Bootstrap Estimation Of The Fixation Index In Taylor Fish Management Of Western Australian Coastal Waters

By

Thandar Lim[1], Gopalan Nair[2], Neil Sumner[3] and Nihal Yatawara[2]

[1]Health Department of WA, 189, Royal Street, East Perth, WA 6004
[2]School of Mathematics and Statistics, Curtin University of Technology, GPO Box, U1987, Perth, WA 6845
[3]Fisheries WA, PO Box 20, North Beach, WA 6020

Abstract The Nonparametric bootstrap is a particularly versatile tool for data analysis. It's properties have been established by theoretical means, through simulation studies and by application to real data. This paper discusses an application of the bootstrap method for estimating a parameter well known in genetics as the fixation index (F_{ST}). The index provides information as to whether one subpopulation of a species is reproductively isolated from other subpopulations of the same species and is estimated using genotype data. The application discussed in the paper arose as a part of an investigation by the Fisheries Department of Western Australia which was mainly concerned with the sharp decline of the Taylor fish in the coastal waters. Using genotype data collected from six different locations we obtain the distribution of F_{ST}, form confidence intervals and test hypothesis concerning F_{ST} to ascertain whether the Taylor fish stocks found in different areas are genetically different from one another.

1. INTRODUCTION

The Taylor fish is an extremely important species in Western Australia. In the 1980's a sharp decline in the Taylor catches in metropolitan waters was noticed. Consequently the Fisheries Research division undertook a major study to provide definitive information which could determine the extent of the decline. As a part of the study the Department of fisheries collected samples of genotype data for the species at six different locations, Shark Bay, Abrolhos, Geraldton, Perth, Wilson Inlet and Sydney. The data contains information as to the genetic differences existing between sub-populations which will evolve in the course of time. The genetic data was obtained by protein electrophoresis, a commonly used procedure, especially for natural populations.

To estimate the genetic variation of a subpopulation with respect to the total population it is desirable to quantify the amount of differentiation. Consider a large randomly mating population as a basis for comparison and a locus at which there is no selection. If the locus has two segregating alleles, A and A' with the respective frequencies of p and q, the expected proportions of the genotypes AA, AA' and $A'A'$ will be Hardy-Weinberg proportions p^2, $2pq$ and q^2. If the population is subdivided and allele frequencies vary among subpopulations, then the proportion of homozygotes in the total population will be greater than the Hardy-Weinberg expectations(Ryman and Utter(1987)). If p_i is the frequency of A in subpopulation i where i=1,2,...,n and that in each subpopulation genotypes occur in the Hardy-Weinberg proportions given by the local allele frequencies then the proportions of AA, AA' and $A'A'$ in the total population can be written as $\bar{p}^2 + F_{ST}\,\bar{p}\,\bar{q}$, $2\bar{p}\,\bar{q}\,(1 - F_{ST})$ and $F_{ST}\,\bar{p}\,\bar{q} + \bar{q}^2$; where $\bar{p} = \dfrac{\sum p_i}{n}$ is the average frequency of A in the total population, $\bar{q} = 1 - \bar{p}$ and F_{ST} is the fixation index parameter introduced by Wright (1921) to characterize the genotypic distribution at a two-

allele locus. It was shown that $F_{ST} = \frac{V_p}{\bar{p}\bar{q}}$ regardless of whether subpopulations are in Hardy-Weinberg proportion or not where V_p is the variance of p_i over subpopulations (Wright (1921)). A convenient formula for F_{ST} can be obtained by introducing

$$H_s = \frac{\sum H_i}{n}, \qquad (1.1)$$

where, the Hardy-Weinberg expectation of heterozygosity in subpopulation *i* is given by

$$H_i = 1 - (p_i^2 + q_i^2), \qquad (1.2)$$

and $$H_T = 1 - \left(\bar{p}^2 + \bar{q}^2\right). \qquad (1.3)$$

A simple calculation gives

$$F_{ST} = \frac{V_p}{\bar{p}\,\bar{q}} = 1 - \frac{H_s}{H_T}. \qquad (1.4)$$

The F_{ST} defined in the above equation (1.4) is a useful measure of the degree of differentiation among subpopulations at a two-allele locus.

