbf{Q}_1^{*r}), \ldots, \mathbf{p}_{n_s}^*(\mathbf{Y}_{n_s}^{*r}, \mathbf{Q}_{n_s}^{*r})$.

   (e) Compute the median of the diet estimates in Step (d): $\mathbf{m}^{*r}$.

6. Compute PERC confidence intervals.

Before presenting our results, a few comments are needed. Realize that the above algorithm was carried out to gain insight into the coverage probabilities and the lengths of the CIs that we may obtain in practice when the PERC method is used. While at this point we recommend using $\bar{\mathbf{X}}$ to estimate the diet in real-life applications, the above algorithm may be an appropriate way of incorporating a potential source of variability arising from the fact that seals do not consume the mean prey signature but rather a sample of the prey signatures.

We chose to examine the RS and AITQ methods at $n_s = 10$. (The more complex MQ method runs much more slowly than the RS and AITQ methods and we would expect it to yield CIs with lengths similar those produced by the AITQ method.) For the AITQ method, we used the 5th, 25th, 50th, 75th and 95th quantiles. Note that we simply used the previously computed estimates of bias to shift the intervals. (It would be straightforward but computationally time consuming to apply the bias adjustment algorithm in Section 4.2 to the diet estimates based on the RS and AITQ methods.) The results are given in Tables 4.15-4.16.

Using the AITQ and RS methods to summarize the prey produced, as expected, longer intervals on average. The result of the longer intervals was an increase in the coverage probabilities to one almost everywhere. (The only exception was the coverage probability of 0.93 for Sandlance in Diet 4.) The average lengths of the AITQ and RS method intervals were very similar and, surprisingly, most often the RS intervals were slightly shorter. Perhaps using a smaller number of quantiles (say 25th, 50th and 75th) in the AITQ method would have yielded shorter intervals.

The results imply that in practice the lengths of the CIs obtained may be considerably longer than the those obtained using pseudo-seals. The extent of the difference in the lengths will depend on the similarity in the within species FA signatures. More insight into the lengths of CIs obtained in practice is presented in Section 4.6, where the PERC method is applied to some real-life data on captive seabirds. As will be seen, for this application, the PERC method performed very well and yielded CIs of informative lengths.

	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
Method	Diet 1	0.30	0.30	0	0.15	0	0.15	0	0
MEAN	No BA	0.99 (0.32)	0.98 (0.37)	1.00 (0.14)	0.83 (0.31)	1.00 (0.14)	0.92 (0.20)	1.00 (0.06)	1.00 (0.09)
	BA 2	0.98 (0.27)	1.00 (0.34)	1.00 (0.12)	0.93 (0.26)	1.00 (0.13)	0.86 (0.21)	1.00 (0.05)	1.00 (0.08)
AITQ	No BA	1.00 (0.57)	1.00 (0.62)	1.00 (0.48)	1.00 (0.62)	1.00 (0.48)	1.00 (0.58)	1.00 (0.27)	1.00 (0.32)
	BA 2	1.00 (0.48)	1.00 (0.58)	1.00 (0.38)	1.00 (0.46)	1.00 (0.38)	1.00 (0.51)	1.00 (0.22)	1.00 (0.26)
RS	No BA	1.00 (0.53)	1.00 (0.59)	1.00 (0.46)	1.00 (0.56)	1.00 (0.44)	1.00 (0.54)	1.00 (0.27)	1.00 (0.30)
	BA 2	1.00 (0.46)	1.00 (0.57)	1.00 (0.38)	1.00 (0.42)	1.00 (0.36)	1.00 (0.49)	1.00 (0.22)	1.00 (0.25)

Table 4.15: Diet 1, AIT: Coverage probabilities (average lengths) of CIs computed using PERC method and m_k with the MEAN, RS and AITQ methods of summarizing the prey. (Based on 100 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients).

4.5.2 Fat Content

Another issue that emerged in the diet estimation problem and that was briefly mentioned in Chapter 1 relates to the fat content of the species. Associated with each prey is a fat content and species with higher fat contents contribute proportionately more to the seal's signature. In practice, the diet estimates should be adjusted to account for the fat content. Iverson *et al* (2004) recommend using the following adjusted estimate of diet for the kth species

$$a_k = \frac{\frac{p_k}{f_k}}{\sum_{k=1}^{I} \frac{p_k}{f_k}},$$

	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
Method	Diet 4	0.09	0	0.09	0	0.45	0.09	0.09	0.09
MEAN	No BA	0.99 (0.31)	1.00 (0.20)	1.00 (0.23)	0.77 (0.32)	0.84 (0.32)	0.97 (0.17)	0.94 (0.12)	0.99 (0.17)
	BA 2	0.99 (0.24)	1.00 (0.16)	0.99 (0.19)	0.85 (0.23)	0.85 (0.25)	0.87 (0.15)	0.92 (0.09)	0.96 (0.14)
AITQ	No BA	1.00 (0.58)	1.00 (0.60)	1.00 (0.53)	1.00 (0.57)	0.91 (0.55)	1.00 (0.54)	1.00 (0.31)	1.00 (0.33)
	BA 2	1.00 (0.45)	1.00 (0.48)	1.00 (0.44)	1.00 (0.42)	0.98 (0.54)	1.00 (0.44)	1.00 (0.27)	1.00 (0.29)
RS	No BA	1.00 (0.56)	1.00 (0.58)	1.00 (0.52)	1.00 (0.53)	0.79 (0.51)	1.00 (0.52)	1.00 (0.32)	1.00 (0.32)
	BA 2	1.00 (0.43)	1.00 (0.46)	1.00 (0.43)	1.00 (0.39)	0.93 (0.51)	1.00 (0.43)	1.00 (0.28)	1.00 (0.29)

Table 4.16: Diet 4, AIT: Coverage probabilities (average lengths) of CIs computed using PERC method and m_k with the MEAN, RS and AITQ methods of summarizing the prey. (Based on 100 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients).

where p_k is the diet estimate for the kth species and f_k is the average fat content for each species.

Note that it would be straightforward to incorporate the variability due to fat content into our CI methods. We could simply re-sample the fat content each time the seal and prey FA signatures are re-sampled. As the fat content is one-dimensional, for a large prey base, this source of variability would not be expected to significantly change the length of the CIs. To verify the effect of fat content, simulations would have to be carried out using the pseudo-seals. Since currently our pseudo-seals do not reflect the fat content, we would have to alter how the pseudo-seals are being generated. While we have not carried out such simulations, for our real-life example in Section 4.6, the diet estimates will be adjusted by the fat content.

4.6 Real-life Example: Captive Seabird Data

The QFASA method of diet estimation has recently been applied by S.J. Iverson, A. M. Springer and A. Kitaysky to seabird data (unpublished) that they obtained from islands in the southeastern Bering Sea (Alaska region). They were interested in opposing trends in various populations of seabirds and seals in the Pribilof Islands

and Bogoslof Island (in the southeastern Bering Sea). While in the Pribilof Islands the populations have declined since the mid-1970s, at Bogoslof Island, the populations have increased. As food limitation is related to changes in many populations, knowledge of the predators' diet is considered to be fundamental in understanding the trends. Recall from Chapter 1 the various advantages of the QFASA method over traditional methods in estimating predators' diet. Using data collected on captive seabirds fed a known diet, Iverson, Springer and Kitaysky determined that the QFASA method was a useful way of estimating the diet of seabirds. (Prior to this research the focus of the QFASA method had been on various types of mammals such as seals.) We applied our percentile (PERC) CI method to the captive seabird data in order to assess its performance on real-life data. We now describe the captive seabird experiment and subsequently present our results.

The predator data consisted of 20 Red-legged Kittiwake FA signatures and 26 Common Murre FA signatures while the prey base contained 10 Herring, 15 Silverside and 15 Smelt FA signatures. The Kittiwake chicks were fed a mixture of Herring and Silverside from hatching until day 15. From days 16-42, half of the Kittiwake chicks were switched to Silverside only and the other half to Smelt only. The tissue from which the FA signatures are determined was collected on day 42. The Murre chicks were all fed only Silverside from hatching until day 10. From days 11-45, half of the Murre chicks were continued to be fed only Silverside and the other half fed only Smelt. The tissue was sampled on day 45. Note that one group of Murre chicks was only ever fed Silverside so that calibration factors could obtained and used in our analysis. The fat content of the prey was also recorded and the diet estimates were adjusted for this as well. The CIs for the true diets are given in Tables 4.17 and 4.18.

Diet				Species		
Day 0-15	Day 16-42	n_s		Herring	Silverside	Smelt
Herring/ Silverside	Silverside	10	Median	0.035	0.964	0.000
			CI	[0.008,0.070]	[0.868,0.963]	[0.000,0.044]
Herring/ Silverside	Smelt	10	Median	0.000	0.000	1.000
			CI	[0.000,0.003]	[0.000,0.009]	[0.990,1.000]

Table 4.17: Red-legged Kittiwake Seabirds: Median diet estimate and PERC CIs (bias corrected).

Diet				Species		
Day 0-10	Day 11-45	n_s		Herring	Silverside	Smelt
Silverside	Silverside	13	Median	0.000	1.000	0.000
			CI	[0.000,0.025]	[0.961,1.000]	[0.000,0.043]
Silverside	Smelt	13	Median	0.000	0.000	1.000
			CI	[0.000,0.003]	[0.000,0.008]	[0.991,1.000]

Table 4.18: Common Murre Seabirds: Median diet estimate and PERC CIs (bias corrected).

Shown in the tables are the median of the diet estimates and the PERC method CIs for the four groups. Note that in Table 4.17, the first interval for Silverside does not contain the sample median 0.964 and we surmise that this is due to the bias adjustment. More specifically, before the bias adjustment was applied to the interval, the upper bound was actually 0.932. Thus the bias correction appeared to help, but perhaps not quite enough.

The results are consistent with the findings of Iverson, Springer and Kitaysky and show that the QFASA diet estimation method is an accurate method of estimating the diet of seabirds. Furthermore, for this application, the PERC CIs provide useful interval estimates as they reflect the true diets and are of reasonable lengths.

Chapter 5

A Measure of Species Contribution to Seal Variability

It would be useful to have a statistic similar to the coefficient of determination, R^2, in regression analysis, that measures how well the variability in the seal FA signatures can be explained by a convex linear combination of the prey FA signatures. In Section 5.1 we define such a statistic and establish that it is a sensible measure of explained variability for the diet estimation problem. In Section 5.2, we consider using this statistic to reduce the number of potential species in the diet.

5.1 Definition

For the linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i,$$

R^2 is defined as follows

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}},$$

where $\text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$, $\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, and $\hat{Y}_i$ denotes the fitted values. To formulate an analogous statistic for the diet estimation problem given a sample of seal FA signatures, $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_s}$, and prey FA signatures, $\mathbf{X}_1, \ldots, \mathbf{X}_I$, we attempted to deduce an appropriate "SST" and "SSE" in terms of $\mathbf{Y}_i$ and $\hat{\mathbf{Y}}_i = \sum_{k=1}^{I} p_k \bar{\mathbf{X}}_k$, where $p_k = p_{k,i}(\mathbf{Y}_i, \bar{\mathbf{X}})$, the DMA (distance minimization algorithm) estimate of diet for species k. A natural choice for our "SSE" is $\sum_{i=1}^{n_s} \text{dist}(\mathbf{Y}_i, \hat{\mathbf{Y}}_i)$, where dist denotes either the KL or AIT distance measure. For "SST", or the uncertainty in predicting $\mathbf{Y}$ when $\mathbf{X}_1 \ldots, \mathbf{X}_I$ are not taken into account, there are various possibilities. Three choices are

1. $\sum_{i=1}^{n_s} \text{dist}(\mathbf{Y}_i, \bar{\mathbf{Y}})$.

2. $\sum_{i=1}^{n_s} \text{dist}(\mathbf{Y}_i, \hat{\mathbf{Y}}_i^r)$ where $\hat{\mathbf{Y}}_i^r = \sum_{k=1}^{I} p_k^r \bar{\mathbf{X}}_k$, $p_k^r = p_{k,i}(\mathbf{Y}_i, \bar{\mathbf{X}}^r)$, and $\bar{\mathbf{X}}^r$ denotes that the prey FA signatures were randomly assigned a species label.

3. $\sum_{i=1}^{n_s} \text{dist}(\mathbf{Y}_i, \hat{\mathbf{Y}}^e)$ where $\hat{\mathbf{Y}}^e = \frac{1}{I} \sum_{k=1}^{I} \bar{\mathbf{X}}_k$. (That is, $p_k = \frac{1}{I}$ for all k.)

If we define our "R^2" or the proportion of variability explained (PVE) as

$$\text{PVE} = 1 - \frac{\text{SSE}}{\text{SST}},$$

where $\text{SSE} = \sum_{i=1}^{n_s} \text{dist}(\mathbf{Y}_i, \hat{\mathbf{Y}}_i)$ and SST is one of the above statistics, then since $\text{SST}, \text{SSE} \geq 0$, $\text{PVE} \leq 1$. Furthermore, by the definition of $\mathbf{p}_i(\mathbf{Y}_i, \bar{\mathbf{X}})$, $\text{dist}(\mathbf{Y}_i, \hat{\mathbf{Y}}_i) \leq \text{dist}(\mathbf{Y}_i, \hat{\mathbf{Y}}_i^r)$ for all i, and similarly for $\text{dist}(\mathbf{Y}_i, \hat{\mathbf{Y}}^e)$. Then, for choices 2. and 3., $\text{SSE} \leq \text{SST}$ and $0 \leq \text{PVE} \leq 1$, an appealing property. However, if the true diet is roughly the same for all species (that is $\pi_k \approx \frac{1}{I}$ for all k), then choice 3. would not be effective. We will consequently prefer choice 2.

When the measure of PVE defined thus far was applied to the diet estimation problem, two modifications were found to be needed. Firstly, it was noticed that frequently the sample of seals contained a FA signature, $\mathbf{Y}_i$, with some of its FAs near zero. For this FA signature, $\text{dist}(\mathbf{Y}_i, \hat{\mathbf{Y}}_i)$ and $\text{dist}(\mathbf{Y}_i, \hat{\mathbf{Y}}_i^r)$ were much larger than the distances computed from the other FA signatures in the sample and this one seal had a substantial effect on the PVE statistic. The following more robust SSE and SST statistics appeared to yield a more sensible PVE statistic

$$\text{SSE} = \text{median}_i \left(\text{dist}(\mathbf{Y}_i, \hat{\mathbf{Y}}_i) \right), \text{ and} \tag{5.1}$$

$$\text{SST} = \text{median}_i \left(\text{dist}(\mathbf{Y}_i, \hat{\mathbf{Y}}_i^r) \right). \tag{5.2}$$

It was also observed that the PVE statistic varied a fair amount as $\bar{\mathbf{X}}^r$ varied. We accordingly modified our PVE statistic to use an average SST over B different random assignments of the prey.

In summary, our PVE statistic for the diet estimation problem is defined as

$$\text{PVE(B)} = 1 - \frac{\text{SSE}}{\frac{1}{B} \sum_{b=1}^{B} \text{SST}_b}, \tag{5.3}$$

where SSE is computed as in Equation 5.1 and SST_b, $b = 1, \ldots, B$ is given by Equation 5.2.

In addition to being an overall measure of how well the prey FA signatures explain the seal FA signatures, the PVE statistic in Equation 5.3 provides some insight into which species are significantly contributing to the seal's diet. This is achieved by adding or removing species and observing the change in the PVE statistic. If species k is removed, we calculate

$$\text{PVE}_{-k}(B) = 1 - \frac{\text{SSE}_{-k}}{\frac{1}{B}\sum_{b=1}^{B}\text{SST}_b},$$

where $\text{SSE}_{-k} = \sum_{i=1}^{n_s}\text{dist}(\mathbf{Y}_i, \hat{\mathbf{Y}}_i^{-k})$, $\hat{\mathbf{Y}}_i^{-k} = \sum_{\{s:s\neq k\}}^{I} p_s^{-k}\bar{\mathbf{X}}_s$ and p_s^{-k} is the sth component of $\mathbf{p}^{-k}(\mathbf{Y}, \bar{\mathbf{X}})$, the diet estimate computed without species k as in Subsection 4.3.5. (We may similarly define $\text{PVE}_{-(k,j)}$ to be the PVE statistic without species k and j.) Observe that we consider $\sum_{b=1}^{B}\text{SST}_b$ to be fixed (as in regression analysis) and it is always computed using all of the species that could possibly be part of the seal's diet. Notice also that in our notation, $\mathbf{p}^{-k}(\mathbf{Y}, \bar{\mathbf{X}})$ is of dimension $I-1$. If instead we consider it to be of dimension I with the kth component zero, then it is easy to see that $\text{dist}(\mathbf{Y}_i, \hat{\mathbf{Y}}_i) \leq \text{dist}(\mathbf{Y}_i, \hat{\mathbf{Y}}_i^{-k})$ (or that $\text{SSE} \leq \text{SSE}_{-k}$) so that $\text{PVE}(B) \geq \text{PVE}_{-k}(B)$. If species k is a significant part of the diet, then we would expect a notable drop in the PVE statistic when it is removed from the prey base. We will see, however, that if two species are similar in their FA signatures, dropping one may not greatly change the PVE statistic. Note finally that unlike $\text{PVE}(B)$, it is possible for $\text{PVE}_{-k}(B) < 0$ though this generally only occurs (and not always) when all but one or two species have been dropped.

To assess the usefulness of the PVE statistic in Equation 5.3, we computed PVE(B) and $\text{PVE}_{-k}(B)$, $k = 1, \ldots, I$ for 100 samples of pseudo-seals having Diets 1 and 4 and using the AIT distance measure. (The samples of pseudo-seals were generated as in Section 3.4 and the sample of prey FA signatures was treated as the population.) The average PVE(B) and $\text{PVE}_{-k}(B)$ are given in Tables 5.1 and 5.2. Note that in a preliminary investigation, we found that to two decimal places $\text{PVE}(50) \approx \text{PVE}(500)$. Since time will be a factor (that is, since each calculation of $\text{PVE}(B)$ requires $n_s(B+1)$ optimizations in I dimensions), we will use $B = 50$. For the remainder of the chapter, we will let $\text{PVE} = \text{PVE}(50)$ and $\text{PVE}_{-k} = \text{PVE}_{-k}(50)$.

Based on Tables 5.1 and 5.2, PVE and PVE_{-k} appear to be sensible measures of species contribution to the variability in the seal FA signatures. Notice that the

	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
	Diet 1	0.30	0.30	0	0.15	0	0.15	0	0
	μ_p	0.283	0.245	0.023	0.224	0.045	0.134	0.017	0.029
n_s	PVE	PVE_{-1}	PVE_{-2}	PVE_{-3}	PVE_{-4}	PVE_{-5}	PVE_{-6}	PVE_{-7}	PVE_{-8}
1	0.702	0.656	0.679	0.699	0.678	0.685	0.679	0.698	0.697
5	0.782	0.726	0.756	0.779	0.752	0.769	0.761	0.777	0.778
10	0.801	0.739	0.773	0.798	0.769	0.795	0.781	0.797	0.799

Table 5.1: Diet 1, AIT distance, $B = 50$: Average PVE statistics for three sample sizes. (Based on 100 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

	Species	COD	HAD	PLC	POL	SAND	SH	WF	YT
	Diet 4	0.09	0	0.09	0	0.45	0.09	0.09	0.09
	μ_p	0.104	0.043	0.069	0.155	0.369	0.117	0.074	0.069
n_s	PVE	PVE_{-1}	PVE_{-2}	PVE_{-3}	PVE_{-4}	PVE_{-5}	PVE_{-6}	PVE_{-7}	PVE_{-8}
1	0.701	0.694	0.697	0.694	0.687	0.506	0.689	0.681	0.695
5	0.782	0.774	0.780	0.774	0.772	0.569	0.772	0.762	0.775
10	0.786	0.777	0.784	0.777	0.773	0.579	0.774	0.765	0.780

Table 5.2: Diet 4, AIT distance, $B = 50$: Average PVE statistics for three sample sizes. (Based on 100 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

PVE increases with n_s and is similar for both diets. The latter implies that perhaps the PVE is somewhat independent of the underlying true diet, a property that may have practical usefulness. For example, biologists may want to compute the PVE for several subsets of FAs to determine the subset for which the prey FA signatures best explain the variability in the seal FA signatures. In this case, it may be beneficial for the PVE statistic not to depend on the true diet of the sample of seals so that the results can be extended to seals with a different diet.