For a locus with more than two alleles, F_{ST} can be defined for each allele by combining the frequencies of all other alleles at the locus. The value of F_{ST} will in general differ among alleles, but at a two-allele locus, F_{ST} will be the same for both alleles. Nei (1973a, 1977) extended the notion of F_{ST} by providing a measure of differentiation called the G_{ST}. It is based on allele frequencies at several multiallelic loci. For a given locus, if p_{ik} is the frequency of allele k in subpopulation *i* then H_I, $\bar{p}_k$ and the H_T are given by,

$$H_i = 1 - \sum_k p_{ik}^2, \qquad (1.5)$$

$$\bar{p}_k = \frac{\sum_i p_{ik}}{n}, \qquad (1.6)$$

and $$H_T = 1 - \sum_k \bar{p}_k^{\,2} \qquad (1.7)$$

where H_T is interpreted as the probability of nonidentity of two-alleles sampled from the total population. Let the average of H_S and H_T with more than one locus be $\bar{H}_T$ and $\bar{H}_S$, then the G_{ST} is defined as,

$$G_{ST} = 1 - \frac{\bar{H}_S}{\bar{H}_T}. \qquad (1.8)$$

For a single locus with two alleles, this becomes the same as in (1.4)(Ryman and Utter, 1987). A connection between the G_{ST} and the F_{ST} can be shown by letting F_{STkl} be the value of F_{ST} for allele k at locus l, and $\bar{p}_{kl}$ be the frequency of this allele averaged over the total population. Then the $\bar{F}_{ST}$ can be defined as the weighted average of F_{STkl} over all alleles, with weights proportional to $\bar{p}_{kl}(1-\bar{p}_{kl})$. This $\bar{F}_{ST}$ is identical to G_{ST} in Equation (1.8) (Wright, 1978),

$$\bar{F}_{ST} = G_{ST} = \frac{\sum_{k,l} \bar{p}_{kl}(1-\bar{p}_{kl})F_{STkl}}{\sum_{k,l} \bar{p}_{kl}(1-\bar{p}_{kl})}. \qquad (1.9)$$

Since the actual distribution of the F_{ST} index in case of Taylor fish is unknown, we will apply a computer based resampling method known as the Bootstrap (Efron 1979) to estimate it. The bootstrap method can also be used to construct confidence intervals and to perform hypothesis tests regarding the F_{ST}. Based on the inference it is possible to make an assesment of the migrating patterns of Taylor fish and this information will be useful to protect the breeding stocks and to recruit adequate number of young fish to sustain the population.

The paper is organised as follows. In section 2, we briefly review the Bootstrap method and the associated statistical inference procedures. In section 3, we discus the application of Bootstrap estimation method for Taylor fish data. Section 4, gives the conclusions of the investigation.

2. STATISTICAL INFERENCE USING THE BOOTSTRAP

2.1 Review of Bootstrap

This section briefly reviews the main features of the bootstrap method. Its relevance to the problem discussed in this paper is quite clear as it allows us to find an approximate distribution for the F_{ST} index which is otherwise unknown. The estimation of a general parameter θ is assumed in the following discussion and the particular

application is discussed in section 3. The basic steps of the bootstrap method are:

(i) The initial construction of an empirical probability distribution (EDF) $\hat{F}(x)$. which is a non-parametric maximum likelihood estimate (MLE) of the population distribution function.

(ii) Drawing simple random samples of size n from the EDF, with replacement. This is a "resample", x_b^*.

(iii) Calculation of the statistic of interest, $\hat{\theta}$, from this resample, yielding $\hat{\theta}_b^*$.

(iv) Repeating the steps (ii) and (iii) B times to estimate the standard error of $\hat{\theta}$.

(v) Construction of a probability distribution from the B, $\hat{\theta}_b^*$ values. This distribution is the bootstrap estimate of the sampling distribution of $\hat{\theta}$, denoted by $\hat{F}^*(\hat{\theta}^*)$, where B is a large number depending on the number of tests to be carried out from data. Typically B should be 50 – 200 (Mooney and Duval (1993), Efron and Tibshirani (1993)).

2.2 Bootstrap Confidence Intervals

The estimated sampling distribution of $\hat{\theta}$ can be used for such inference procedures as estimating the bias of $\hat{\theta}$ and for obtaining confidence intervals around θ. An estimate of the bias of $\hat{\theta}$ is given by

$$\text{Bias}(\hat{\theta}) = \hat{\theta} - \hat{\theta}^*_{(.)} \qquad (2.1)$$

where $\hat{\theta}^*_{(.)} = \dfrac{\sum_{b=1}^{B} \hat{\theta}^*_b}{B}$, with an estimated standard error,

$$se = \sqrt{\frac{\sum_{b=1}^{B} \left(\hat{\theta}^*_b - \hat{\theta}^*_{(.)}\right)^2}{B-1}}. \qquad (2.2)$$

The development of confidence intervals around θ uses information from the estimated sampling distribution of $\hat{\theta}$. An α- level bootstrap confidence interval of θ is defined as those values of $\hat{\theta}$ an analyst feels [(1-α) × 100]% certain will include the true value of θ, given the variability in the sample and the shape of $\hat{\theta}$'s sampling distribution. Some well known methods available for developing bootstrap confidence intervals are the Normal Approximation Method, Percentile Method, Bias-Corrected Percentile (BC) Method and the Percentile-t Method (see, Efron and Tibshirani,(1986)).

2.3 Bootstrap Hypothesis Testing

In nonparametric bootstrap hypothesis testing, resampling can be done in a way that reflects the null hypotheses, even when the true hypothesis is distant from the null. Also the tests should use methods that are already recognized as having good features in the closely related problem of confidence interval construction.

Suppose we are interested in testing the hypothesis that H_0: $\theta = \theta_0$ against the two sided alternative H_1: $\theta \neq \theta_0$. Let $\hat{\theta}$ be a function of the sample $x_1, x_2, \ldots, x_n$. The bootstrap estimate of θ, $\hat{\theta}^*$ can be computed from a resample $x_1^*, x_2^*, \ldots, x_n^*$ drawn from the sample with replacement. Consequently, the bootstrap distribution of $\dfrac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}^*}$ can be used to obtain critical values for the test, where $\hat{\sigma}^*$ is the value of $\hat{\sigma}$ from the resample. For example to test a hypothesis at the 5% level, compute a number κ such that

$$\Pr^*\left(\frac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}^*} > \kappa\right) = 0.05,$$

where $\Pr^*$ denotes the probability measure under the bootstrap distribution.

Then the null hypothesis H_0 will be rejected if

$$\frac{|\hat{\theta} - \theta_0|}{\hat{\sigma}} > \kappa.$$

3. ESTIMATION OF THE INDEX BY BOOTSTRAP METHOD

This section provides the main application of the bootstrap method to Taylor fish data. In particular we will show how to estimate the parameters F_{ST} and G_{ST} from the genotype data.

3.1 Data Description

The genotype data for Taylor fish is obtained from the Fisheries Department of Western Australia which is collected from six different places - Shark Bay, Abrolhos, Geraldton, Perth, Wilson Inlet and Sydney.

For each fish there are genotypes for five different loci (eg AA AB AA AA AB) that are coded for each of the five enzymes. The five different enzymes in this particular data are

ADH-1, AAT-1, LGP-1, VLP-1 and MEP-2. The five different loci for each fish represent the alleles for each of those enzymes. In genetics, each locus has two alleles for diploid animals. A diploid animal (for example, most fish and all humans) has two chromosomes. The gene at a particular locus will have two alleles one from each chromosome. These alleles work together to provide the genetic instructions to produce a particular form of a protein or enzyme.
The AA or AB values represent the A allele on one chromosome and the A or B allele on the other chromosome. There are seven combinations AA, AB, AC, AD, AE, BB and BD in the data. However, AA and AB appear most of the time and the rest of the genotype alleles appear only once or twice. The frequency of these genotypes may be related to the geographical location. The sample sizes at the different locations are as shown in Table 3.1

Table 3.1: Sample size of the data at different Locations.