When species k is removed, PVE_{-k} also behaves in a desirable manner and tends to decrease if $\pi_k > 0$ and generally by an amount related to the size of π_k. As an example, observe the decrease in PVE when Sandlance ($\pi_k = 0.45$) is removed from Diet 4 (Table 5.2). For some species, however, the decrease in PVE_{-k} may not be completely representative of the size of π_k. In Table 5.1 for example, although $\pi_k = 0.30$ for Haddock and $\pi_k = 0.15$ for Pollock, when $n_s = 10$, the decrease in PVE is roughly the same. Two potential explanations for these types of occurrences are as follows: 1) As previously mentioned, some of the species have similar FA signatures (see hierarchical cluster analysis in Figure 3.3) and, consequently, removing one of these species may not greatly affect PVE, and 2) we have difficulty accurately estimating the diet of some species. As discussed in Section 3.3, the diet estimates, $\mathbf{p}(\mathbf{Y}, \boldsymbol{\mu}_\mathbf{X})$, are actually biased and the decrease in PVE for species k often reflects the size of $\mu_{p_k} = \text{E}_\mathbf{Y}[p_k(\mathbf{Y}, \boldsymbol{\mu}_\mathbf{X})]$ rather than π_k. As an example consider Pollock in Diet 1. Although $\pi_k = 0.15$, $\mu_{p_k} = 0.224$ and this may help to explain the larger than expected decrease in PVE for this species.

5.2 Application

A potential application of the PVE statistic is the reduction of the number of possible species in the diet, not unlike the selection of predictor variables in regression analysis. In this section we discuss a procedure that attempts to conservatively reduce the number of potential species in the diet. When the number of potential species is large but only a few are contributing to the diet, applying this procedure before one of the interval estimation methods of Chapter 4 could significantly decrease the computational time and yield shorter intervals.

To determine which species, if any, can be removed from the prey base, we implemented a backward elimination type procedure. The procedure begins with all potential species being included and determines which species' removal causes the smallest change in PVE. If species s corresponds to the smallest change in PVE, the algorithm tests

$$\begin{aligned} H_0 &: \text{ Species } s \text{ has no effect} \\ H_1 &: \text{ Species } s \text{ has an effect.} \end{aligned} \tag{5.4}$$

If H_0 is *not* rejected then species s is removed from the prey base and we then examine the change in PVE_{-s} when the remaining species are removed. This process continues until either the species in question cannot be dropped or until only one species remains. (Note that unlike backward elimination in regression, only $I-1$ species can be dropped and our final prey base must always contain at least one species.) The overall backward elimination algorithm is as follows:

Backward Elimination Algorithm

1. Let $s_1 = 0$.
2. *for* k=1:(I-1)
   (a) Compute $PVE_{-(s_1,\ldots,s_k)}$ and $PVE_{-(s_1,\ldots,s_k,j)}$, $j \in \{1,\ldots,I\} - \{s_1,\ldots,s_k\}$, where $PVE_{-0} = PVE$ and $PVE_{-(0,j)} = PVE_{-j}$.
   (b) Compute $s_{k+1} = \arg\min_j \left(PVE_{-(s_1,\ldots,s_k)} - PVE_{-(s_1,\ldots,s_k,j)}\right)$, $j \in \{1,\ldots,I\} - \{s_1,\ldots,s_k\}$.
   (c) Test whether species s_{k+1} may be dropped. If not, exit *for* loop.

To carry out Step (c) in the backward elimination algorithm, an effective testing procedure based on the PVE statistic is required. We examined a variety of ways of obtaining a P-value for the hypothesis test in Equation 5.4. Our preferred method is a nonparametric approach in which samples of pseudo-seals are generated under the null and the observed value of $T = PVE - PVE_{-s}$ is compared to the bootstrap distribution of T under the null. Note that the procedure is similar to the NONPAR method of obtaining a P-value (Subsection 4.3.5) with the null hypothesis, in this case, being $\pi_k = 0$. The procedure, in detail, is as follows:

1. Compute $T = \text{PVE} - \text{PVE}_{-s}$.

2. For each seal, compute the null diet: $p_i^{-s}(\mathbf{Y}_i, \bar{\mathbf{X}})$, $i = 1, \ldots, n_s$.

3. *for* $r = 1 : R$

   (a) Generate n_s seals, $\mathbf{Y}_1^{*r}, \ldots, \mathbf{Y}_{n_s}^{*r}$, where the ith seal has null diet $p_i^{-s}(\mathbf{Y}_i, \bar{\mathbf{X}})$.

   (b) Using $\mathbf{Y}_1^{*r}, \ldots, \mathbf{Y}_{n_s}^{*r}$, compute $T^{*r} = \text{PVE}^{*r} - \text{PVE}_{-s}^{*r}$.

4. Compute

$$p^{\text{boot}} = \frac{\#\{T^{*r} \geq T\}}{R}.$$

Note that in the backward elimination algorithm, to test whether species s_{k+1} may be dropped, our test statistic is actually $T = \text{PVE}_{-(s_1,\ldots,s_k)} - \text{PVE}_{-(s_1,\ldots,s_k,s_{k+1})}$. Additionally, realize that the above procedure does not account for the variability due to the prey but this could easily be incorporated by re-sampling the prey signatures in Step 3.

Before presenting some results, it should be mentioned that due to the number of optimizations involved in the backward elimination procedure with the above testing procedure, the algorithm is extremely computationally intensive. We therefore also investigated the possibility of using some less time consuming testing procedures. A simple approach is to simply assume that $T = \text{PVE} - \text{PVE}_{-s}$ is approximately normally distributed and to estimate the variance of T using Davison and Hinkley's (1997) jackknife approximation to the variance, which does not require further optimizations. The jackknife variance is given by

$$v_{\text{jack}} = \frac{1}{n_s(n_s-1)}\left[\sum_{i=1}^{n_s} l_{\text{jack},i}^2 - n_s\left(-\frac{1}{n_s}\sum_{i=1}^{n_s} l_{\text{jack},i}\right)^2\right],$$

where $l_{\text{jack},i}$ are the jackknife empirical influence values defined in Subsection 4.3.5. Essentially $l_{\text{jack},i}$ is computed by removing $\text{dist}(\mathbf{Y}_i, \hat{\mathbf{Y}}_i)$ and $\text{dist}(\mathbf{Y}_i, \hat{\mathbf{Y}}_i^T)$ in the calculation of SSE and SST respectively. We then assume that $\frac{T}{\sqrt{v_{\text{jack}}}} \dot{\sim} \mathcal{N}(0,1)$ under H_0 and reject H_0 if $T > z_{1-\alpha}$.

To assess the validity of the changes in PVE being approximately normally distributed, histograms of $\text{PVE}-\text{PVE}_{-k}$, $k = 1, \ldots, I$ for the 100 samples of size $n_s = 10$

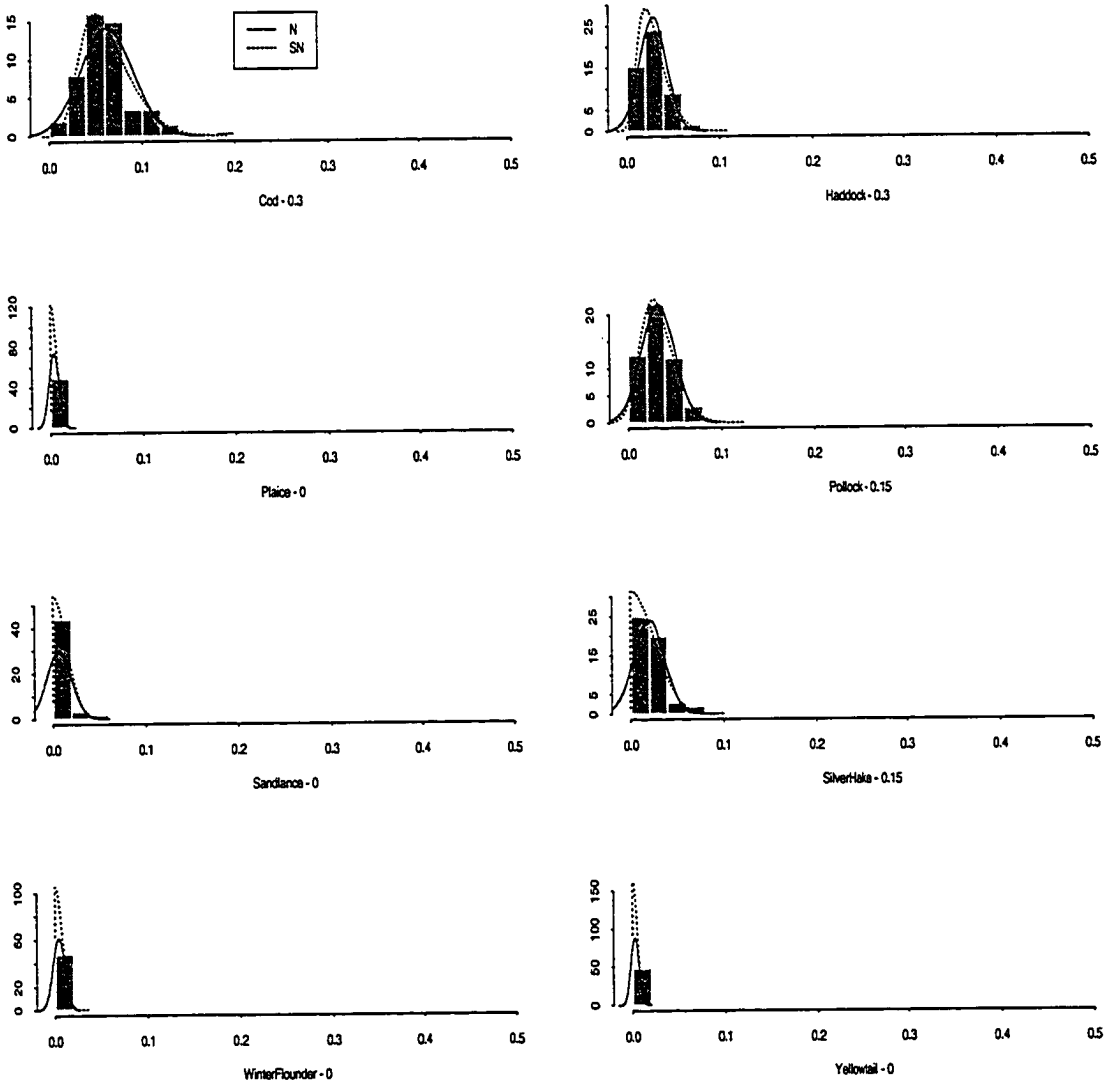

Figure 5.1: Diet 1, $n_s = 10$, AIT distance, $B = 50$: Distribution of PVE $-$ PVE$_{-k}$. (Based on 100 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

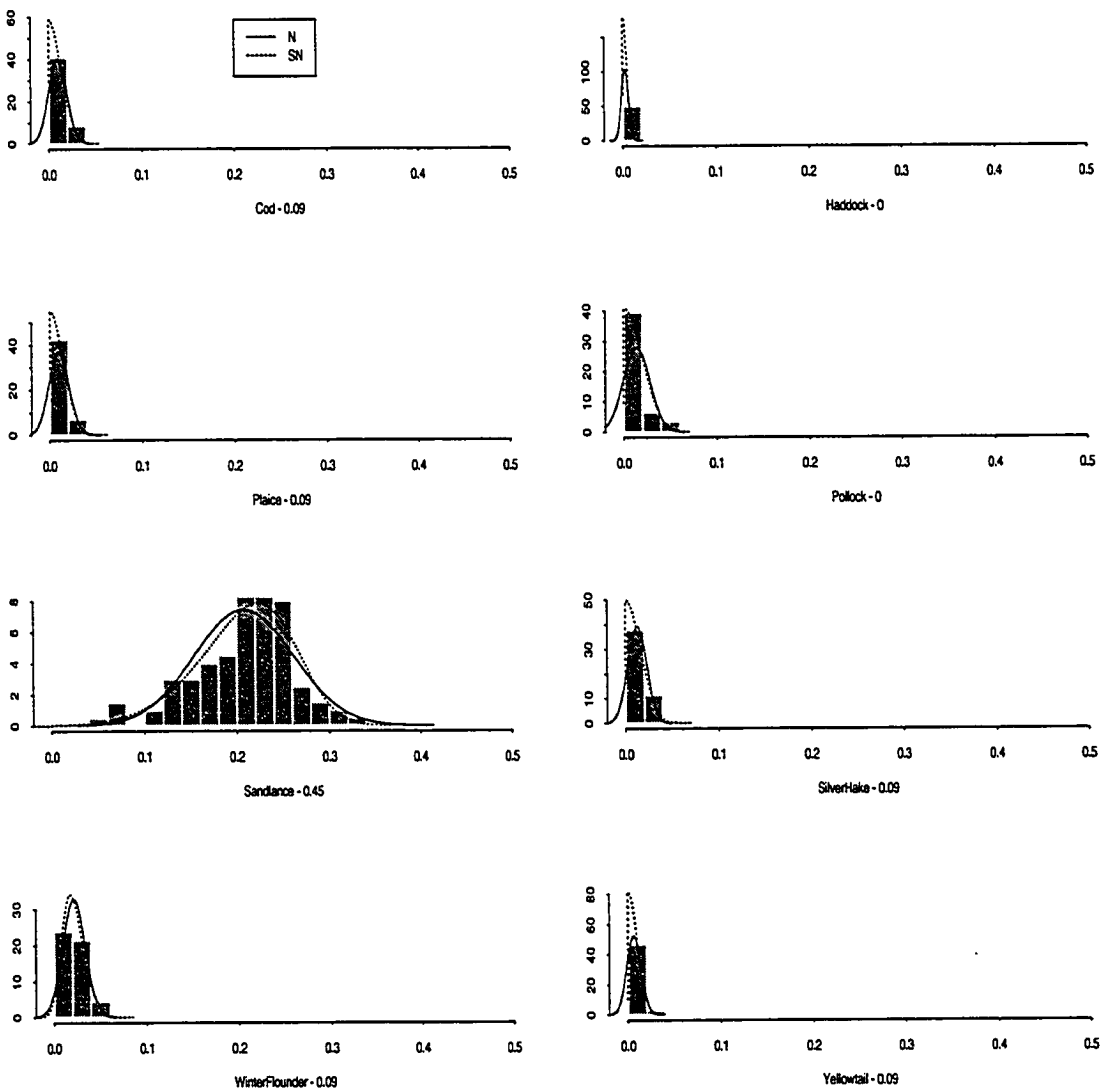

Figure 5.2: Diet 4, $n_s = 10$, AIT distance, $B = 50$: Distribution of $\text{PVE} - \text{PVE}_{-k}$. (Based on 100 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

used to produce Tables 5.1 and 5.2 are given in Figures 5.1 and 5.2. From the Figures, the normal approximation appears to be somewhat reasonable when $\pi_k \geq 0.30$ but the data is often very skewed to the right, particularly when $\pi_k < 0.15$. As a result, in these cases our probability of Type I error may actually be larger than α and we may not drop some species when we should. If the purpose of the elimination procedure is viewed as a first step to interval estimation, then not removing a species that should be removed may not be considered a crucial error. While a variety of transformations were applied to the data, none appeared to greatly improve the fit of the normal distribution. The univariate skew-normal ($\mathcal{SN}(\mu_k, \sigma_k, \alpha_k)$) distribution (defined in Chapter 2) was also fit to the differences and the fit is shown in Figures 5.1 and 5.2 as well. The plots show the data to generally be well fit by the $\mathcal{SN}$ distribution. If time was not an issue, the nuisance parameters σ_k and α_k could be estimated using a bootstrap approach and then a parametric bootstrap P-value computed by generating differences under the null skew distribution. This parametric method would be comparable in computational intensity to the nonparametric method and, due to time constraints, we could not obtain results from both methods. Furthermore, although the normal approximation method is much faster than the other methods, it is still a fairly time consuming algorithm to run. We therefore chose only to examine the performance of the backward elimination algorithm with the nonparametric testing procedure. We used a prey base similar to that used in Iverson *et al* (2004) that contained 27 species (see Figure 5.3 and Appendix A for the species that were used) and from this prey base generated 20 samples of pseudo-seals from each diet (Diet 1 and Diet 4) and both distance measures (AIT and KL). (The samples were generated as in the previous section.) The results are given in Tables 5.3-5.6.

The Tables essentially contain the power associated with our nonparametric testing procedure as they give the proportion of time the species with non-zero diets are not dropped. Note that the power calculations are based only on 20 samples of pseudo-seals, $B = 5$ and $R = 50$ due to the slowness of the algorithm and are therefore only approximate. For this reason and because not dropping a species when we should (and therefore committing a Type I error) isn't considered crucial, we have used $\alpha = 0.2$ instead of the usual values of α such as 0.01, 0.05 or 0.1. We have also

I	Average No.Species Dropped	Average PVE	Species	Diet 1	μ_p	Power
27	17.95	0.900	Cod	0.30	0.221	1.00
			Haddock	0.30	0.040	0.35
			Pollock	0.15	0.246	1.00
			SilverHake	0.15	0.038	0.60

Table 5.3: Diet 1, $n_s = 10$, AIT distance, $B = 5$, $R = 50$, $\alpha = 0.2$: Results of backward elimination procedure applied to 27 species. (Based on 20 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

I	Average No.Species Dropped	Average PVE	Species	Diet 1	μ_p	Power
27	19.70	0.966	Cod	0.30	0.149	1.00
			Haddock	0.30	0.183	0.90
			Pollock	0.15	0.186	0.90
			SilverHake	0.15	0.095	0.90

Table 5.4: Diet 1, $n_s = 10$, KL distance, $B = 5$, $R = 50$, $\alpha = 0.2$: Results of backward elimination procedure applied to 27 species. (Based on 20 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

I	Average No.Species Dropped	Average PVE	Species	Diet 4	μ_p	Power
27	19.45	0.905	Cod	0.09	0.046	0.35
			Plaice	0.09	0.029	0.60
			Sandlance	0.45	0.326	1.00
			SilverHake	0.09	0.037	0.55
			WinterFlounder	0.09	0.063	0.90
			YellowTail	0.09	0.038	0.20

Table 5.5: Diet 4, $n_s = 10$, AIT distance, $B = 5$, $R = 50$, $\alpha = 0.2$: Results of backward elimination procedure applied to 27 species. (Based on 20 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

I	Average No.Species Dropped	Average PVE	Species	Diet 4	$\mu_{\mathbf{p}}$	Power
27	18.25	0.959	Cod	0.09	0.020	0.20
			Plaice	0.09	0.005	0.10
			Sandlance	0.45	0.361	1.00
			SilverHake	0.09	0.047	0.75
			WinterFlounder	0.09	0.076	1.00
			YellowTail	0.09	0.047	0.30

Table 5.6: Diet 4, $n_s = 8$, KL distance, $B = 5$, $R = 50$, $\alpha = 0.2$: Results of backward elimination procedure applied to 27 species. (Based on 20 samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients.)

included $\mu_{\mathbf{p}} = \mathrm{E}_{\mathbf{Y}}[\mathbf{p}(\mathbf{Y}, \mu_{\mathbf{X}})]$ in the tables which was computed as in Chapter 3 and is the average estimated diet of 1000 pseudo-seals generated with Diet 1 or Diet 4. Also shown is the average PVE statistic of the 20 samples of pseudo-seals as well as the average number of species dropped.

The backward elimination algorithm was first applied with the AIT distance measure (see Tables 5.3 and 5.5). Although it appears that the algorithm is performing poorly for all but a few species, $\mu_{\mathbf{p}}$ shows that the AIT distance is not yielding accurate estimates of the diet (see Haddock in Diet 1, for example) implying that some species are not being distinguished from others. A hierarchical cluster analysis is given in Figure 5.3 and helps to explain the large bias in the diet estimates computed with the AIT distance measure, specifically in Diet 1. As an example, consider Haddock in Figure 5.3 which appears to be similar to SeaRaven, RedHake and WhiteHake. While $\mu_{p_k} = 0.040$ for Haddock instead of 0.30, the sum of the components of μ_{p_k} corresponding to SeaRaven, RedHake and WhiteHake (not shown) is approximately 0.15 instead of 0. This suggests that the FA signatures from these three species are perhaps not being distinguished from the FA signatures of Haddock. Also, note the similarity between Pollock and SilverHake in Figure 5.3. This is in accordance with Table 5.3 where Pollock is over-estimated and SilverHake underestimated. For Diet 4, the results are perhaps not as unreasonable but it appears to be the case that if π_k is small ($\pi_k < 0.10$), species k will often be dropped.

The backward elimination algorithm was also applied with the KL distance as the

bias in the estimates, when all 27 species were used, was found to be much less than with the AIT distance measure, particularly for Diet 1. Note the larger average PVE statistic that is obtained when the KL distance is used, suggesting that a better fit occurs with this distance measure. Also with the KL distance, the performance of the backward elimination algorithm is similar to the AIT distance results for Diet 4 but much better with Diet 1. From Figure 5.3, with the KL distance Haddock appears to be more dissimilar to RedHake and WhiteHake then with the AIT distance.

Although a larger number of samples of pseudo-seals, B and R would be required to fully assess the procedure, our results show that overall the backward elimination algorithm tends to drop a species from the diet if $\mu_{p_k} \leq 0.06$. Consequently, if it is believed that our estimates are fairly accurate (such as with the KL distance in the above simulations) and if, roughly, $\pi_k \geq 0.15$, the elimination algorithm can be useful in significantly reducing the number of potential species in the diet. In Table 5.4 where the algorithm generally did not drop the non-zero species, on average, it reduced the number of species from 27 to about 7.

In summary, the backwards elimination procedure presented appears to be a potentially useful way of greatly reducing the number of possible species in the diet. From the results of our small simulation study, the procedure must be used with some caution, however, since its usefulness is related to the accuracy of the diet estimates.

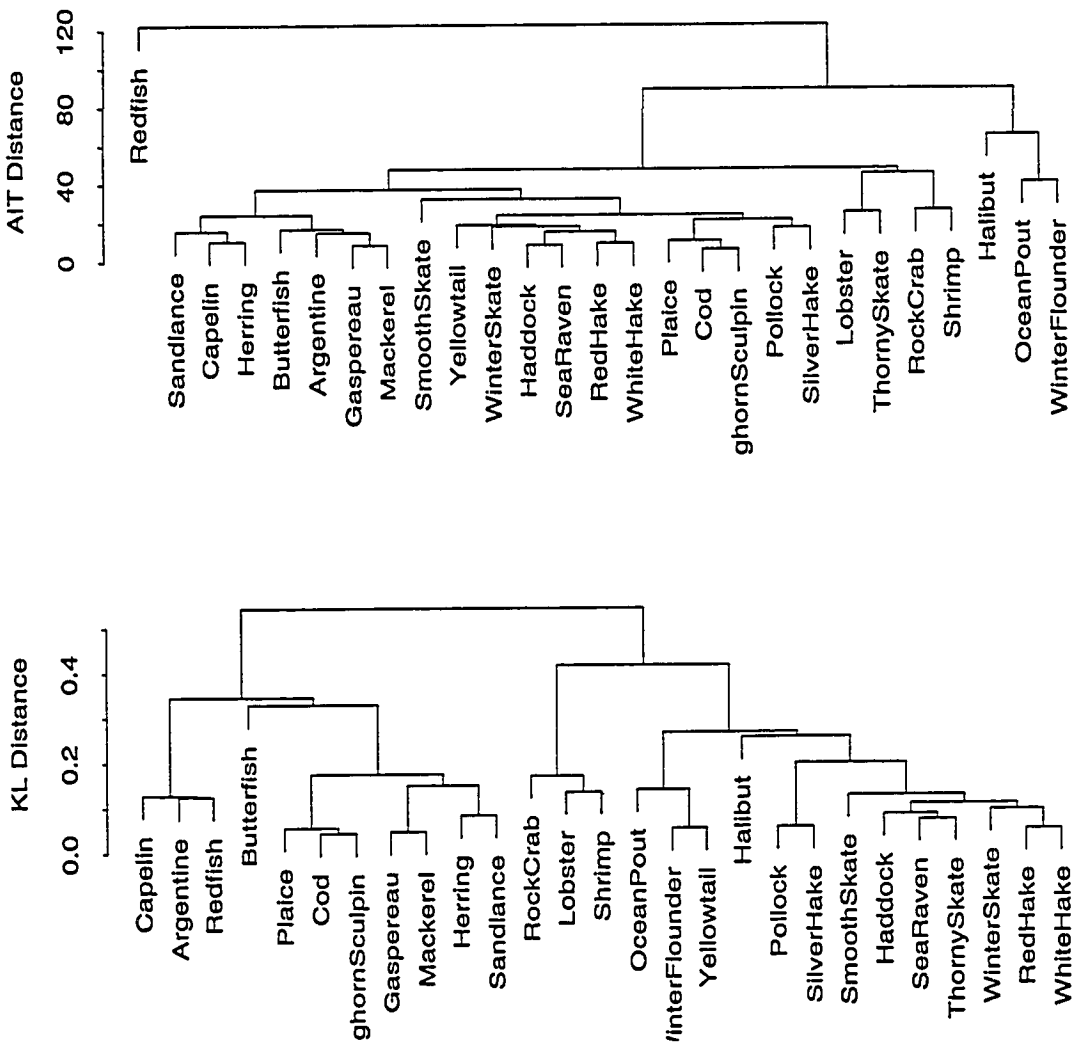

Figure 5.3: Hierarchical cluster analysis on $\bar{\mathbf{X}}_k$, $k = 1, \ldots, 27$, using *hclust* in S-PLUS with both the AIT and KL distance measures, and the average linkage method.

Chapter 6

Testing for a Difference in Diet

In this chapter we explore methods of determining whether 1) two independent populations of seals have different diets and 2) the diet for a single population of seals has changed. For both problems we develop testing procedures for the case when the data consists only of samples of seal FA signatures (either independent or paired), and when, in addition to the seal FA signatures, corresponding prey bases (or a common prey base) are also available. In the latter case, we will use the prey bases to estimate the diets (using the methods of Chapter 3) and will base the hypothesis test on the diet estimates rather than on the FA signatures. As the data will therefore either be FA signatures or diet estimates, appropriate tests for differences in compositional data are needed. Although Aitchison (1986) and (2003) discuss such tests, their methods are not always directly applicable to our problems as they require the number of observations in the samples to be larger than their dimension, and often we will have, for example, the number of seals being much smaller than the number of FAs. This is the case for our real-life example presented in Section 6.4. Additionally, when the sample sizes are small, Aitchison's methods require parametric assumptions while we will attempt to use nonparametric permutation tests.

6.1 Preliminary Issues

In addition to the challenges of the number of seals in our sample potentially being small and the data being compositional, a few issues arose during our analysis. To simplify the discussions, we will consider the issues involved in comparing the diet of two independent populations of seals as they are essentially the same for paired comparisons.

Although our primary interest is usually whether the two populations of seals have the same diet, given samples of seal FA signatures only, a crucial issue is that we are

limited to testing for a difference in the FA signatures. If no difference is found in the FA signatures, then we will assume that there is also no difference in the diets. The difficulty is when the FA signatures are found to be significantly different since this may or may not imply a difference in the diets. Consider two populations of seals eating only one species, say Cod, from two different regions. If the FA signatures of Cod from the two regions are different then so will be the seal FA signatures and we may correctly reject the null hypothesis that the seal FA signatures are the same even though the seals have the same diet. As will be seen in our simulation study, seemingly small differences in the prey populations are sufficient to result in the conclusion that the seals have different FA signatures even when their diets are the same.

Another issue that arose was the "zero problem" since it is possible for the seal FA signatures and the diet estimates to contain zeros. This is problematic as our preferred test statistics involve logarithms. Recall that two types of zeros, namely *essential* and *rounded* zeros, were defined in Section 2.1 based on Martín-Fernández *et al* (2003). We will assume that if only a few zeros are present in the seal FA signatures, that they are *rounded* zeros and simply fall below detection limit. In this case we will apply the previously discussed multiplicative replacement strategy (Martín-Fernández *et al*, 2003) to a seal FA $\mathbf{Y}$ with n_z zeros. That is , we will replace the jth component Y_j by

$$Y_j' = \begin{cases} \delta, & \text{if } Y_j = 0, \\ (1 - n_z\delta)Y_j, & \text{if } Y_j > 0, \end{cases} \tag{6.1}$$

where we let $\delta = \frac{\text{smallest non-zero FA in sample}}{10}$.

The zeros in the diet estimates are considered to be *essential* zeros since they indicate a true absence of the species from the diet. This issue will influence our choice of test statistic and will be discussed further when testing procedures are considered in detail.

6.2 Comparison of Two Independent Populations

6.2.1 Analysis Based on Seal FA Signatures Only

Suppose we have two independent samples of seal FA signatures, say $\mathbf{Y}_{1i}$, $i = 1, \ldots, n_{s1}$ and $\mathbf{Y}_{2i}$, $i = 1, \ldots, n_{s2}$, of dimension n_{FA} with true diets π_1 and π_2 respectively. Let Y_{bij} denote the jth component of the ith observation from the bth sample, $j = 1, \ldots, n_{FA}$, $i = 1, \ldots, n_{sb}$ and $b = 1, 2$. To simplify the notation that follows, let $N_{FA} = n_{\text{FA}} - 1$. In this subsection we assume that prey bases, say $\mathbf{X}_1$ and $\mathbf{X}_2$ (or a common prey base, $\mathbf{X}$), representative of the prey populations from which the seals are eating, are unknown. Consider first testing for a difference in the FA signatures from the two populations. The usefulness of the test when we are actually interested in testing for a difference in the underlying diets is discussed subsequently.

In accordance with Aitchison (1986), suppose that

$$\begin{aligned} \mathbf{Y}_1 &\sim \mathcal{L}^{N_{FA}}(\mu_1, \Sigma_1), \\ \mathbf{Y}_2 &\sim \mathcal{L}^{N_{FA}}(\mu_2, \Sigma_2), \end{aligned}$$

or that,

$$\begin{aligned} \mathbf{Z}_1 = \log\left(\frac{\mathbf{Y}_{1-n_{FA}}}{Y_{1n_{FA}}}\right) &\sim \mathcal{N}^{N_{FA}}(\mu_1, \Sigma_1), \\ \mathbf{Z}_2 = \log\left(\frac{\mathbf{Y}_{2-n_{FA}}}{Y_{2n_{FA}}}\right) &\sim \mathcal{N}^{N_{FA}}(\mu_2, \Sigma_2). \end{aligned} \tag{6.2}$$

If we are interested in whether a difference in the two populations of FA signatures exists, we may test

$$\begin{aligned} \mathrm{H}_0 &: \mu_1 = \mu_2 \\ \mathrm{H}_1 &: \mu_1 \neq \mu_2, \end{aligned} \tag{6.3}$$

using standard multivariate techniques, provided both n_{s1} and n_{s2} are sufficiently large. Let $\hat{\mu}_b = \bar{\mathbf{Z}}_b = \frac{1}{n_s}\sum_{i=1}^{n_s} \mathbf{Z}_{bi}$, $b = 1, 2$. Then, for example, if we assume that $\Sigma_1 = \Sigma_2$, we would reject H_0 if

$$\begin{aligned} T^2 &= (\hat{\mu}_1 - \hat{\mu}_2)' \left[\left(\frac{1}{n_{s1}} + \frac{1}{n_{s2}}\right) \mathrm{S}_{\text{pool}}\right]^{-1} (\hat{\mu}_1 - \hat{\mu}_2) \\ &> \frac{(n_{s1} + n_{s2} - 2)N_{FA}}{n_{s1} + n_{s2} - N_{FA} - 1} F_{N_{FA}, n_{s1}+n_{s2}-N_{FA}-1}(1 - \alpha), \end{aligned}$$

where

$$\mathrm{S}_{\text{pool}} = \frac{n_{s1}-1}{n_{s1}+n_{s2}-2}\mathrm{S}_1 + \frac{n_{s2}-1}{n_{s1}+n_{s2}-2}\mathrm{S}_2$$

and $\mathrm{S}_b = \frac{1}{n_{sb}-1}\sum_{i=1}^{n_{sb}}(\mathbf{Z}_{bi}-\hat{\boldsymbol{\mu}}_b)(\mathbf{Z}_{bi}-\hat{\boldsymbol{\mu}}_b)'$, $b = 1, 2$.

If $n_{s1} - N_{FA}$ and $n_{s2} - N_{FA}$ are large, then the normality assumptions are less critical and we would reject H_0 if

$$T^2 = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)'\left(\frac{1}{n_{s1}}\mathrm{S}_1 + \frac{1}{n_{s2}}\mathrm{S}_2\right)^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) > \chi^2_{N_{FA}}(1-\alpha).$$

For n_{s1} and n_{s2} less than N_{FA}, we have chosen to use an approach similar to Davison and Hinkley's (1997) univariate nonparametric permutation test for the comparison of two means. They argue that in the univariate case, for certain forms of the underlying densities (in which the null hypothesis implies a common density for the two populations), the null hypothesis sufficient statistic is the order statistics for the pooled sample. In these cases, a P-value can be computed by pooling the n_{s1} and n_{s2} observations and considering permutations of the concatenation of the two random samples. For each permutation, the first n_{s1} components of the concatenation vector give the first sample and the remaining n_{s2} observations the second sample. An appropriate test statistic is calculated for R permutations and the P-value is computed in a manner similar to the bootstrap P-value. Realize that if we assume that the normality assumption in Equation 6.2 is valid, then, in the univariate case, the permutation test is justified if $\sigma_1^2 = \sigma_2^2$. Note also that Davison and Hinkley's (1997) nonparametric bootstrap test for the comparison of means problem is similar to the permutation test but the sampling is done with replacement.

It is straightforward to extend the algorithm to the multivariate case since all that is required is a suitable test statistic. In our simulations (to be discussed) we examined the performance of various test statistics. To simplify the notation, for vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ of dimension D, we define the following distance measures

$$\begin{aligned}
\mathrm{ABS}(\mathbf{X}_1, \mathbf{X}_2) &= \sum_{j=1}^{D} |X_{1j} - X_{2j}|, \\
\mathrm{REL}(\mathbf{X}_1, \mathbf{X}_2) &= \sum_{j=1}^{D} r_j, \text{ where } r_j = \begin{cases} 0 & \text{if } X_{1j} = X_{2j}, \\ \frac{|X_{1j}-X_{2j}|}{\max(X_{1j}, X_{2j})} & \text{otherwise,} \end{cases}
\end{aligned}$$

$$\mathrm{SQ}(\mathbf{X}_1, \mathbf{X}_2) \;=\; \sum_{j=1}^{D} (X_{1j} - X_{2j})^2 .$$

Our chosen test statistics may be expressed as functions of the above defined distance measures and are as follows

1. $T_1 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \mathrm{REL}(\mathbf{Y}_{1i_1}, \mathbf{Y}_{2i_2})$
2. $T_2 = \mathrm{median}_{i_1,i_2} \left(\mathrm{REL}(\mathbf{Y}_{1i_1}, \mathbf{Y}_{2i_2})\right)$
3. $T_3 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \mathrm{ABS}(\mathbf{Z}_{1i_1}, \mathbf{Z}_{2i_2})$
4. $T_4 = \mathrm{median}_{i_1,i_2} \left(\mathrm{ABS}(\mathbf{Z}_{1i_1}, \mathbf{Z}_{2i_2})\right)$
5. $T_5 = \sqrt{\mathrm{SQ}(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2)}.$

The test statistics given in 1. and 2. are useful when there are essential zeros in the data and are perhaps the simplest to interpret. (Note that our relative distance definition, REL, allows for the possibility of essential zero components by dividing by the maximum of the two components. If both components are zero, then the relative difference is set to zero.) While we could have chosen say, $T = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \mathrm{ABS}(\mathbf{Y}_{1i_1}, \mathbf{Y}_{2i_2})$ as a possible test statistic, test statistics based on relative differences such as those given in 1.-5. above are more appropriate for compositional data. Furthermore, the biologists with whom we have been collaborating have indicated that, in fact, relative differences in the FA signatures or the diet estimates are of most interest.

The multivariate permutation test algorithm, in detail, is as follows:

<u>Multivariate Permutation Test</u>

1. Compute the test statistic, T.
2. *for* $r = 1, \ldots, R$
   (a) Permute the $n_{s1} + n_{s2}$ observations to obtain: $\mathbf{Y}_i^{*r}$, $i = 1, \ldots, n_{s1} + n_{s2}$.
   (b) Let $\mathbf{Y}_{1i}^{*r} = \mathbf{Y}_i^{*r}$, $i = 1, \ldots, n_{s1}$ and $\mathbf{Y}_{2i}^{*r} = \mathbf{Y}_i^{*r}$ $i = n_{s1} + 1, \ldots, n_{s1} + n_{s2}$.
   (c) If required, compute $\mathbf{Z}_i^{*r} = \log\left(\frac{\mathbf{Y}_{i-n_{FA}}^{*r}}{Y_{in_{FA}}^{*r}}\right)$.

(d) Compute the test statistic, T^{*r}, using the generated samples in (b) or (c).

3. Compute

$$p^{\text{perm}} = \frac{\#\{T^{*r} \geq T\}}{R}.$$

As previously discussed, we may be more concerned with whether or not the diets of the two populations of seals differ rather than whether the FA signatures differ. Recall from Section 6.1 that the performance of say, the multivariate permutation test, in testing

$$\begin{aligned} \mathrm{H}_0 &: \boldsymbol{\pi}_1 = \boldsymbol{\pi}_2 \\ \mathrm{H}_1 &: \boldsymbol{\pi}_1 \neq \boldsymbol{\pi}_2, \end{aligned} \tag{6.4}$$

would depend on the extent of the differences in the two prey populations. If the prey populations are different, we may reject $\mathrm{H}_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ when, in fact, $\mathrm{H}_0 : \boldsymbol{\pi}_1 = \boldsymbol{\pi}_2$ is true, consequently committing a Type I Error if we are really interested in the latter. A simulation study was carried out in which we computed the P[Type I Error] when two samples of pseudo-seals were generated with the same diet from 1) separate prey bases and 2) the same (full) prey base. To create separate prey bases we simply randomly (and evenly) divided our full prey base in two. We used the following simulation algorithm to compute the P[Type I Error] associated with the multivariate permutation test when the hypotheses of interest are given by Equation 6.4.

1. *for* $m = 1, \ldots, M$

   (a) Choose a diet, say $\boldsymbol{\pi}$.

   (b) Either

      i. Randomly and evenly split prey base into $\mathbf{X}_1$ and $\mathbf{X}_2$.

      ii. Generate a sample of n_{s1} and n_{s2} pseudo-seals with diets $\boldsymbol{\pi}$ from $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively.

   or

      i. Generate a sample of n_{s1} and n_{s2} pseudo-seals with diets $\boldsymbol{\pi}$ from the full prey base.

   (c) With the generated samples, say, $\mathbf{Y}_{11}, \ldots, \mathbf{Y}_{1n_{s1}}$ and $\mathbf{Y}_{21}, \ldots, \mathbf{Y}_{2n_{s2}}$, compute the multivariate permutation test P-value: $p^{\text{perm},m}$.

2. Compute

$$P[\text{Type I Error}] = \frac{\#\{p^{\text{perm},m} \leq \alpha\}}{M}.$$

We chose $\alpha = 0.01,\ 0.05$ and 0.1, $M = 100$ and $n_{s1} = n_{s2} = 10$ (since standard multivariate techniques can be applied when n_{s1} and n_{s2} are large). Further, to assess the magnitude of the difference in samples of seals generated from two separate prey bases versus one common prey base, two samples of 1000 pseudo-seals were generated with and without splitting and $\text{REL}(\text{E}[\mathbf{Y}_1], \text{E}[\mathbf{Y}_2]))$ and $\sqrt{\text{SQ}(\mu_1, \mu_2)}$ computed in both cases. (With $n_{s1} = n_{s2} = 1000$, we assume that $\text{E}[\mathbf{Y}_b] \approx \bar{\mathbf{Y}}_b$, and that $\mu_b \approx \hat{\mu}_b$ $b = 1, 2$). Note that these distances are only for one particular split. The results are given in Table 6.1.