Location	Locus				
	1	2	3	4	5
Shark Bay	150	150	150	150	150
Abrolhos	94	94	94	94	84
Geraldton	78	78	78	78	76
Perth	84	84	84	84	84
Wilson Inlet	106	106	106	106	104
Sydney	58	58	58	58	58

3.2 The F_{ST} and the G_{ST} Calculation

The bootstrap analysis assumes that the loci are independent. This was confirmed by a Chi-squared test performed on the data.
The calculation of the F_{ST} for allele A in locus1 is shown below. The procedure is repeated to obtain the F_{ST} for alleles B, C, D and E in locus 1 and at all other loci 2, 3, 4 and 5.

Location	No of Allele A
Shark Bay	140
Abrolhos	90
Geraldton	73
Perth	76
Wilson Inlet	97
Sydney	56

To be able to use equations 1.1 to 1.9 the required values of p_i's and q_i's are calculated from the table as follows:

$p_1 = \frac{140}{150} = 0.93333$, $p_2 = \frac{90}{94} = 0.95745$,

$p_3 = \frac{73}{78} = 0.9359$, $p_4 = \frac{76}{84} = 0.90476$,

$p_5 = \frac{97}{106} = 0.91509$, $p_6 = \frac{56}{58} = 0.96552$,

$q_1 = 1-p_1 = 0.06667$, $q_2 = 1-p_2 = 0.04255$,

$q_3 = 1-p_3 = 0.0641$, $q_4 = 1-p_4 = 0.09524$,

$q_5 = 1-p_5 = 0.08491$, $q_6 = 1-p_6 = 0.03448$.

From (1.2) the H_i's can be calculated as follows:

$$H_1 = 1-(p_1^2+q_1^2) = 0.12444,$$
$$H_2 = 1-(p_2^2+q_2^2) = 0.08148,$$
$$H_3 = 1-(p_3^2+q_3^2) = 0.11999,$$
$$H_4 = 1-(p_4^2+q_4^2) = 0.17234,$$
$$H_5 = 1-(p_5^2+q_5^2) = 0.15539,$$
$$H_6 = 1-(p_6^2+q_6^2) = 0.06659.$$

Thus the H_s is given by

$$H_s = \frac{\sum H_i}{n} = \frac{\sum H_i}{6} = 0.120955 \text{ (see, (1.1))}.$$

Now the average values of p and q are given by

$$\bar{p} = \frac{\sum p_i}{n} = \frac{\sum p_i}{6} = 0.935342, \text{ and}$$

$$\bar{q} = 1-\bar{p} = 0.064658.$$

Thus from (1.3) we get,

$H_T = 1-(\bar{p}^2+\bar{q}^2) = 0.120955$ and hence,

from(1.4), $F_{ST} = 1 - \frac{H_S}{H_T} =$ **0.007575.**

The F_{ST} values calculated for each locus and each allele are summarised in table 3.2 below.

Note that the F_{ST} value is set to zero for any particular Locus type which does not contain certain alleles. The G_{ST} calculated from (1.9) yields a value of **0.0197705.**

3.3 Analysis of the Results

Using bootstrap samples, 1000, F_{ST} values for each allele in each loci and 1000, G_{ST} values were computed using a customised C Programme. The entire set of results is not presented here due to space constraints and in the foregoing we only present, a few distributions of F_{ST} and the G_{ST}. The G_{ST} index is a weighted

Table 3.2: The Fst Values using Hardy-Weinberg Expectation.