						P[Type I Error]		
π_1	π_2	Prey Base	REL $(\text{E}[\mathbf{Y}_1], \text{E}[\mathbf{Y}_2])$	$\sqrt{\text{SQ}(\mu_1, \mu_2)}$	T	α 0.01	α 0.05	α 0.1
Diet 1	Diet 1	Split	1.626	0.398	T_1	0.88	0.96	0.97
					T_2	0.67	0.92	0.95
					T_3	0.66	0.86	0.93
					T_4	0.55	0.77	0.85
					T_5	0.69	0.88	0.90
		Full	0.100	0.0224	T_1	0.01	0.05	0.09
					T_2	0.02	0.04	0.08
					T_3	0.01	0.04	0.08
					T_4	0.01	0.01	0.07
					T_5	0	0.04	0.07
Diet 4	Diet 4	Split	1.856	0.569	T_1	0.99	1.00	1.00
					T_2	0.99	1.00	1.00
					T_3	0.82	0.98	1.00
					T_4	0.73	0.93	0.98
					T_5	0.80	0.96	0.99
		Full	0.048	0.0162	T_1	0.02	0.05	0.08
					T_2	0.02	0.06	0.10
					T_3	0.02	0.06	0.09
					T_4	0.02	0.04	0.07
					T_5	0.02	0.07	0.12

Table 6.1: P[Type I Error] associated with the multivariate permutation test at $n_{s1} = n_{s2} = 10$. $T_1 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \text{REL}(\mathbf{Y}_{1i_1}, \mathbf{Y}_{2i_2})$, $T_2 = \text{median}_{i_1,i_2}(\text{REL}(\mathbf{Y}_{1i_1}, \mathbf{Y}_{2i_2}))$, $T_3 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \text{ABS}(\mathbf{Z}_{1i_1}, \mathbf{Z}_{2i_2})$, $T_4 = \text{median}_{i_1,i_2}(\text{ABS}(\mathbf{Z}_{1i_1}, \mathbf{Z}_{2i_2}))$, $T_5 = \sqrt{\text{SQ}(\hat{\mu}_1, \hat{\mu}_2)}$. (Based on $M = 100$ samples of pseudo-seals with $\epsilon = 10\%$ and $n^p = 30$, and $R = 300$.)

Observe that when the prey base is split, $\sqrt{\mathrm{SQ}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)}$ is large (and similarly for $\mathrm{REL}(\mathrm{E}[\mathbf{Y}_1], \mathrm{E}[\mathbf{Y}_2])$) compared to the distance when the full prey base is used, even though $\boldsymbol{\pi} = \boldsymbol{\pi}_1 = \boldsymbol{\pi}_2$. The P[Type I Error] is accordingly significantly affected by the prey base from which the pseudo-seals are generated. When the prey base is split, the P[Type I Error] is large and we almost always conclude that $\boldsymbol{\pi}_1 \neq \boldsymbol{\pi}_2$. When the full prey base is used, the permutation test works well for all of the test statistics as the P[Type I Error] is similar to α. Recall however that these samples of pseudo-seals constructed with the full prey base may have some prey signatures in common and are not entirely independent. Consequently, if in practice it is known that the two samples of seals are eating from the same prey populations, the P[Type I Error] may be slightly larger than what is shown in Table 6.1.

Note that, based on the distances ($\sqrt{\mathrm{SQ}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)}$ and $\mathrm{REL}(\mathrm{E}[\mathbf{Y}_1], \mathrm{E}[\mathbf{Y}_2])$) which measure the difference between two populations of FA signatures, the permutation test appears to work well as a test of $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. For example, when the prey base is split, we apparently generate two different populations of FA signatures and this is usually detected by our test. If we are actually interested in testing $H_0 : \boldsymbol{\pi}_1 = \boldsymbol{\pi}_2$, then we must interpret a small P-value cautiously. If the seals are eating two different populations of prey, then even if they are eating the prey in the same proportions, we will likely obtain a small P-value. If the seals are eating from the same populations of prey, then it may be appropriate to interpret a small P-value as evidence against $H_0 : \boldsymbol{\pi}_1 = \boldsymbol{\pi}_2$.

The power of the permutation test was also investigated by generating pseudo-seals under the alternative hypothesis in Equation 6.4. We essentially used the simulation algorithm that was used to compute the P[Type I Error], but generated n_{s1} and n_{s2} pseudo-seals with diets $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ respectively, for various choices of $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$. Note that we considered only the case where both samples of pseudo-seals are generated from the full prey base. (When the prey base is split, the power will be high based on our previous results where we found that even if $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2$, H_0 is usually rejected.) For samples of pseudo-seals generated with diets $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ the power is computed as follows:

$$\text{Power} = \beta(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) = 1 - \frac{\#\{p^{\mathrm{perm},m} > \alpha\}}{M}.$$

Tables 6.2 and 6.3 and Figure 6.1 contain the results.

In our simulations, π_1 was fixed to be either Diet 1 or Diet 4 and π_2 was modified in two different ways. We examined the effect of interchanging a non-zero species ($\pi_k > 0$) with a zero species ($\pi_k = 0$), as well as the effect of increasing or decreasing π_k in the non-zero species. (Note that we actually modified the original diet while the tables contain the modified diet adjusted for the noise factor.) Let $\pi_2(1)$ denote the modified diet corresponding to 1 in the tables, and similarly for the other five diets. In both Tables 6.2 and 6.3 diets $\pi_2(1)$ and $\pi_2(2)$ correspond to diets where non-zero and zero species have been interchanged whereas $\pi_2(3) - \pi_2(6)$ contain the same non-zero species as Diet 1 or Diet 4 but with the non-zero components increased or decreased. We have also computed $\text{ABS}(\pi_1, \pi_2)$ and $\text{REL}(\pi_1, \pi_2)$ for the various combinations of π_1 and π_2 as measures of the effect size.

Consider first the effect of interchanging a zero species with a non-zero species. We presume that the power will depend on the similarity between the two species being interchanged as well as the magnitude of their contribution. When $\pi_1 =$ Diet 1 we first interchanged Haddock ($\pi_k = 0.30$) and Plaice ($\pi_k = 0$) and then Pollock ($\pi_k = 0.15$) and Sandlance ($\pi_k = 0$). Even though these pairs of species are similar (recall Figure 3.3), the power is very high for all test statistics when $\alpha \geq 0.05$ and for some test statistics (all but T_2 and T_4 which are based on the median) when $\alpha = 0.01$. When $\pi_1 =$ Diet 4, Sandlance ($\pi_k = 0.45$) and Pollock ($\pi_k = 0$) are interchanged as well as Haddock ($\pi_k = 0$) and Plaice ($\pi_k = 0.09$). In the former case the power is 1 for all test statistics and all levels of α while in the latter case the power is very low. (Realize that because of the 10% noise used in generating the pseudo-seals, the true proportion of Haddock is likely closer to 0.05 so that the change to 0.09 is fairly small.) Our results suggest that when a zero species is interchanged with a non-zero species, the magnitude of the non-zero species has a larger effect on the power than the similarity between the species. It appears to be the case that when a non-zero species with $\pi_k \geq 0.15$ is interchanged with a zero species, the change should be detected by the permutation test with $\alpha \geq 0.05$ and any of the test statistics. Note that the ABS distance measure is a more informative measure of effect size when non-zero and zero species are interchanged.

When non-zero species are increased or decreased ($\pi_2(3)-\pi_2(6)$), where the power is large, it is generally large for all test statistics. There are, however, some exceptions. In Table 6.2, at $\alpha = 0.10$ and $\pi_2(4)$, $T_3 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \text{ABS}(\mathbf{Z}_{1i_1}, \mathbf{Z}_{2i_2})$ gives a more reasonable power than the other test statistics. In Table 6.3, at $\alpha = 0.10$ for $\pi_2(4)$ and $\pi_2(6)$, $T_1 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \text{REL}(\mathbf{Y}_{1i_1}, \mathbf{Y}_{2i_2})$ and $T_2 = \text{median}_{i_1,i_2}\,(\text{REL}(\mathbf{Y}_{1i_1}, \mathbf{Y}_{2i_2}))$ perform better than the other test statistics. To help assess the magnitude of change in the non-zero species that can be detected by this test, a plot of power versus effect size is given in Figure 6.1 at $\alpha = 0.1$ using $T_1 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \text{REL}(\mathbf{Y}_{1i_1}, \mathbf{Y}_{2i_2})$. (Note that we have not plotted the effect size corresponding to $\pi_1(1)$ and $\pi(2)$ where non-zero and zero diets are interchanged.) Overall, it appears that when non-zero species are changed by an absolute difference of roughly 0.20 or more, or a relative difference of one or more, the test will usually conclude that there is a difference in the FA signatures when T_1 is used. A notable exception is in Table 6.2 with $\pi_2(6)$. In this case the absolute difference is 0.252 and the relative difference is 1.10 but the power is only 0.43 (for T_1). This is most probably because the species being increased and decreased are very similar.

6.2.2 Analysis Based on Seal and Prey FA Signatures

In this subsection it is assumed that prey bases containing samples of prey from which the two independent populations of seals are consuming are known and that we are interested in testing

$$\begin{aligned} \text{H}_0 &: \pi_1 = \pi_2 \\ \text{H}_1 &: \pi_1 \neq \pi_2. \end{aligned} \tag{6.5}$$

Let $\mathbf{X}_b$ be the prey base corresponding to seals $\mathbf{Y}_b$, $b = 1, 2$, or let $\mathbf{X}$ denote a common prey base if the seals are eating from the same population of prey. With the prey bases known, we may estimate the diet of each seal using the methods of Chapter 3. Let $\mathbf{p}_{bi}$ be the diet estimate for the ith seal from the bth population and let I denote the dimension of $\mathbf{p}_{bi}$. (Note that I corresponds to the number of species in the prey bases and we assume that, if two prey bases are given, that they contain the same number of species.) Apart from the possible zero estimates, the diet estimates are essentially compositions and the methods of Subsection 6.2.1 can be applied to the

Species	π_1	π_2					
	Diet 1	1	2	3	4	5	6
Cod	0.30	0.30	0.30	0.318	0.345	0.390	0.363
Haddock	0.30	0	0.30	0.318	0.345	0.390	0.237
Plaice	0	0.30	0	0	0	0	0
Pollock	0.15	0.15	0	0.132	0.105	0.060	0.213
Sandlance	0	0	0.15	0	0	0	0
SilverHake	0.15	0.15	0.15	0.132	0.105	0.060	0.087
WinterFlounder	0	0	0	0	0	0	0
YellowTail	0	0	0	0	0	0	0
$\text{ABS}(\pi_1, \pi_2)$		0.600	0.300	0.072	0.180	0.360	0.252
$\text{REL}(\pi_1, \pi_2)$		2.00	2.00	0.353	0.861	1.66	1.10
$\alpha = 0.01$	T_1	1.00	0.87	0.05	0.28	0.81	0.11
	T_2	1.00	0.67	0.02	0.20	0.60	0.09
	T_3	1.00	0.89	0.09	0.51	0.90	0.05
	T_4	1.00	0.73	0.07	0.44	0.79	0.06
	T_5	1.00	0.90	0.05	0.39	0.84	0.03
$\alpha = 0.05$	T_1	1.00	0.98	0.14	0.56	0.95	0.31
	T_2	1.00	0.89	0.08	0.37	0.84	0.21
	T_3	1.00	0.99	0.23	0.74	1.00	0.18
	T_4	1.00	0.88	0.20	0.60	0.93	0.16
	T_5	1.00	0.99	0.19	0.67	0.96	0.22
$\alpha = 0.10$	T_1	1.00	0.99	0.24	0.73	0.99	0.43
	T_2	1.00	0.92	0.15	0.53	0.93	0.34
	T_3	1.00	0.99	0.35	0.86	1.00	0.33
	T_4	1.00	0.93	0.29	0.70	0.98	0.25
	T_5	1.00	0.99	0.27	0.76	0.99	0.32

Table 6.2: Power, $\beta(\pi_1, \pi_2)$, of the multivariate permutation test for $\pi_1 =$ Diet 1 and various choices of π_2, at $n_{s1} = n_{s2} = 10$. $T_1 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \text{REL}(\mathbf{Y}_{1i_1}, \mathbf{Y}_{2i_2})$, $T_2 = \text{median}_{i_1,i_2}(\text{REL}(\mathbf{Y}_{1i_1}, \mathbf{Y}_{2i_2}))$, $T_3 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \text{ABS}(\mathbf{Z}_{1i_1}, \mathbf{Z}_{2i_2})$, $T_4 = \text{median}_{i_1,i_2}(\text{ABS}(\mathbf{Z}_{1i_1}, \mathbf{Z}_{2i_2}))$, $T_5 = \sqrt{\text{SQ}(\hat{\mu}_1, \hat{\mu}_2)}$. (Based on $M = 100$ samples of pseudo-seals with $\epsilon = 10\%$ and $n^p = 30$, and $R = 300$.)

Species	π_1	π_2					
	Diet 4	1	2	3	4	5	6
Cod	0.09	0.09	0.09	0.099	0.108	0.117	0.072
Haddock	0	0	0.09	0	0	0	0
Plaice	0.09	0.09	0	0.099	0.108	0.117	0.072
Pollock	0	0.45	0	0	0	0	0
Sandlance	0.45	0	0.45	0.405	0.360	0.315	0.540
SilverHake	0.09	0.09	0.09	0.099	0.108	0.117	0.072
WinterFlounder	0.09	0.09	0.09	0.099	0.108	0.117	0.072
YellowTail	0.09	0.09	0.09	0.099	0.108	0.117	0.072
$\text{ABS}(\pi_1, \pi_2)$		0.900	0.180	0.009	0.180	0.270	0.180
$\text{REL}(\pi_1, \pi_2)$		2.00	2.00	0.555	1.033	1.454	1.167
$\alpha = 0.01$	T_1	1.00	0.03	0.03	0.56	0.97	0.77
	T_2	1.00	0.02	0.04	0.48	0.83	0.60
	T_3	1.00	0.02	0.04	0.14	0.34	0.11
	T_4	1.00	0.00	0.02	0.12	0.32	0.07
	T_5	1.00	0.03	0.01	0.14	0.32	0.13
$\alpha = 0.05$	T_1	1.00	0.12	0.08	0.78	1.00	0.95
	T_2	1.00	0.10	0.13	0.70	0.98	0.79
	T_3	1.00	0.11	0.08	0.36	0.75	0.54
	T_4	1.00	0.05	0.06	0.34	0.63	0.37
	T_5	1.00	0.12	0.07	0.31	0.66	0.47
$\alpha = 0.10$	T_1	1.00	0.19	0.13	0.96	1.00	1.00
	T_2	1.00	0.15	0.17	0.77	1.00	0.94
	T_3	1.00	0.18	0.12	0.55	0.90	0.75
	T_4	1.00	0.11	0.09	0.51	0.81	0.59
	T_5	1.00	0.18	0.13	0.47	0.87	0.68

Table 6.3: Power, $\beta(\pi_1, \pi_2)$, of the multivariate permutation test for $\pi_1 =$ Diet 4 and various choices of π_2, at $n_{s1} = n_{s2} = 10$. $T_1 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \text{REL}(\mathbf{Y}_{1i_1}, \mathbf{Y}_{2i_2})$, $T_2 = \text{median}_{i_1,i_2}(\text{REL}(\mathbf{Y}_{1i_1}, \mathbf{Y}_{2i_2}))$, $T_3 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \text{ABS}(\mathbf{Z}_{1i_1}, \mathbf{Z}_{2i_2})$, $T_4 = \text{median}_{i_1,i_2}(\text{ABS}(\mathbf{Z}_{1i_1}, \mathbf{Z}_{2i_2}))$, $T_5 = \sqrt{\text{SQ}(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2)}$. (Based on $M = 100$ samples of pseudo-seals with $\epsilon = 10\%$ and $n^p = 30$, and $R = 300$.)

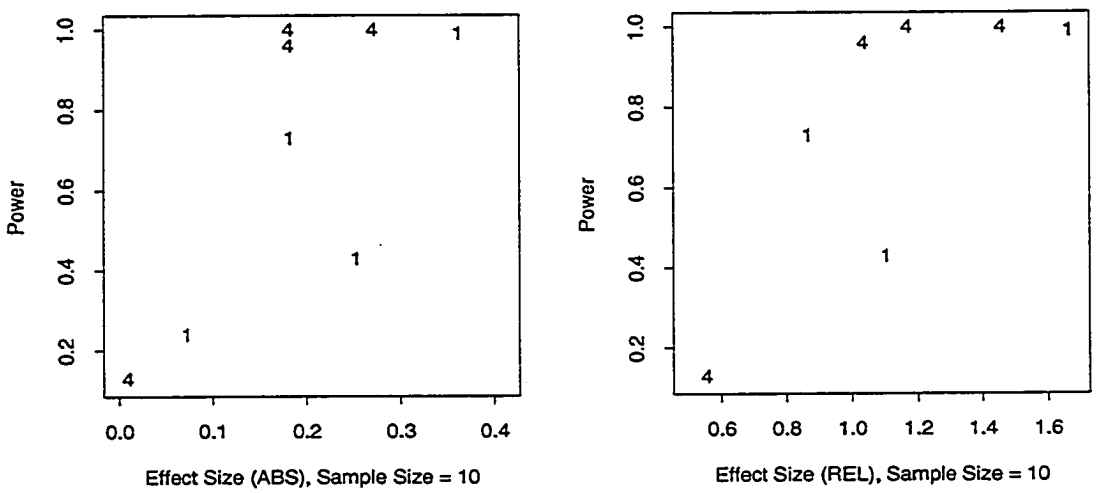

Figure 6.1: Plots of power versus effect size at $\alpha = 0.1$, using $T_1 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \mathrm{REL}(\mathbf{Y}_{1i_1}, \mathbf{Y}_{2i_2})$ with $n_{s1} = n_{s2} = 10$. Effect size is measured as $\mathrm{ABS}(\pi_1, \pi_2)$ or $\mathrm{REL}(\pi_1, \pi_2)$. π_1 =Diet 1 is denoted by 1 and π_2 =Diet 4 is denoted by 4.

diet estimates, as they were to the FA signatures, with adjustments to manage the zeros.

Because of the essential zeros in the diet estimates, it is not clear how to even extend the large sample procedure discussed in Subsection 6.2.1. A P-value can easily be computed, however, with the multivariate permutation test if a suitable test statistic is used and if n_{s1} and n_{s2} are greater than one. Based on the satisfactory results obtained in the previous subsection using test statistics involving relative distances (even without a log transformation), we will use the following two test statistics

1. $T_1 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \text{REL}(\mathbf{p}_{1i_1}, \mathbf{p}_{2i_2})$
2. $T_2 = \text{median}_{i_1,i_2} \left(\text{REL}(\mathbf{p}_{1i_1}, \mathbf{p}_{2i_2})\right)$.

To examine the performance of the multivariate permutation test in testing Equation 6.5 using samples of diet estimates, we again carried out simulations using pseudo-seals. The following algorithm was used to generate samples of pseudo-seals with diets $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ and to estimate their diets:

1. *for* $i = 1, \ldots, n_{s1}$
   (a) Randomly and evenly split prey base into $\mathbf{X}_1$ and $\mathbf{X}_2$.
   (b) Generate a pseudo-seal, $\mathbf{Y}_{1i}$, from $\mathbf{X}_1$ with diet $\boldsymbol{\pi}_1$.
   (c) Estimate the diet of the pseudo-seal using $\mathbf{X}_2$ to obtain $\mathbf{p}_{1i}(\mathbf{Y}_{1i}, \bar{\mathbf{X}}_2)$.
2. *for* $i = 1, \ldots, n_{s2}$
   (a) Randomly and evenly split prey base into $\mathbf{X}_1$ and $\mathbf{X}_2$.
   (b) Generate a pseudo-seal, $\mathbf{Y}_{2i}$, from $\mathbf{X}_1$ with diet $\boldsymbol{\pi}_2$.
   (c) Estimate the diet of the pseudo-seal using $\mathbf{X}_2$ to obtain $\mathbf{p}_{2i}(\mathbf{Y}_{2i}, \bar{\mathbf{X}}_2)$

We first carried out simulations to compute the P[Type I Error] by setting $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2$ in the above algorithm. To date we have only carried out simulations with the AIT distance measure. Table 6.4 contains the results.

			P[Type I Error]		
π_1	π_2	T	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
Diet 1	Diet 1	T_1	0.00	0.06	0.12
		T_2	0.01	0.06	0.13
Diet 4	Diet 4	T_1	0.00	0.05	0.16
		T_2	0.01	0.06	0.16

Table 6.4: P[Type I Error] associated with the multivariate permutation test at $n_{s1} = n_{s2} = 10$. $T_1 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \text{REL}(\mathbf{p}_{1i_1}, \mathbf{p}_{2i_2})$, $T_2 = \text{median}_{i_1,i_2}(\text{REL}(\mathbf{p}_{1i_1}, \mathbf{p}_{2i_2}))$. (Based on $M = 100$ samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients, and $R = 300$.)

At $\alpha = 0.01$ and $\alpha = 0.05$, the P[Type I Error] is close to α but at $\alpha = 0.10$, is slightly large, particularly for Diet 4. Both test statistics appear to yield roughly equivalent results. Note that our samples of diet estimates are not completely independent (since there may be some overlap in the prey used to generate the seals) and, in practice, the P[Type I Error] may be larger than that given in Table 6.4.

Although not shown, we also carried out simulations using modifications of our algorithm for generating the seals and diet estimates. One such modification involved splitting the prey base into $\mathbf{X}_1$ and $\mathbf{X}_2$ and generating n_{s1} seals from $\mathbf{X}_1$ with diet π_1, and n_{s2} seals from $\mathbf{X}_2$ with diet π_2. We then used $\mathbf{X}_1$ to estimate the diet of the n_{s1} seals and $\mathbf{X}_2$ to estimate the diet of the n_{s2} seals. The P[Type I Error] was surprisingly quite high when the simulations were carried out in this manner with $\pi_1 = \pi_2$. This was likely due to using the same split to generate all seals since the P[Type I Error] was much more reasonable when a different split was used to generate each seal as we have done.

Power calculations were also carried out and the results are given in Tables 6.5-6.6 and Figure 6.2. We again examined the effect of interchanging a zero and non-zero species and also the effect of increasing or decreasing the non-zero species. Diets $\pi_2(1)$ and $\pi_2(2)$ have a non-zero and zero species interchanged while diets $\pi_2(3) - \pi_2(8)$ contain modifications of the non-zero species in diet π_1. Observe that while $\pi_2(1)$ and $\pi_2(2)$ are as before, the various modifications $\pi_2(3) - \pi_2(8)$ are much "farther" from π_1 than in the previous subsection as we found that the power was not as good with the diet estimates. We surmise that this is due to the diet estimates being

more variable than the FA signatures. For this reason we examined two sample sizes ($n_{sb} = 10$ and $n_{sb} = 30$, $b = 1, 2$) to determine whether a decent power is obtained at the larger sample size. Note that $n_{s1} = n_{s2} = 30$ might still be considered relatively small since the dimension of the diet estimates, I, will usually be large. (In our case I=8.)

At $n_{s1} = n_{s2} = 10$, when the non-zero and zero species are interchanged, it now appears to be the case that a difference will usually only be detected when a non-zero species with $\pi_k \geq 0.30$ is interchanged with a zero species. When Haddock (0.30) and Plaice (0) are interchanged in Table 6.5 (and similarly for Pollock, $\pi_k = 0$, and Sandlance, $\pi_k = 0.45$, in Table 6.6) the power is lower than before but still adequate at $\alpha = 0.1$ when $T_1 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \text{REL}(\mathbf{p}_{1i_1}, \mathbf{p}_{2i_2})$ is used. At $n_{s1} = n_{s2} = 30$, the power is much higher. In particular, when Pollock ($\pi_k = 0.15$) and Sandlance ($\pi_k = 0$) are interchanged, the power increases from 0.29 to 0.70 at $n_{s1} = n_{s2} = 30$. We might therefore conclude that at sample sizes of 30, if a non-zero species with $\pi_k > 0.15$ is interchanged with a zero species, we will usually detect the change.

The effect on the power when the non-zero species are increased or decreased is shown in Figure 6.2 at $\alpha = 0.1$ and using $T_1 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \text{REL}(\mathbf{p}_{1i_1}, \mathbf{p}_{2i_2})$. (Note that T_1 usually yields a higher power than T_2.) It appears that a relative distance of at least three is needed to ensure a power of roughly 0.8 or more at $n_{s1} = n_{s2} = 10$. In terms of the absolute difference, when $\pi_1 =$ Diet 4, an absolute difference of about 0.6 is needed while for $\pi_1 =$ Diet 1, a difference of over 1 is needed. Recall that Diet 1 was originally chosen to be a difficult case since the non-zero species are similar in their FA signatures and this is likely affecting the power. By $n_{s1} = n_{s2} = 30$, an absolute difference of about 0.54 is usually detected as is a relative difference of approximately 2.2. (This is not the case when $\pi_1 =$ Diet 4 and $\pi_2 = \pi(6)$ since the relative difference is 2.833 but the power is only 0.68.)

6.2.3 Conclusions

The results of our simulation studies suggest that the multivariate permutation test is a very useful test for testing for a difference in the FA signatures of seals at relatively small sample sizes of 10. If the prey from which the samples of seals are eating are

Species	π_1	π_2							
	Diet 1	1	2	3	4	5	6	7	8
Cod	0.30	0.30	0.30	0.345	0.435	0.435	0.45	0.12	0.03
Haddock	0.30	0	0.30	0.345	0.435	0.165	0.09	0.12	0.03
Plaice	0	0.30	0	0	0	0	0	0	0
Pollock	0.15	0.15	0	0.105	0.015	0.285	0.36	0.33	0.42
Sandlance	0	0	0.15	0	0	0	0	0	0
SilverHake	0.15	0.15	0.15	0.105	0.015	0.015	0	0.33	0.42
WinterFlounder	0	0	0	0	0	0	0	0	0
YellowTail	0	0	0	0	0	0	0	0	0
ABS(π_1, π_2)		0.600	0.300	0.180	0.540	0.540	0.720	0.720	1.080
RELdist(π_1, π_2)		2.000	2.000	0.861	2.421	2.134	2.617	2.291	3.086
$\alpha = 0.01$	T_1	0.46	0.10	0.01	0.12	0.10	0.24	0.10	0.63
		1.00	0.38	0.03	0.51	0.42	0.87	0.80	1.00
	T_2	0.32	0.08	0.01	0.08	0.07	0.18	0.08	0.47
		0.98	0.28	0.03	0.33	0.26	0.76	0.63	0.99
$\alpha = 0.05$	T_1	0.73	0.21	0.10	0.3	0.25	0.46	0.34	0.87
		1.00	0.61	0.14	0.83	0.72	0.99	0.96	1.00
	T_2	0.59	0.19	0.08	0.15	0.19	0.32	0.29	0.75
		1.00	0.41	0.15	0.60	0.63	0.89	0.85	1.00
$\alpha = 0.10$	T_1	0.85	0.29	0.18	0.46	0.36	0.61	0.45	0.92
		1.00	0.70	0.26	0.93	0.85	0.99	1.00	1.00
	T_2	0.70	0.26	0.19	0.29	0.36	0.49	0.42	0.84
		1.00	0.54	0.22	0.74	0.73	0.96	0.89	1.00

Table 6.5: Power, $\beta(\pi_1, \pi_2)$, of the multivariate permutation test for π_1 = Diet 1 and various choices of π_2, and at two samples sizes: $n_{s1} = n_{s2} = 10$ (first row) and $n_{s1} = n_{s2} = 30$ (second row). $T_1 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \text{REL}(\mathbf{p}_{1i_1}, \mathbf{p}_{2i_2})$, $T_2 = \text{median}_{i_1,i_2}(\text{REL}(\mathbf{p}_{1i_1}, \mathbf{p}_{2i_2}))$. (Based on $M = 100$ samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients, and $R = 300$.)

Species	π_1	π_2							
	Diet 4	1	2	3	4	5	6	7	8
Cod	0.09	0.09	0.09	0.153	0.162	0.171	0.045	0.036	0.027
Haddock	0	0	0.09	0	0	0	0	0	0
Plaice	0.09	0.09	0	0.153	0.162	0.171	0.045	0.036	0.027
Pollock	0	0.45	0	0	0	0	0	0	0
Sandlance	0.45	0	0.45	0.135	0.09	0.045	0.675	0.72	0.765
SilverHake	0.09	0.09	0.09	0.153	0.162	0.171	0.045	0.036	0.027
WinterFlounder	0.09	0.09	0.09	0.153	0.162	0.171	0.045	0.036	0.027
YellowTail	0.09	0.09	0.09	0.153	0.162	0.171	0.045	0.036	0.027
ABS(π_1, π_2)		0.900	0.180	0.630	0.720	0.810	0.450	0.540	0.630
RELdist(π_1, π_2)		2.000	2.000	2.759	3.022	3.268	2.833	3.375	3.912
$\alpha = 0.01$	T_1	0.51	0	0.27	0.29	0.49	0.06	0.32	0.32
		1.00	0.01	0.98	1.00	1.00	0.22	0.99	0.99
	T_2	0.40	0	0.16	0.15	0.38	0.04	0.16	0.19
		0.99	0.02	0.90	0.97	0.99	0.19	0.93	0.98
$\alpha = 0.05$	T_1	0.72	0.07	0.55	0.62	0.79	0.18	0.66	0.6
		1.00	0.03	1.00	1.00	1.00	0.49	1.00	1.00
	T_2	0.62	0.05	0.40	0.47	0.67	0.10	0.43	0.40
		1.00	0.05	0.95	1.00	1.00	0.39	0.98	0.99
$\alpha = 0.1$	T_1	0.83	0.12	0.70	0.77	0.88	0.28	0.81	0.77
		1.00	0.13	1.00	1.00	1.00	0.68	1.00	1.00
	T_2	0.7	0.13	0.54	0.67	0.79	0.18	0.64	0.51
		1.00	0.08	0.98	1.00	1.00	0.53	1.00	1.00

Table 6.6: Power, $\beta(\pi_1, \pi_2)$, of the multivariate permutation test for π_1 = Diet 4 and various choices of π_2, and at two sample sizes: $n_{s1} = n_{s2} = 10$ (first row) and $n_{s1} = n_{s2} = 30$ (second row). $T_1 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \text{REL}(\mathbf{p}_{1i_1}, \mathbf{p}_{2i_2})$, $T_2 = \text{median}_{i_1,i_2}(\text{REL}(\mathbf{p}_{1i_1}, \mathbf{p}_{2i_2}))$. (Based on $M = 100$ samples of pseudo-seals with $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients, and $R = 300$.)

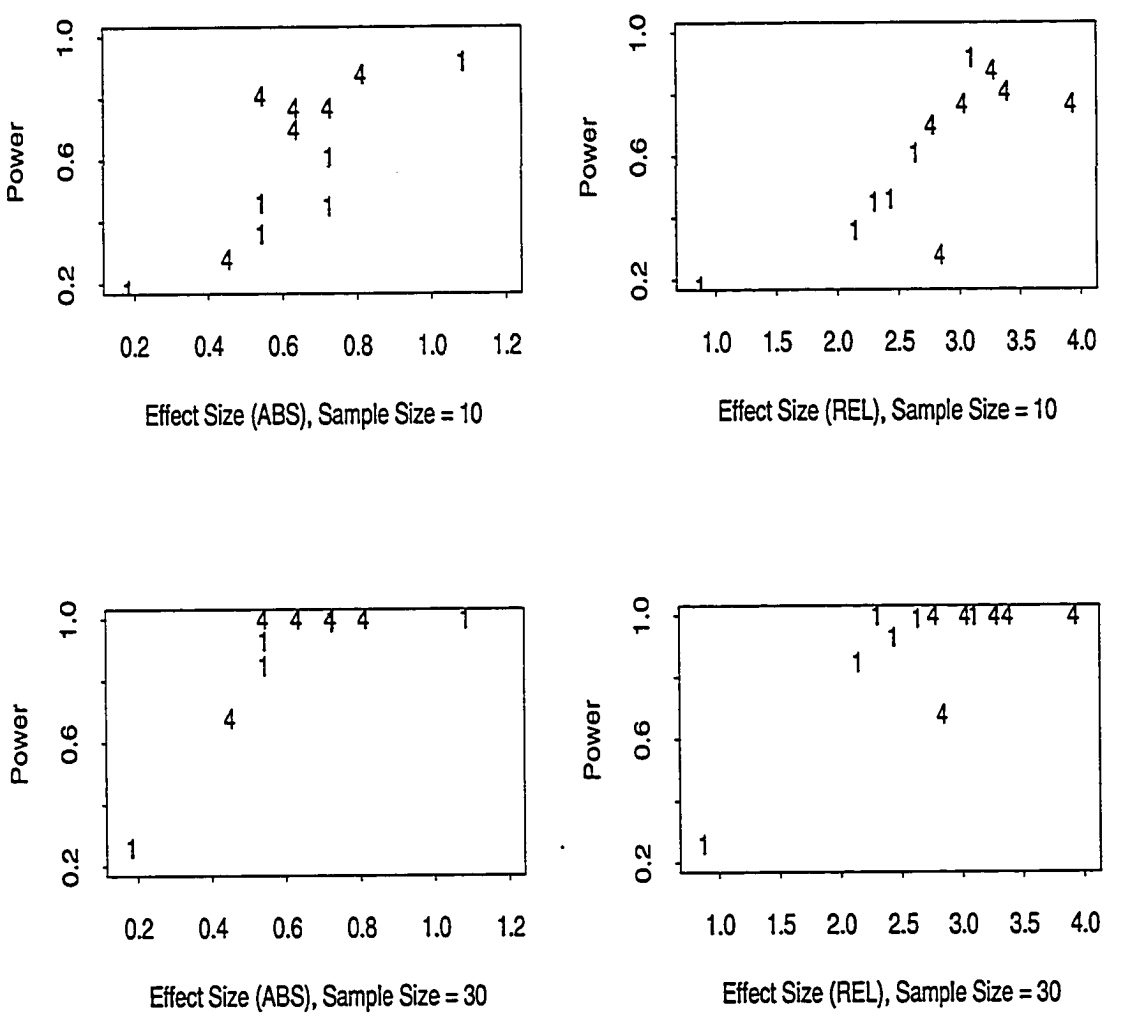

Figure 6.2: Plots of power versus effect size at $\alpha = 0.1$, using $T_1 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \mathrm{REL}(\mathbf{p}_{1i_1}, \mathbf{p}_{2i_2})$ with two sample sizes: $n_{s1} = n_{s2} = 10$ and $n_{s1} = n_{s2} = 30$. Effect size is measured as $\mathrm{ABS}(\pi_1, \pi_2)$ or $\mathrm{REL}(\pi_1, \pi_2)$. π_1 =Diet 1 is denoted by 1 and π_2 =Diet 4 is denoted by 4.

essentially the same then this test may also be used to test for a difference in diet, given only the signatures. Otherwise, to test for a difference in diet, a prey base is needed to estimate the diet estimates and the multivariate permutation test can be applied to the diet estimates. In this case, although the P[Type I Error] was reasonable for sample sizes of 10, the test was not particularly powerful. Sample sizes close to 30 appear to be needed to obtain decent power when the diet estimates are used. While a variety a test statistics can be used with the multivariate permutation test, we recommend the sum of relative differences on the untransformed compositions, namely $T_1 = \sum_{i_1=1}^{n_{s1}} \sum_{i_2=1}^{n_{s2}} \mathrm{REL}(\cdot, \cdot)$.

6.3 Paired Comparison

6.3.1 Analysis Based on Seal FA Signatures Only

We now consider the case where we are given paired samples of FA signatures, say $\mathbf{Y}_{Bi}$ and $\mathbf{Y}_{Ai}$, $i = 1, \ldots, n_s$, with respective true diets π_B and π_A. (B =Before, A =After.) Recall that the dimension of the signatures is denoted by n_{FA} and $N_{FA} = n_{FA} - 1$. As in Subsection 6.2.1, given only the FA signatures, we will actually test for a difference in the before and after FA signatures and, through simulations, will examine the extent to which the test may be used to test for a change in the diets of the seals. Let

$$\begin{aligned} \mathbf{Z}_B &= \log\left(\frac{\mathbf{Y}_{B-n_{FA}}}{Y_{Bn_{FA}}}\right), \\ \mathbf{Z}_A &= \log\left(\frac{\mathbf{Y}_{A-n_{FA}}}{Y_{An_{FA}}}\right), \\ \mathbf{D} &= \mathbf{Z}_A - \mathbf{Z}_B, \end{aligned}$$

and consider testing

$$\begin{aligned} \mathrm{H}_0 &: \boldsymbol{\mu}_D = 0 \\ \mathrm{H}_1 &: \boldsymbol{\mu}_D \neq 0, \end{aligned} \tag{6.6}$$

where $\boldsymbol{\mu}_D = \mathrm{E}[\mathbf{D}]$. Let $\mathbf{D}_i = \mathbf{Z}_{Ai} - \mathbf{Z}_{Bi}$, if $n_s > N_{FA}$ and

$$\mathbf{D} \sim \mathcal{N}^{N_{FA}}(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D),$$

then standard multivariate procedures can be applied and we would reject H_0 if

$$T^2 = n_s \bar{\mathbf{D}}' \mathbf{S}_{\mathbf{D}}^{-1} \bar{\mathbf{D}} > \frac{(n_s - 1)N_{FA}}{n_s - N_{FA}} F_{N_{FA}, n_s - N_{FA}}(1 - \alpha), \tag{6.7}$$

where $\bar{\mathbf{D}} = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbf{D}_i$ and $\mathbf{S}_{\mathbf{D}} = \frac{1}{n_s - 1} \sum_{i=1}^{n_s} (\mathbf{D}_i - \bar{\mathbf{D}})(\mathbf{D}_i - \bar{\mathbf{D}})'$. If $n_s - N_{FA}$ is large then the normality assumption is not needed and $\frac{(n_s-1)N_{FA}}{n_s - N_{FA}} F_{N_{FA}, n_s - N_{FA}}(1 - \alpha) \approx \chi^2_{N_{FA}}(1 - \alpha)$.

It should be mentioned that Aitchison and Ng (2003) use a different but equivalent approach for the general problem of testing for change in compositions. They argue that perturbations are the appropriate measure of difference in compositions. Recall that if $\mathbf{P}$ and $\mathbf{V}$ are two D-part compositions then the perturbation $\mathbf{P} \circ \mathbf{V}$ is defined by

$$\mathbf{U} = \mathbf{P} \circ \mathbf{V} = \mathcal{C}\left[P_1 V_1, \ldots, P_D V_D\right],$$

where $\mathcal{C}$ is the constraining operator that divides each component by the sum of the components. The inverse operation is then

$$\mathbf{P} = \mathbf{U} \ominus \mathbf{V} = \mathcal{C}\left[\frac{U_1}{V_1}, \ldots, \frac{U_D}{V_D}\right].$$

To test for a difference in the before and after FA signatures, Aitchison and Ng (2003) consider the hypotheses

$$\begin{aligned} H_0 &: \mu_Q = 0 \\ H_1 &: \mu_Q \neq 0, \end{aligned} \tag{6.8}$$

where $\mu_Q = \mathrm{E}\left[\log\left(\frac{P_{-n_{FA}}}{P_{n_{FA}}}\right)\right]$ and $\mathbf{P} = \mathbf{Y}_B \ominus \mathbf{Y}_A$. It is straightforward to show that $\mu_Q = \mu_D$.

For $n_s < N_{FA}$, dimension reduction techniques such as principal component analysis could be applied to the transformed differences. Alternatively, Aitchison's (1986) methods such as log contrast principal component analysis or subcompositional analysis could be applied to the perturbations. We would then apply the test in Equation 6.7 to the reduced data. A drawback to this method is that it still requires the multivariately normal distribution assumption if n_s is small relative to the reduced dimension.

Another approach to testing Equation 6.9 which we will call the "Regression Method" involves fitting n_s straight lines to the n_s before and after (transformed) signatures. That is, for $i = 1 \ldots, n_s$, assume that

$$Z_{Bij} = \beta_{0i} + \beta_{1i} Z_{Aij} + \epsilon_{ij},$$

for the pairs $(Z_{Bij}, Z_{Aij}) = \left(\log\left(\frac{Y_{Bij}}{Y_{Bijn_{FA}}}\right), \log\left(\frac{Y_{Aij}}{Y_{Aijn_{FA}}}\right)\right)$. For this analysis, orthogonal least squares (as opposed to ordinary least squares) is used to estimate β_{0i} and β_{1i} so that the $\mathbf{Z}_{Ai}$ and $\mathbf{Z}_{Bi}$ variables are treated equally. If we let $\mathbf{b}_i = [b_{0i}, b_{1i}]$ be the vector containing the estimates of β_{0i} and β_{1i} respectively then we will test

$$\mathrm{H}_0 : \begin{pmatrix} \mu_{b_0} \\ \mu_{b_1} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\mathrm{H}_1 : \begin{pmatrix} \mu_{b_0} \\ \mu_{b_1} \end{pmatrix} \neq \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

using

$$T^2 = n_s \left(\bar{\mathbf{b}} - [0,1]\right)' \mathbf{S}_{\mathbf{b}}^{-1} \left(\bar{\mathbf{b}} - [0,1]\right),$$

where $\bar{\mathbf{b}} = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbf{b}_i$ and $\mathbf{S}_{\mathbf{b}} = \frac{1}{n_s - 1} \sum_{i=1}^{n_s} (\mathbf{b}_i - \bar{\mathbf{b}})(\mathbf{b}_i - \bar{\mathbf{b}})'$. We reject H_0 if $T^2 > \frac{(n_s-1)2}{n_s-2} F_{2,n_s-2}(1-\alpha)$. Again, unless $n_s - 2$ is large, the appropriateness of the test relies on $\mathbf{b}$ being multivariately normal.

We have also considered an extension to the univariate matched-pair randomization P-value discussed in Problem 7 (page 186) in Davison and Hinkley (1997). Their univariate method involves re-sampling under the null hypothesis of no difference by computing $D_i^* = S_i D_i$, $i = 1, \ldots, n_s$ where D_i are the computed differences in the original data set and S_i are independent and equally likely to be +1 and -1. For each generated sample, a suitable test statistic is computed.

In the multivariate setting, while we could compute $D_{ij}^* = S_{ij} D_{ij}$ for each FA, we would then not be taking into account the possible correlation between the variables. For example, if the transformed FAs j and k were positively correlated, then we might expect d_{ij} and d_{ik} to have the same sign. If they were negatively correlated, they would likely have opposite signs. Therefore, to preserve the signs, we have

chosen to compute $\mathbf{D}_i^* = S_i\mathbf{D}_i$. Note that when the untransformed data is used (such as when essential zeros are present), we will actually use relative differences. That is, let $\mathbf{R}_i$ have jth component

$$r_{ij} = \begin{cases} 0 & \text{if } Y_{Bij} = Y_{Aij}, \\ \frac{Y_{Aij} - Y_{Bij}}{\max(Y_{Aij}, Y_{Bij})} & \text{otherwise,} \end{cases}$$

Also, let $\bar{\mathbf{R}} = \frac{1}{n_s}\sum_{i=1}^{n_s}\mathbf{R}_i$.

Possibilities for a test statistic in the multivariate case are:

1. $T_1 = \sum_{j=1}^{n_{FA}} |\bar{R}_j(\mathbf{Y})|$
2. $T_2 = \sum_{j=1}^{n_{FA}} |\text{median}_i(R_j(\mathbf{Y}))|$
3. $T_3 = \sum_{j=1}^{N_{FA}} |\bar{D}_j(\mathbf{Z})|$
4. $T_4 = \sum_{j=1}^{N_{FA}} |\text{median}_i(D_j(\mathbf{Z}))|$
5. $T_5 = \sqrt{\sum_{j=1}^{N_{FA}} \bar{D}_j^2(\mathbf{Z})}$

where $R_j(\mathbf{Y})$ denotes that the (relative) differences are taken on the original data while $D_j(\mathbf{Z})$ indicates differences in the transformed data. Recall also that $N_{FA} = n_{FA} - 1$.

In summary, our multivariate randomization P-value is computed using the following algorithm

Multivariate Randomization Test

1. Compute the differences and denote these by $\mathbf{D}(\cdot)$ or $\mathbf{R}(\cdot)$.
2. Compute the test statistic, T, using differences in 1.
3. *for* $r = 1, \ldots, R$
   (a) For the ith observation, randomly select +1 or -1 and call this S_i^{r*} $i = 1, \ldots, n_s$.
   (b) Compute $\mathbf{D}_i^{*r} = S_i^{r*}\mathbf{D}_i$, where $\mathbf{D}_i$ denotes the ith row of $\mathbf{D}$, $i = 1, \ldots, n_s$. (And similarly for $\mathbf{R}_i$.)

(c) Compute T^{*r} using $\mathbf{D}^{*r}(\cdot)$. (And similarly for $\mathbf{R}$.)

4. Compute

$$p^{\text{rand}} = \frac{\#\{T^{*r} \geq T\}}{R}.$$

To evaluate and compare the principal component (PC), regression (REG) and randomization (RAND) P-value methods through simulations, it is necessary to be able to generate paired samples of pseudo-seals. While there may be various ways of accomplishing this, we have chosen to adjust our "after" seals as follows: let

$$\mathbf{Y}_A^* = (1-\delta)\mathbf{Y}_A + \delta\mathbf{Y}_B,$$

where $\mathbf{Y}_B$ and $\mathbf{Y}_A$ are the "before" and "after" pseudo-seals with diets π_B and π_A respectively, and $\mathbf{Y}_A^*$ is our adjusted "after" seal. In our simulations, we set δ equal to 0.2 and 0.5.

As in Subsection 6.2.1, simulations were carried out to determine whether, given only the seal FA signatures, the above described tests would be useful in testing

$$\begin{aligned} \mathrm{H}_0 &: \pi_B = \pi_A \\ \mathrm{H}_1 &: \pi_B \neq \pi_A. \end{aligned} \tag{6.9}$$

Note that we did not consider splitting the prey base since in Section 6.2 it was concluded that if the prey populations from which the seals were eating were different, then the P[Type I Error] would be large and we expect similar results in the paired comparison case. Additionally, we now do not require independent samples. Note further that by not splitting the prey base, we are essentially assessing the usefulness of the methods in testing for a difference in the FA signatures. That is, testing for a difference in FA signatures should be roughly equivalent to testing for a difference in the diets when the prey populations are similar.

The following simulation algorithm was used to compute the P[Type I Error] and/or the power, $\beta(\pi_B, \pi_A)$, associated with our methods:

1. *for* $m = 1, \ldots, M$

   (a) Choose π_B and π_A.

(b) Using the full prey base, generate n_s pseudo-seals with diet π_B, $\mathbf{Y}_{B1}, \ldots, \mathbf{Y}_{Bn_s}$, and n_s pseudo-seals with diet π_A, $\mathbf{Y}_{A1}, \ldots, \mathbf{Y}_{An_s}$.

(c) Adjust the "after" seals: $\mathbf{Y}^*_{Ai} = (1-\delta)\mathbf{Y}_{Ai} + \delta\mathbf{Y}_{Bi}$, $i = 1, \ldots, n_s$.

(d) Compute the PC, REG, or RAND P-value, p^m, using $\mathbf{Y}_{B1}, \ldots, \mathbf{Y}_{Bn_s}$ and $\mathbf{Y}^*_{A1}, \ldots, \mathbf{Y}^*_{An_s}$.

2. If $\pi_B = \pi_A$, compute

$$P[\text{Type I Error}] = \frac{\#\{p^m \leq \alpha\}}{M}.$$

If $\pi_B \neq \pi_A$, compute

$$\text{Power} = \beta(\pi_B, \pi_A) = 1 - \frac{\#\{p^m > \alpha\}}{M}.$$

We chose to use $\alpha = 0.01,\ 0.05$ and 0.1, $\delta = 0.2$ and 0.5, $M = 100$, and $n_s = 10$. The results are given in Tables 6.7-6.11 and in Figure 6.3.

π_1	π_2	# PCs	$\delta = 0.20$	$\delta = 0.50$
Diet 1	Diet 1	3	0.786	0.788
		4	0.869	0.870
Diet 4	Diet 4	3	0.814	0.814
		4	0.882	0.883

Table 6.7: Mean proportion of total variance explained by first 3 and 4 principal components with $n_s = 10$. (Based on $M = 100$ samples of pseudo-seals with $\epsilon = 10\%$ and $n^p = 30$.)

Firstly, the results corresponding to $\delta = 0.2$ and $\delta = 0.5$ were almost identical and we have therefore not chosen to illustrate the power results for $\delta = 0.2$. Also, although from Table 6.7, three and four principal components explained a large proportion of the total variance, the PC method is consistently out-performed by the other methods. Consequently, the comments that follow are focused primarily on the REG and RAND methods. Based on our simulation results, the PC method is not recommended for small sample sizes.

The $P[\text{Type I Error}] \approx \alpha$ (Tables 6.8 and 6.9) for both the REG and RAND methods. Apart from $T_5 = \sqrt{\sum_{j=1}^{N_{FA}} \bar{D}_j^2(\mathbf{Z})}$ giving a slightly small $P[\text{Type I Error}]$

		P[Type I Error]		
δ	Method	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
$\delta = 0.2$	3 PC	0	0	0
	4 PC	0	0	0
	REG	0.02	0.04	0.10
	RAND - T_1	0.01	0.05	0.10
	RAND - T_2	0.01	0.03	0.09
	RAND - T_3	0.01	0.05	0.11
	RAND - T_4	0.01	0.04	0.08
	RAND - T_5	0.01	0.02	0.07
$\delta = 0.5$	3 PC	0	0	0
	4 PC	0	0	0
	REG	0.02	0.04	0.10
	RAND - T_1	0.01	0.05	0.08
	RAND - T_2	0.01	0.03	0.10
	RAND - T_3	0.01	0.04	0.11
	RAND - T_4	0.01	0.04	0.09
	RAND - T_5	0.01	0.03	0.05

Table 6.8: $\pi_B = \pi_A =$ Diet 1: P[Type I Error] for PC,REG and RAND methods at $n_s = 10$. $T_1 = \sum_{j=1}^{n_{FA}} |\bar{R}_j(\mathbf{Y})|$, $T_2 = \sum_{j=1}^{n_{FA}} |\text{median}_i(R_j(\mathbf{Y}))|$, $T_3 = \sum_{j=1}^{N_{FA}} |\bar{D}_j(\mathbf{Z})|$, $T_4 = \sum_{j=1}^{N_{FA}} |\text{median}_i(D_j(\mathbf{Z}))|$, and $T_5 = \sqrt{\sum_{j=1}^{N_{FA}} \bar{D}_j(\mathbf{Z})^2}$. (Based on $M = 100$ samples of pseudo-seals with $\epsilon = 10\%$ and $n^p = 30$, and $R = 300$.)

		P[Type I Error]		
δ	Method	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
0.2	3 PC	0	0	0.01
	4 PC	0	0	0
	REG	0	0.02	0.08
	RAND - T_1	0.02	0.08	0.12
	RAND - T_2	0.03	0.04	0.12
	RAND - T_3	0.01	0.07	0.11
	RAND - T_4	0.02	0.05	0.10
	RAND - T_5	0.01	0.06	0.10
0.5	3 PC	0	0	0.01
	4 PC	0	0	0
	REG	0	0.03	0.11
	RAND - T_1	0.02	0.08	0.11
	RAND - T_2	0.03	0.04	0.11
	RAND - T_3	0.01	0.07	0.10
	RAND - T_4	0.02	0.06	0.11
	RAND - T_5	0.01	0.06	0.10

Table 6.9: $\pi_B = \pi_A$ =Diet 4: P[Type I Error] for PC, REG, and RAND methods at $n_s = 10$. $T_1 = \sum_{j=1}^{n_{FA}} |\bar{R}_j(\mathbf{Y})|$, $T_2 = \sum_{j=1}^{n_{FA}} |\text{median}_i(R_j(\mathbf{Y}))|$, $T_3 = \sum_{j=1}^{N_{FA}} |\bar{D}_j(\mathbf{Z})|$, $T_4 = \sum_{j=1}^{N_{FA}} |\text{median}_i(D_j(\mathbf{Z}))|$, and $T_5 = \sqrt{\sum_{j=1}^{N_{FA}} \bar{D}_j(\mathbf{Z})^2}$. (Based on $M = 100$ samples of pseudo-seals with $\epsilon = 10\%$ and $n^p = 30$, and $R = 300$.)

at $\alpha = 0.1$ when $\pi_B = \pi_A =$ Diet 1, the test statistics all performed similarly and yielded appropriate P[Type I Error].

In Tables 6.10 and 6.11 the power of the test is given for the various methods. As in Section 6.2.1, π_B was set to be either Diet 1 or Diet 4 and π_A to one of the six previously discussed modified diets. Overall, the RAND method tends to give a larger power then the REG method. This is most noticeable when $\pi_B =$ Diet 4. For example, in Table 6.11, for $\pi_A(4) - \pi_A(6)$, the power for the RAND method (especially for $T_1 = \sum_{j=1}^{n_{FA}} |\bar{R}_j(\mathbf{Y})|$) is large ($> 0.9$) and is small ($< 0.26$) for the REG method.

As in the independent case, at $\alpha \geq 0.05$, the RAND and REG methods appear to be able to detect a difference when a non-zero species is interchanged with a zero species having diet $\pi_k \geq 0.15$. From Figure 6.3, it appears that generally a change in the diet by an absolute difference of 0.20 or a relative difference of approximately 1 is detected by the RAND method with $T_1 = \sum_{j=1}^{n_{FA}} |\bar{R}_j(\mathbf{Y})|$. An exception is when $\pi_A = \pi_A(6)$ as this change was not usually detected.

Overall we recommend the RAND method with test statistic $T_1 = \sum_{j=1}^{n_{FA}} |\bar{R}_j(\mathbf{Y})|$.

6.3.2 Analysis Based on Seal and Prey FA Signatures

As in Subsection 6.2.2, in this subsection we assume that prey bases ($\mathbf{X}_A$ and $\mathbf{X}_B$) or a single prey base ($\mathbf{X}$) are available to estimate the before and after diets of the seals. Since the multivariate randomization P-value (RAND) method appeared, from our simulation study, to be the most appropriate testing procedure we will examine its performance in testing:

$$\begin{aligned} \mathrm{H}_0 &: \pi_B = \pi_A \\ \mathrm{H}_1 &: \pi_B \neq \pi_A. \end{aligned} \tag{6.10}$$

using the diet estimates.

Recall from the previous subsection that test statistics involving relative differences taken on the untransformed data performed sufficiently well. Since these test statistics do not involve log transformations they may be applied to the diet estimates which will likely contain essential zeros. As functions of our diet estimates, these test statistics are as follows:

Species	π_B	π_A					
	Diet 1	1	2	3	4	5	6
Cod	0.30	0.30	0.30	0.318	0.345	0.39	0.363
Haddock	0.30	0	0.30	0.318	0.345	0.39	0.237
Plaice	0	0.30	0	0	0	0	0
Pollock	0.15	0.15	0	0.132	0.105	0.06	0.213
Sandlance	0	0	0.15	0	0	0	0
SilverHake	0.15	0.15	0.15	0.132	0.105	0.06	0.087
WinterFlounder	0	0	0	0	0	0	0
YellowTail	0	0	0	0	0	0	0
ABS(π_1, π_2)		0.600	0.300	0.072	0.180	0.360	0.252
REL(π_1, π_2)		2.00	2.00	0.353	0.861	1.66	1.10
$\alpha = 0.01$	3 PC	0.43	0.07	0.00	0.00	0.04	0.00
	4 PC	0.27	0.00	0.00	0.00	0.00	0.00
	REG	0.77	0.48	0.02	0.25	0.55	0.04
	RAND - T_1	0.98	0.77	0.03	0.25	0.70	0.10
	RAND - T_2	0.98	0.70	0.03	0.20	0.49	0.06
	RAND - T_3	0.97	0.73	0.07	0.4	0.81	0.02
	RAND - T_4	0.93	0.71	0.05	0.36	0.68	0.01
	RAND - T_5	0.98	0.80	0.06	0.33	0.81	0.02
$\alpha = 0.05$	3 PC	0.88	0.41	0.01	0.06	0.27	0.00
	4 PC	0.82	0.22	0.00	0.00	0.12	0.00
	REG	0.96	0.88	0.12	0.50	0.90	0.10
	RAND - T_1	1.00	0.97	0.13	0.52	0.95	0.24
	RAND - T_2	1.00	0.93	0.11	0.39	0.86	0.24
	RAND - T_3	1.00	0.94	0.23	0.79	0.98	0.13
	RAND - T_4	1.00	0.92	0.22	0.75	0.97	0.15
	RAND - T_5	1.00	0.96	0.16	0.74	0.98	0.14
$\alpha = 0.10$	3 PC	0.93	0.60	0.02	0.21	0.58	0.01
	4 PC	0.96	0.47	0.00	0.05	0.38	0.00
	REG	0.99	0.93	0.24	0.68	0.95	0.14
	RAND - T_1	1.00	0.99	0.26	0.74	1.00	0.40
	RAND - T_2	1.00	0.99	0.21	0.58	0.95	0.39
	RAND - T_3	1.00	0.99	0.39	0.93	0.99	0.19
	RAND - T_4	1.00	0.98	0.32	0.86	0.99	0.23
	RAND - T_5	1.00	1.00	0.28	0.86	0.99	0.27

Table 6.10: Power, $\beta(\pi_B, \pi_A)$, for PC, REG and RAND methods with π_B = Diet 1 and various choices of π_A, at $n_s = 10$. $T_1 = \sum_{j=1}^{n_{FA}} |\bar{R}_j(\mathbf{Y})|$, $T_2 = \sum_{j=1}^{n_{FA}} |\text{median}_i(R_j(\mathbf{Y}))|$, $T_3 = \sum_{j=1}^{N_{FA}} |\bar{D}_j(\mathbf{Z})|$, $T_4 = \sum_{j=1}^{N_{FA}} |\text{median}_i(D_j(\mathbf{Z}))|$, and $T_5 = \sqrt{\sum_{j=1}^{N_{FA}} \bar{D}_j(\mathbf{Z})^2}$. (Based on $M = 100$ samples of pseudo-seals with $\delta = 0.5$, $\epsilon = 10\%$ and $n^p = 30$, and $R = 300$.)

Species	π_B	π_A					
	Diet 4	1	2	3	4	5	6
Cod	0.09	0.09	0.09	0.099	0.108	0.117	0.072
Haddock	0	0	0.09	0	0	0	0
Plaice	0.09	0.09	0	0.099	0.108	0.117	0.072
Pollock	0	0.45	0	0	0	0	0
Sandlance	0.45	0	0.45	0.405	0.36	0.315	0.54
SilverHake	0.09	0.09	0.09	0.099	0.108	0.117	0.072
WinterFlounder	0.09	0.09	0.09	0.099	0.108	0.117	0.072
YellowTail	0.09	0.09	0.09	0.099	0.108	0.117	0.072
$\text{ABS}(\pi_1, \pi_2)$		0.900	0.180	0.009	0.180	0.270	0.180
$\text{REL}(\pi_1, \pi_2)$		2.00	2.00	0.555	1.033	1.454	1.167
$\alpha = 0.01$	3 PC	1.00	0.00	0.00	0.00	0.00	0.00
	4 PC	0.89	0.00	0.00	0.00	0.00	0.00
	REG	1.00	0.00	0.01	0.04	0.01	0.03
	RAND - T_1	0.99	0.01	0.00	0.56	0.81	0.52
	RAND - T_2	0.99	0.01	0.01	0.50	0.73	0.53
	RAND - T_3	0.99	0.00	0.01	0.14	0.28	0.15
	RAND - T_4	0.97	0.00	0.01	0.12	0.25	0.15
	RAND - T_5	0.99	0.00	0.00	0.17	0.34	0.20
$\alpha = 0.05$	3 PC	1.00	0.00	0.00	0.01	0.02	0.04
	4 PC	1.00	0.00	0.00	0.00	0.00	0.01
	REG	1.00	0.04	0.06	0.17	0.10	0.07
	RAND - T_1	1.00	0.07	0.08	0.83	0.96	0.81
	RAND - T_2	1.00	0.08	0.07	0.75	0.91	0.74
	RAND - T_3	1.00	0.06	0.04	0.32	0.63	0.37
	RAND - T_4	1.00	0.09	0.04	0.37	0.65	0.40
	RAND - T_5	1.00	0.04	0.04	0.37	0.69	0.42
$\alpha = 0.10$	3 PC	1.00	0.00	0.00	0.06	0.07	0.09
	4 PC	1.00	0.00	0.00	0.02	0.12	0.06
	REG	1.00	0.15	0.14	0.26	0.21	0.15
	RAND - T_1	1.00	0.12	0.14	0.92	0.98	0.94
	RAND - T_2	1.00	0.17	0.12	0.87	0.97	0.89
	RAND - T_3	1.00	0.16	0.08	0.58	0.81	0.58
	RAND - T_4	1.00	0.17	0.10	0.56	0.82	0.54
	RAND - T_5	1.00	0.15	0.11	0.59	0.84	0.59

Table 6.11: Power, $\beta(\pi_B, \pi_A)$, for PC, REG and RAND methods with π_B = Diet 4 and various choices of π_A, at $n_s = 10$. $T_1 = \sum_{j=1}^{n_{FA}} |\bar{R}_j(\mathbf{Y})|$, $T_2 = \sum_{j=1}^{n_{FA}} |\text{median}_i(R_j(\mathbf{Y}))|$, $T_3 = \sum_{j=1}^{N_{FA}} |\bar{D}_j(\mathbf{Z})|$, $T_4 = \sum_{j=1}^{N_{FA}} |\text{median}_i(D_j(\mathbf{Z}))|$, and $T_5 = \sqrt{\sum_{j=1}^{N_{FA}} \bar{D}_j(\mathbf{Z})^2}$. (Based on $M = 100$ samples of pseudo-seals with $\delta = 0.5$, $\epsilon = 10\%$ and $n^p = 30$, and $R = 300$.)

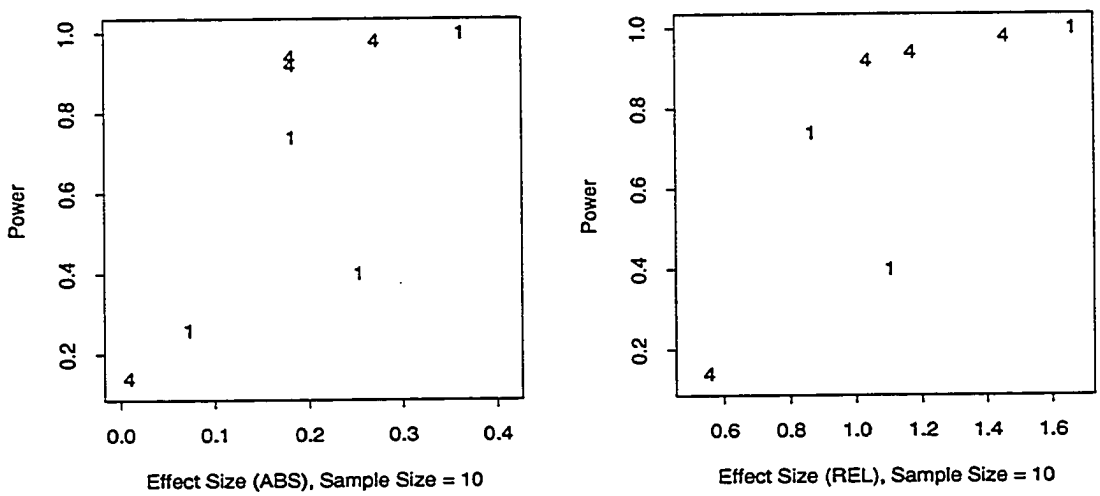

Figure 6.3: Plots of power versus effect size at $\alpha = 0.1$ using $T_1 = \sum_{j=1}^{n_{FA}} |\bar{R}_j(\mathbf{Y})|$ with $n_s = 10$. Effect size is measured as ABS(π_1, π_2) or REL(π_1, π_2). π_1 =Diet 1 is denoted by 1 and π_2 =Diet 4 is denoted by 4.

1. $T_1 = \sum_{k=1}^{I} |\bar{R}_k(\mathbf{p})|$

2. $T_2 = \sum_{k=1}^{I} |\text{median}_i(R_k(\mathbf{p}))|$

To assess the P[Type I Error] and power of the RAND method applied to the diet estimates, we generated paired samples of pseudo-seals as in the previous subsection but now estimated the diet of each of the pseudo-seals in our sample using the full prey base. Results on the P[Type I Error] and power are given in Tables 6.12 - 6.14 and in Figure 6.2. Note that we used $\delta = 0.5$ and the AIT distance measure.

From Table 6.12, the P[Type I Error] associated with this test is usually larger than α. For Diet 1, $T_1 = \sum_{k=1}^{I} |\bar{R}_k(\mathbf{p})|$ gives a smaller P[Type I Error] than $T_2 = \sum_{k=1}^{I} |\text{median}_i(R_k(\mathbf{p}))|$ while for Diet 4, T_2 gives the smaller P[Type I Error]. In practice, to ensure that P[Type I Error] $\leq \alpha$, we should consider rejecting H_0 if the $P-$value $\leq \alpha/2$.

Tables 6.13-6.14 are analogous to the power tables in the previous subsections and the choices of π_B and π_A are identical to the choices of π_1 and π_2 in Subsection 6.2.2 where a test based on independent samples of diet estimates was examined.

Compared to the power calculations when the test was applied to the FA signatures, the power is much lower when the diet estimates are used. Recall that this drop in power also occurred in the independent case and is most probability due to the diet estimates being more variable than the FA signatures. We therefore again examined the power at two sample sizes: $n_s = 10$ and $n_s = 30$.

The test appears to be most powerful when $T_1 = \sum_{k=1}^{I} |\bar{R}_k(\mathbf{p})|$ is used as the test statistic. At $n_s = 10$, the test appears to only detect an interchanging of a non-zero and zero species when the non-zero diet is close to 0.30. At $n_s = 30$, the test almost always detects a difference when a non-zero species with $\pi_k \geq 0.15$ is interchanged with a zero species.

The extent to which a change in the diet of non-zero species can be detected is depicted in Figure 6.4. At $n_s = 10$, an absolute difference of roughly 0.8 and a relative difference of at least 3 is needed to obtain decent power at $\alpha = 0.1$. By $n_s = 30$, the power has greatly improved and it appears that generally an absolute difference of roughly 0.6 and a relative difference of about 2.3 yield high power at $\alpha \geq 0.05$. One exception is when $\pi_B =$ Diet 4 and $\pi_A = \pi(6)$.

6.3.3 Conclusions

The RAND method appears to be an effective method of testing for a difference in paired FA signatures at a sample size of 10. If we reject the null hypothesis that the FA signatures are the same, then, provided the prey from which the seals were eating was the same throughout the experiment, we may also conclude that the before and after diet has changed. If prey bases are available to estimate the diets then we may apply the RAND method to the diet estimates to yield a more appropriate test for a difference in diet. Based on our simulation study, this test must be used cautiously at small sample sizes. When the sample size is small we should probably only reject our null hypothesis of equal diets if the P-value$\leq \frac{\alpha}{2}$ (to ensure that the P[Type I Error] $\approx \alpha$) and in which case our test may not be exceptionally powerful. As in the independent case, using the test on the diet estimates may be much more appropriate with sample sizes closer to 30. Finally, our preferred test statistic is $T_1 = \sum_{k=1}^{I} |\bar{R}_k(\cdot)|$ which is somewhat analagous to the test statistic chosen for the independent case.

			P[Type I Error]		
π_1	π_2	T	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
Diet 1	Diet 1	T_1	0.01	0.07	0.14
		T_2	0.04	0.09	0.19
Diet 4	Diet 4	T_1	0.01	0.11	0.24
		T_2	0.03	0.09	0.17

Table 6.12: P[Type I Error] for RAND method at $n_s = 10$. $T_1 = \sum_{k=1}^{I} |\bar{R}_k(\mathbf{p})|$, $T_2 = \sum_{k=1}^{I} |\text{median}_i(R_k(\mathbf{p}))|$. (Based on $M = 100$ samples of pseudo-seals with $\delta = 0.5$, $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients, and $R = 300$.)

6.4 Real-life Example: Before and After Seal Data

In this section, paired data consisting of before and after seal FA signatures, collected by Margi Cooper (Dalhousie University), is analyzed. In particular, for each of two separate experiments, one conducted in the Fall of 1999 and the other in the Spring of 2000, the FA signatures of seals before and after a feeding experiment were recorded. The data consists of 8 seals for the 1999 data set and 10 seals for the 2000 data sets.

Species	π_B	π_A							
	Diet 1	1	2	3	4	5	6	7	8
Cod	0.30	0.30	0.30	0.345	0.435	0.435	0.45	0.12	0.03
Haddock	0.30	0	0.30	0.345	0.435	0.165	0.09	0.12	0.03
Plaice	0	0.30	0	0	0	0	0	0	0
Pollock	0.15	0.15	0	0.105	0.015	0.285	0.36	0.33	0.42
Sandlance	0	0	0.15	0	0	0	0	0	0
SilverHake	0.15	0.15	0.15	0.105	0.015	0.015	0	0.33	0.42
WinterFlounder	0	0	0	0	0	0	0	0	0
YellowTail	0	0	0	0	0	0	0	0	0
ABS(π_B, π_A)		0.600	0.300	0.180	0.540	0.540	0.720	0.720	1.080
RELdist(π_B, π_A)		2.000	2.000	0.861	2.421	2.134	2.617	2.291	3.086
$\alpha = 0.01$	T_1	0.40	0.22	0.02	0.12	0.09	0.15	0.19	0.49
		1	0.82	0.16	0.73	0.35	0.71	0.77	0.98
	T_2	0.15	0.19	0.03	0.07	0.04	0.11	0.19	0.27
		0.87	0.74	0.29	0.86	0.54	0.82	0.86	0.98
$\alpha = 0.05$	T_1	0.76	0.45	0.13	0.37	0.22	0.32	0.44	0.80
		1.00	0.95	0.43	0.97	0.64	0.92	0.97	1.00
	T_2	0.42	0.41	0.12	0.36	0.13	0.24	0.38	0.57
		1.00	0.86	0.59	0.98	0.79	0.98	0.97	1.00
$\alpha = 0.10$	T_1	0.89	0.57	0.27	0.59	0.31	0.44	0.64	0.92
		1.00	0.97	0.59	0.98	0.76	0.94	0.97	1.00
	T_2	0.64	0.59	0.22	0.54	0.29	0.43	0.51	0.75
		1.00	0.93	0.72	1.00	0.90	0.99	1.00	1.00

Table 6.13: Power, $\beta(\pi_B, \pi_A)$, for RAND methods for π_B = Diet 1 and various choices of π_A, and at two sample sizes: $n_s = 10$ (first row) and $n_s = 30$ (second row) . $T_1 = \sum_{k=1}^{I} |\bar{R}_k(\mathbf{p})|$, $T_2 = \sum_{k=1}^{I} |\text{median}_i(R_k(\mathbf{p}))|$. (Based on $M = 100$ samples of pseudo-seals with $\delta = 0.5$, $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients, and $R = 300$.)

Species	π_B	π_A							
	Diet 4	1	2	3	4	5	6	7	8
Cod	0.09	0.09	0.09	0.153	0.162	0.171	0.045	0.036	0.027
Haddock	0	0	0.09	0	0	0	0	0	0
Plaice	0.09	0.09	0	0.153	0.162	0.171	0.045	0.036	0.027
Pollock	0	0.45	0	0	0	0	0	0	0
Sandlance	0.45	0	0.45	0.135	0.09	0.045	0.675	0.72	0.765
SilverHake	0.09	0.09	0.09	0.153	0.162	0.171	0.045	0.036	0.027
WinterFlounder	0.09	0.09	0.09	0.153	0.162	0.171	0.045	0.036	0.027
YellowTail	0.09	0.09	0.09	0.153	0.162	0.171	0.045	0.036	0.027
ABS(π_B, π_A)		0.900	0.180	0.630	0.720	0.810	0.450	0.540	0.630
REL(π_B, π_A)		2.000	2.000	2.759	3.022	3.268	2.833	3.375	3.912
$\alpha = 0.01$	T_1	0.29	0.06	0.46	0.62	0.70	0.01	0.21	0.21
		0.95	0.34	1.00	1.00	1.00	0.27	0.89	0.90
	T_2	0.12	0.05	0.30	0.46	0.42	0.01	0.13	0.14
		0.67	0.20	1.00	1.00	1.00	0.23	0.80	0.86
$\alpha = 0.05$	T_1	0.50	0.21	0.75	0.86	0.88	0.13	0.51	0.44
		0.99	0.55	1.00	1.00	1.00	0.46	0.99	0.97
	T_2	0.36	0.20	0.69	0.67	0.71	0.11	0.35	0.37
		0.93	0.41	1.00	1.00	1.00	0.46	0.98	0.96
$\alpha = 0.1$	T_1	0.71	0.32	0.86	0.92	0.93	0.24	0.66	0.64
		1.00	0.77	1.00	1.00	1.00	0.65	1.00	0.98
	T_2	0.50	0.31	0.80	0.86	0.82	0.19	0.51	0.53
		0.97	0.49	1.00	1.00	1.00	0.66	1	0.98

Table 6.14: Power, $\beta(\pi_B, \pi_A)$, for RAND methods for π_B = Diet 4 and various choices of π_A, and at two sample sizes: $n_s = 10$ (first row) and $n_s = 30$ (second row). $T_1 = \sum_{k=1}^{I} |\bar{R}_k(\mathbf{p})|$, $T_2 = \sum_{k=1}^{I} |\text{median}_i(R_k(\mathbf{p}))|$. (Based on $M = 100$ samples of pseudo-seals with $\delta = 0.5$, $\epsilon = 10\%$, $n^p = 30$ and calibration coefficients, and $R = 300$.)

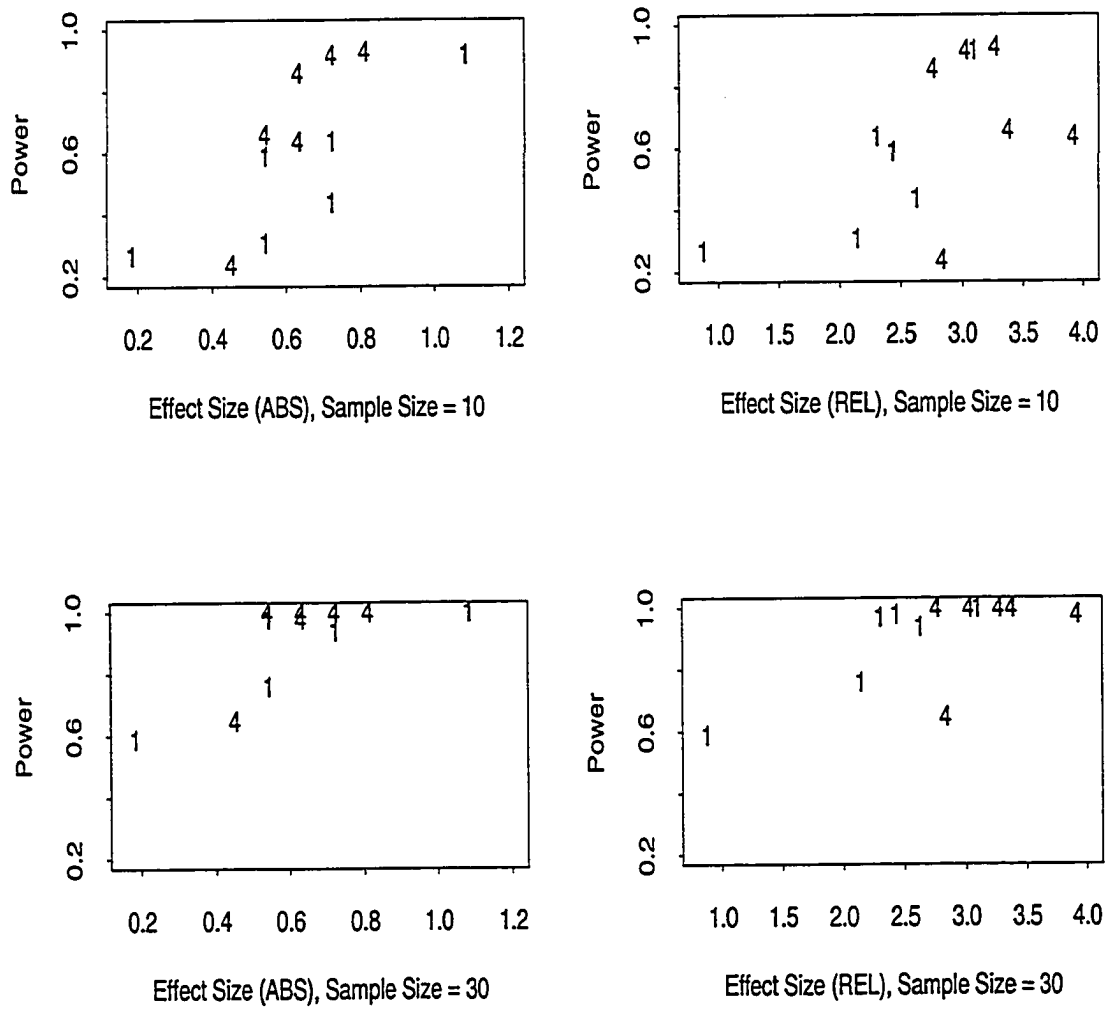

Figure 6.4: Plots of power versus effect size at $\alpha = 0.1$, using $T_1 = \sum_{k=1}^{I} |\bar{R}_k(\mathbf{p})|$, with two sample sizes: $n_{s1} = n_{s2} = 10$ and $n_{s1} = n_{s2} = 30$. Effect size is measured as ABS(π_1, π_2) or REL(π_1, π_2). π_1 =Diet 1 is denoted by 1 and π_2 =Diet 4 is denoted by 4.

The raw FA signatures contained 69 FAs but since only the extended dietary FAs (a subset of 39 FAs) provide information about the diet, these FAs only were used in the analysis. Recall that given only the FA signatures of the seals, we are limited to testing for a change in the FA signatures. For this experiment, of interest is whether or not the FA signatures have changed to reflect the known change in the diet.

Before applying the techniques in Subsection 6.3.1, the zeros in the data set, which we assume to be rounded zeros, had to be adjusted. We proceeded as follows: FAs that had zero entries for all seals (that is, c8.0 and c16.3w1 for the 1999 data and c8.0 for the 2000 data) were removed while FAs that had some zero entries (c22.2w6 for the 1999 data and c16.3w1 and c22.2w6 for the 2000 data) were modified using the multiplicative replacement strategy (MRS) discussed in Section 6.1. (For the 1999 data, $\delta = 1.736 \times 10^{-6}$ while for the 2000 data, $\delta = 1.328 \times 10^{-7}$.)

Based on the results of our simulation study, we chose to use $T_1 = \sum_{j=1}^{n_{FA}} |\bar{R}_j(\mathbf{Y})|$ and the randomization P-value (RAND) method. The P-values are given in Table 6.15.

	1999	2000
n_s	8	10
P-value	0.007	0.003

Table 6.15: P-values obtained using the RAND method (with test statistic $T_1 = \sum_{j=1}^{n_{FA}} |\bar{R}_j(\mathbf{Y})|$) applied to the 1999 and 2000 before and after FA signatures.

For both data sets the P-values are very small. At $\alpha = 0.05$, we may conclude, for both data sets, that the FA signatures of the seals significantly changed during the feeding experiment.

Chapter 7

Conclusions

In this final chapter we first present a summary of the results and recommendations from the previous chapters, and then a discussion of future research. In accordance with the three main topics of this thesis, we divided our summary section into three subsections: Confidence Intervals, Measuring Species Contribution to Seal Variability and Testing for a Difference in Diet.

7.1 Summary

Our research was motivated by the quantitative FA signature analysis (QFASA) method of estimating the diet of predators presented in Iverson *et al* (2004). The method is based on the knowledge that predator FA signatures reflect the signatures of their prey. Given a sample of predator and prey FA signatures from the various potential species in the predators' diet, the proportion of each species in the predators' diet is estimated. In Iverson *et al* (2004), diet estimates were chosen to be the weights that minimized the "distance" between the predator FA signatures and a weighted mixture of the FA signatures of the prey species. We examined the Kulback-Leibler (KL) distance measure (suggested in Iverson *et al* (2004)) as well as another distance measure (AIT) proposed by Aitchison (1992). While a maximum likelihood approach was examined in Section 3.6, it was found to be problematic.

In this thesis we explored in more detail the QFASA model from a statistical perspective and developed QFASA based methods to gain further insight into the diet of predators. These methods are now summarized and we begin with our primary goal of developing confidence intervals for the true diet.

7.1.1 Confidence Intervals

Developing confidence interval (CI) methods for the true diet of a predator or common diet of a group of predators based on QFASA essentially required an examination into ways of parametric modeling the QFASA diet estimates, "parameterizing" the true diet, and estimating the standard error of the diet estimates. As the data, and in particular the diet estimates themselves, encountered in QFASA are compositional, we began in Chapter 2 with a review of the fundamental concepts in compositional data analysis. (This review was based on Aitchison's (1986) book dedicated to compositional data analysis.)

A major challenge in modeling the diet estimates was that they contained essential zeros and, consequently, the models presented in Aitchison (1986) could not be used without some modifications to allow for zero components. Since the dimension of the diet estimate corresponds to the potential number of species in the diet and could be large, we chose to derive marginal distributions for the components, and subsequently individual confidence intervals instead of a confidence region for the true diet. We proposed a mixture distribution obtained by dividing our population of diet estimates into sub-populations according to where the zeros occurred. The nonzero components in each population were modeled by Aitchison's multiplicative logistic normal distribution. Recall that although we derived marginal distributions, these distributions did utilize information from the other components because of how the sub-populations were defined. By making certain assumptions, we saw that this distribution could be simplified which yielded another potential way of modeling the diet estimates. This simpler distribution, however, does not use any information from the other components. We also considered a modification to the simpler distribution in which the multiplicative logistic normal distribution was replaced by the multiplicative logistic skew-normal distribution. When all three distributions were fit to the diet estimates (Chapter 3), we found that the simpler distribution often produced similar fits to those obtained by the more complex mixture distribution but that overall the skew-normal distribution provided the best fit. While our aim was to model the diet estimates, these distributions could potentially be used to model any population of compositional data with a significant number of zero components.

We also examined the distribution of aggregate point estimators of the diet for a sample of independent seal FA signatures such as the sample mean or median. We found that for $n_s \geq 10$, the point estimators were approximately normally distributed. When both the sample size of the predators (n_s) and of the prey (n_k) are large, we proved mathematically that these point estimators are approximately normally distributed. This fact and the adequate fits of the parametric models to the diet estimates allowed us to develop parametric CI methods for the true diet.

Our approach to CI estimation was to derive CIs for certain measures of location (MOLs) of the diet estimates that should be close to the true diet and then to shift our CIs by an estimated amount. In addition to the (population) mean and median of the diet estimates, MOLs that might potentially be better suited for compositional data were discussed in Section 2.4. In Section 3.3 these MOLs were applied to the diet estimates to see if any of the MOLs tended to be closer to the true diet than others. Since a diet estimate could be defined in various ways depending on how the prey are summarized (see Section 3.2), applying the MOLs to the diet estimates involved first defining the population of diet estimates. We concluded that given a FA signature $\mathbf{Y}$ from a seal with true diet π_0, $\mathrm{p}(\mathbf{Y}, \boldsymbol{\mu}_{\mathbf{X}})$ was the most appropriate diet estimate of the true diet and that the population of diet estimates would then be all possible diet estimates that could be obtained from all possible seal FA signatures when the true diet is π_0. (Recall that in practice, we estimate $\boldsymbol{\mu}_{\mathbf{X}}$ by $\bar{\mathbf{X}}$ and therefore have to incorporate this source of variability into our CI methods.) We carried out a small simulation study using pseudo-seals with a known diet to examine the closeness of the MOLs of the simulated population of diet estimates to the true diet. The MOLs were generally close to the true diet and there did not appear to be one best MOL. The median however was an exceptionally good parameterization of the diet when the true diet was zero.

In Chapter 4 we examined various ways of obtaining CI methods for the MOLs and presented a bootstrap based method of estimating the difference or "bias" between the MOLs and the true diet. When the MOL is the mean, the algorithm estimates the bias in the QFASA diet estimates. We divided our interval methods

into four groups: large sample, parametric, semi-parametric and nonparametric intervals. The large sample intervals were simply based on a normal approximation while the more complex parametric and semi-parametric intervals were modifications of CIs discussed in Chapter 2. Recall that in Chapter 2, parametric CIs for MOLs of general compositional data with some zero components were derived based on our proposed parametric mixture models. To apply these intervals to the diet estimates, bootstrap procedures had to be used for two reasons: 1) these intervals involved certain nuisance parameters which would be poorly estimated for small sample sizes and 2) because $\boldsymbol{\mu}_{\mathbf{X}}$ is unknown and the variability due to the prey needs to be incorporated. We also examined two bootstrap based nonparametric methods, one of which was the percentile method (Davison and Hinkley (1997)) and the other of which involved inverting a hypothesis test in which seals are generated under the null hypothesis and bootstrap P-values computed. (Note that the technique of inverting the hypothesis test was also used in the parametric and semi-parametric intervals.)

To compare these methods a simulation study was carried out in which the coverage probabilities and average lengths of the intervals were computed at various sample sizes, for two known diets, and with the AIT and KL distance measure. Because obtaining each diet estimate required an optimization in I dimensions, our CI methods could be very slow and we could not consider every combination of sample size, diet and distance measure that we desired. Instead we carried out a preliminary simulation study and, based on the results, selected a few methods for which to carry out further simulations.

Our overall recommended interval method is the (nonparametric) basic percentile method used with the median point estimator of diet. The basis for this choice was not only its decent coverage probabilities and lengths but also the fact that this method is relatively easy to implement compared to the other methods. Furthermore, although this method can be slow in S-PLUS for larger sample sizes, it is faster than some of the other methods. As the percentile method is not too complex, it would likely be fairly straightforward to convert the S-PLUS code to Fortran to reduce computational time. Additionally, the percentile method incorporates both the variability due to the prey and seals and could easily be extended to include the variability due to fat content.

Recall, however, that we only examined the percentile method for $n_s \leq 25$ due to (computational) time constraints. Although the CI results were good at these sample sizes, for two species (Pollock and SilverHake), the coverage probabilities showed a downward trend. For these species it appeared that the intervals were getting shorter more quickly than the bias estimate was becoming accurate. For Pollock, however, the issue may also be related to the 10% noise that was used. Although the percentile method is preferred, when $n_s \geq 25$, except for Pollock and SilverHake, a normal approximation (without bootstrapping) might suffice.

Note that our CIs suggested that the AIT distance measure was somewhat superior to the KL distance measure but we surmise that this will not always be the case.

We applied the percentile method to real-life captive seabird data and obtained very useful results. The percentile method produced intervals of desirable lengths and appeared to reflect the true diet.

As a final remark, recall that in computing a diet estimate we could summarize the prey in various ways other than by the sample mean prey FA signature. These other methods (discussed in detail in Section 3.2) served two purposes in this thesis: 1) they were used to adjust the variability in the diet estimates obtained using pseudo-seals to approximate the variability expected in real-life and 2) they may be used to incorporate a potential source of variability arising from the fact that seals do not consume the mean prey signature but rather a sample of the prey signatures.

7.1.2 Measuring Species Contribution to Seal Variability

In accordance with the requests of the biologists with whom we have been collaborating, in Chapter 5 we defined a measure of species contribution to the variability in the seal FA signatures. Our statistic was analogous to the coefficient of determination, R^2, used in regression analysis and we called it "PVE" (proportion of variability explained). We defined our "SSE" to be the distance (AIT or KL) from the seal and fitted seal and our "SST" to be the distance from the seal to a fitted seal obtained by randomly assigning prey FA signatures a species label. A desirable property of our PVE statistic is that it is always between zero and one.

We carried out simulations to assess the usefulness of the PVE statistic. We

considered removing species from the prey base and computing the PVE statistic. When species k was removed, we found that generally PVE decreased in accordance with the magnitude of the true diet for species k.

It should be mentioned that although we have not applied our PVE statistic to real-life data in this thesis, the biologists have made use of it.

We also examined a potential "backward elimination" type procedure to reduce the number of possible species in the diet of a predator. Because our approach was bootstrap based it was time consuming to run but initial results (applied to a prey base with 27 species) seemed to indicate that provided a accurate estimates of diets could be obtained with the prey base, the procedure could be used to significantly reduce the number of species in the prey base. If the true proportion of a species in the diet is less than or equal to 0.15, however, the species may be incorrectly dropped. More research into this method is needed.

7.1.3 Testing for a Difference in Diet

Motivated by real-life data on before and after samples of FA signatures of seals, in Chapter 6 we examined methods of testing for a difference in the signatures. We considered tests for both independent and paired samples. For each setting we examined the case where only the seal FA signatures were known and where the prey FA signatures were also known. (In the latter case the tests were carried out on the diet estimates.) Although ultimately it may be the diet of the predators (in our case, seals) that was of interest, if the FA signatures of the predator only were supplied, we found that a significant difference in the signatures may not imply a difference in the underlying diets.

Recall that the major challenge for carrying out such tests was that the number of seals in the sample may be smaller than the dimension of the FA signatures or of the diet estimates, and standard multivariate techniques (modified to deal with compositional data) could not be used. For the case of independent samples, we proposed a multivariate permutation test and for the paired samples a multivariate randomization test. We carried out simulation studies to investigate the P[Type I Error]

and power of the tests. In testing for a difference in FA signatures both tests performed well at sample sizes of 10. When the tests were applied to the diet estimates, it was determined that sample sizes closer to 30 were needed to ensure appropriate P[Type I Error] and power.

In this chapter we also applied our multivariate randomization test to two real-life independent samples of before and after seal FA signatures. Because the sample sizes in the data were much smaller than the dimension of the FA signatures, standard multivariate techniques could not be applied and our test proved to be highly useful.

7.2 Future Research

While QFASA allows accurate point estimates of diet to be obtained in a relatively straightforward manner, we have seen that extending QFASA beyond point estimation presents many statistical challenges. The basis for these challenges include, for example, the data involved being compositional and often containing a fair number of zeros, the potential for small sample sizes compared to the data dimension, variability arising from many sources, the computational burden when several diet estimates need to be computed etc... Some of these issues have been addressed in this thesis; some are still unresolved or require further investigation.

A major issue has been the computational time required to carry out our devised methods in S-PLUS. Although in practice when presumably only a few results would be required, this may not be hugely problematic. We were, however, somewhat limited to what could be investigated through simulations, particularly with respect to our CIs and PVE statistic. Having our code converted to Fortran or C++ might be a useful next step in our research. This would allow, for example, the CIs to be assessed when there is large number of species in the prey base. Additionally, there may exist ways of improving our estimates of bias, such as using a double bootstrap, and we could examine these alternative approaches more efficiently.

Another issue requiring addressing involves incorporating the additional sources of variability into our methods. As previously mentioned, the variability due to fat content is one such source. Another source is the variability due to calibration as we have been treating the calibration factors as known constants.

Throughout our investigations we have found that for data sets with several rounded zeros, our procedures are often largely affected by how we treat these zeros. This suggests that a more thorough examination into the zero issue is needed. Furthermore, it would be very beneficial to have procedures that are robust against outlying FAs.

Thus far we have not considered the PVE statistic and associated backwards elimination procedure in great detail. Accurate modeling of changes in the PVE statistic and the development of a straightforward test for a significant change in the PVE statistic would advantageous. We would also like to examine asymptotic properties of the PVE statistic.

Finally, the biologists with whom we have been working suggested that it would be useful if we could extend our test for a difference in before and after FA signatures to successive periods of time. We could also consider tests for comparing more than two independent populations.

Appendix A

Prey Base

The prey base (see Budge *et al*, 2002) contained FA signatures of prey collected from various areas surrounding the Scotian shelf. Specifically, we selected prey from the following regions: 4Vn, 4Vs, 4VsW, and 4W. (A map of these locations can be found at http://www.nafo.ca.)

Two variations of this prey base were used in this thesis by selecting certain species of interest. The larger of the two was obtained by selecting 27 species corresponding to those in Iverson *et al* (2004). A second prey base, which was referred to as the reduced prey base, contained 8 selected species. The species and their corresponding sample sizes are given in Tables A.1-A.2.

Species	Sample Size	Species	Sample Size
Argentine	25	RedHake	25
Butterfish	33	Redfish	11
Capelin	132	RockCrab	37
Cod	84	Sandlance	96
Gaspereau	48	SeaRaven	18
Haddock	88	Shrimp	96
Halibut	8	SilverHake	43
Herring	157	SmoothSkate	5
Lobster	9	ThornySkate	36
LonghornSculpin	45	WhiteHake	39
Mackerel	34	WinterFlounder	14
OceanPout	11	WinterSkate	15
Plaice	53	Yellowtail	81
Pollock	39		

Figure A.1: Large Prey Base

Species	Sample Size
Cod (COD)	84
Haddock (HAD)	88
Plaice (PLC)	53
Pollock (POL)	39
Sandlance (SAND)	96
SilverHake (SH)	43
WinterFlounder (WF)	14
Yellowtail (YT)	81

Figure A.2: Reduced Prey Base

Appendix B

Pseudo-Seals

Given a prey base with $\mathbf{X}_k$ denoting the prey FA signatures from species k, the following algorithm (analogous to the algorithm in Iverson *et al* (2004)) was used to generate a single pseudo-seal and its QFASA diet estimate:

1. Choose at random one of the 8 vectors of calibration coefficients (used to adjust the FAs as some may always be higher or lower in the predator than in the prey). Let $\mathbf{c}$ denote the chosen vector.

2. Choose the true diet vector π, the amount of noise ϵ, and the number of prey to be sampled n^p.

3. Randomly split $\mathbf{X}_k$ $(k = 1, \ldots, I)$ into a simulation set $\mathbf{X}_k^s$ and a modeling set $\mathbf{X}_k^m$. Note that the splitting assigned 1/3 of the prey signatures to $\mathbf{X}_k^s$ and 2/3 to $\mathbf{X}_k^m$. The splitting process is only carried out if $n_k > 5$.

4. Sample with replacement $n^p \times \pi_k$ times from $\mathbf{X}_k^s$ to obtain $\mathbf{X}_k^*$.

5. Sample with replacement $n^p \times \epsilon$ times from species which are not part of the true diet, π to obtain ϵ^*.

6. Compute
$$\mathbf{Y}^{ps} = \frac{1}{(1+\epsilon)n^p}\mathbf{c}' \odot \left(\sum_{k=1}^{I} \sum_{l=1}^{n^p \times \pi_k} \mathbf{X}_{kl}^* + \sum_{l=1}^{n^p \times \epsilon} \epsilon_l^* \right).$$
The "pseudo-seal" is then $\mathbf{Y}^{ps}$ normalized so that the jth FA of $\mathbf{Y}^{ps}$ is
$$Y_j^{ps} = \frac{Y_j^{ps}}{\sum_{j=1}^{n_{FA}} Y_j^{ps}}.$$

7. Calibrate $\mathbf{Y}^{ps}$ using the mean of the 7 vectors of calibration coefficients not used in Step 1.

8. Select the extended dietary FAs from $\mathbf{Y}^{ps}$ and re-normalize.

9. Select the extended dietary FAs from $\mathbf{X}_k^m$, $k = 1, \ldots, I$ and re-normalize.

10. Summarize $\mathbf{X}_k^m$ (by $\bar{\mathbf{X}}_k^m$, for example), $k = 1, \ldots, I$.

11. Compute $\mathbf{p}(\mathbf{Y}^{ps}, \bar{\mathbf{X}}^m)$.

A sample of n_s pseudo-seals and corresponding diet estimates were computed by repeating Steps 1. - 11. n_s times.

Appendix C

Resampling Techniques

Resampling the Predator FA signatures:

Method 1: Pseudo-Seal Method

1. *for* $r = i, \ldots, n_s$

   (a) Compute the diet estimate for the ith seal $\mathbf{p}(\mathbf{Y}_i, \bar{\mathbf{X}})$.

   (b) Generate a pseudo-seal, $\mathbf{Y}_i^*$, using the algorithm in Appendix B with diet $(1-\epsilon)\mathbf{p}(\mathbf{Y}_i, \bar{\mathbf{X}})$. [1]

Method 2: Nonparametric Bootstrap

1. Sample with replacement from $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_s}$ to obtain $\mathbf{Y}_1*, \ldots, \mathbf{Y}_{n_s}^*$. [1]

Method 3: Parametric Bootstrap

1. Transform $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_s}$ using the additive logratio transformation to obtain $\mathbf{T}_1, \ldots, \mathbf{T}_{n_s}$.
2. Compute $\bar{\mathbf{T}}$ and $\mathbf{S_T}$ where $\bar{\mathbf{T}} = \frac{1}{n_s}\sum_{i=1}^{n_s} \mathbf{T}_i$, and $\mathbf{S_T} = \frac{1}{n_s - 1}\sum_{i=1}^{n_s}(\mathbf{T}_i - \bar{\mathbf{T}})(\mathbf{T}_i - \bar{\mathbf{T}})'$.
3. Sample from a multivariate normal distribution with $\boldsymbol{\mu} \equiv \bar{\mathbf{T}}$ and $\boldsymbol{\Sigma} \equiv \mathbf{S_T}$ to obtain $\mathbf{T}_1^{*r}, \ldots, \mathbf{T}_{n_s}^{*r}$.
4. Transform $\mathbf{T}_1^{*r}, \ldots, \mathbf{T}_{n_s}^{*r}$ to $\mathbf{Y}_1^{*r}, \ldots, \mathbf{Y}_{ns}^{*r}$ using the additive logistic transformation. [1]

[1]Note that if a sample containing more than one calibration vector is supplied, each generated pseudo-seal, $\mathbf{Y}_i^*$, can be assigned a calibration vector with which to be calibrated. In Method 1, the assigned calibration vector could be the mean of the calibration vectors not used in generating the pseudo-seal. In Methods 2 and 3, $\mathbf{Y}_i$ could be assigned a calibration vector chosen by sampling with replacement from the sample of calibration vectors.

Resampling the Prey FA signatures:

Method 1: Nonparametric Bootstrap

For each prey type k,

1. Sample from the rows of $\mathbf{X}_k$ with replacement n_k times to obtain $\mathbf{X}_k^{*r}$.

Method 2: Parametric Bootstrap

For each prey type k,

1. Transform the rows of $\mathbf{X}_k$ using the additive logratio transformation to obtain $\mathbf{Z}_k$.
2. Compute $\bar{\mathbf{Z}}_k$ and $\mathbf{S}_{\text{pool}}(\mathbf{Z}) = \frac{1}{\sum_{k=1}^{I} n_k - I} \sum_{k=1}^{I} (n_k - 1)\mathbf{S}_k$, where $\bar{\mathbf{Z}} = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbf{Z}_i$, and $\mathbf{S}_k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (\mathbf{Z}_i - \bar{\mathbf{Z}})(\mathbf{Z}_i - \bar{\mathbf{Z}})'$.
3. Sample from a multivariate normal distribution with $\boldsymbol{\mu}_k \equiv \bar{\mathbf{Z}}_k$ and $\boldsymbol{\Sigma} \equiv \mathbf{S}_{\text{pool}}(\mathbf{Z})$ n_k times to obtain $\mathbf{Z}_k^{*r}$.
4. Transform $\mathbf{Z}_k^{*r}$ to $\mathbf{X}_k^{*r}$ using the additive logistic transformation.

Bibliography

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman and Hall.

Aitchison, J. (1989). Measures of location of compositional data sets. *Mathematical Geology 21*, 787–790.

Aitchison, J. (2000). Logratio analysis and compositional distance. *Mathematical Geology 32*, 271–275.

Aitchison, J. and Bacon-Shone, J. (1999). Convex linear combinations of compositions. *Biometrika 32*, 354–364.

Aitchison, J. and Ng, K. W. (2003). Compositional hypothesis of subcompositional stability and specific perturbation change and their testing. Compositional Data Analysis Workshop.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *J.R. Stat. Soc., Ser. B 61*(3), 579–602.

Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika 83*(4), 715–726.

Bickel, P. J. and Doksum, K. A. (2001). *Mathematical Statistics*, Volume 1 of *Basic Ideas and Selected Topics.* Prentice Hall.

Budge, S. M., Iverson, S. J., Bowen, W. D., and Ackman, R. G (2002). Among-and within-species variation in fatty acid signatures of marine fish and invertebrates on the Scotian Shelf, Georges Bank and southern Gulf of St. Lawrence. *Canadian Journal of Fisheries and Aquatic Sciences 59*, 886–898.

Casella, G. and Berger, R. (1990). *Statistical Inference.* Duxbury Press.

Casella, G. and Berger, R. (2002). *Statistical Inference.* Duxbury Press.

Chakraborty, B. (2001). On affine equivariant multivariate quantiles. *Annals of the Institute of Statistical Mathematics 53*(2), 380–403.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association 91*(434), 862–872.

Davison, A. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application.* Cambridge University Press.

DiCiccio, T. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science 11*(3), 189–228.

DiCiccio, T. and Romano, J. (1990). Nonparametric confidence-limits by resampling methods and least favorable families. *International Statistical Review 58*(1), 59–76.

Gupta, R. D. and Richards, D. St. P. (2001). The history of the dirichlet and liouville distributions. *International Statistical Review 69*(3), 433–446.

Iverson, S. J., Field, C., Bowen, D. W., and Blanchard, W. (2004). Quantitative fatty acid signature analysis: A new method of estimating predator diets. *Ecological Monographs 72*(2), 211–235.

Ma, Y. and Genton, M. G. (2004). Flexible class of skew-symmetric distributions. *Scandinavian Journal of Statistics 31*(3), 459–468.

Mateu-Figueras, G., Barceló, C., and Pawlowsky-Glahn, V. (1998). Modeling compositional data with multivariate skew-normal distributions. Communication in the IAMG98:Congress of the International Association for Mathematical Geology.

Martín Fernández, J., Barceló-Vidal, C., and Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology 35*(3), 253–278.

Welsh, A. H. (1996). *Aspects of Statistical Inference*. John Wiley and Sons. Inc.