Locus	Allele				
	A	B	C	D	E
1	0.007575	0.016699	0.007874	0.028902	0.005562
2	0.026087	0.026467	0.005562	-	-
3	0.010186	0.019619	0.012469	0.008977	-
4	0.016256	0.008165	0.01127	-	-
5	0.026193	0.026193	-	-	-

average of F_{ST} values over all alleles in every loci, hence, construction of confidence intervals and hypothesis testing were resorted just for this index.

3.3.1 The Distribution of F_{ST}

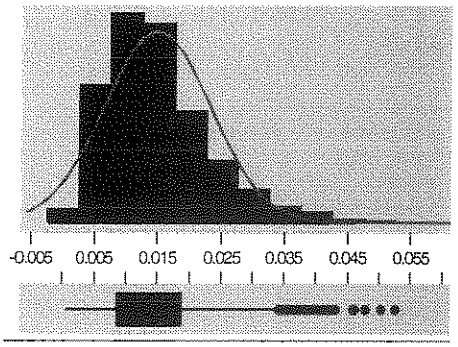

Figure 3.1 The distribution of F_{ST} for Allele A in locus one

The F_{ST} distribution curve for alleles A and B are almost identical and their distributions are similar to a linearly transformed chisquared distribution. If we increase the bootstrap sample size, the curve will be smoother especially for locus three and four. Alleles C, D and E rarely appear in every loci and therefore we could not say much about their distribution curves.

3.3.2 The Distribution of G_{ST}

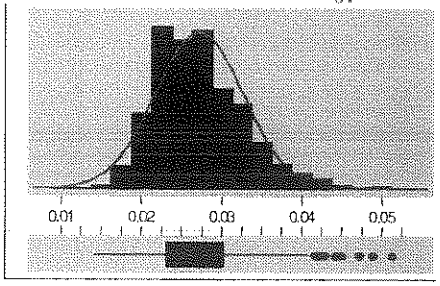

Figure 3.2 The distribution of the G_{ST}

The distribution curve of G_{ST} is similar to a transformed chisquare distribution. The curve will be smoother if the bootstrapping sample size is increased. The mean of the G_{ST} value from Figure 3.2 appears to be different to calculated G_{ST} value of 0.0197705. This is due to the bias in the bootstrapping analysis. The estimation of bias will be carried out in section 3.3.5.

3.3.3 Confidence interval for G_{ST}

A 95% confidence interval for G_{ST} obtained from the percentile method gives ($0.018515 < G_{ST} < 0.039602$), while the Chi-squared approximation gives ($0.018593281 < G_{ST} < 0.039794736$). It is clear that the two are approximately equal suggesting that the Chi-squared approximation is reasonable.

3.3.4 A test of Hypothesis for G_{ST}

A test of hypothesis that the G_{ST} is significantly different from zero is carried out in this section following a method due to Hall and Wilson(1991).To test,

H_0: $G_{ST} = 0$, H_a: $G_{ST} > 0$.

Calculate the test statistic value as,

test statistic = $\left| \hat{G}_{ST} - 0 \right| = 0.0197705$,

where $\hat{G}_{ST}$ is an estimate of the G_{ST} which is obtained using genotype data.

When $\alpha = 0.05$, the critical value of $\left| \hat{G}^{*}_{ST} - \hat{G}_{ST} \right|$ obtained from the percentile method is approximately 0.0176695.Thus,there is evidence in the data to reject H_0, implying that the G_{ST} is significantly different from zero.

3.3.5 Bias of G_{ST}

Equation (2.1) gives the estimated Bias of G_{ST} as , $[\text{Bias}(\hat{G}_{ST})] = \hat{G}_{ST} - \hat{G}^{*}_{ST(.)}$,

where $\hat{G}^{*}_{ST(.)} = \dfrac{\sum_{b=1}^{B} \hat{G}^{*}_{STb}}{B}$

$= \dfrac{27.145}{1000} = 0.027145.$

i.e, estimate $[\text{Bias}(\hat{G}_{ST})] = -0.0073745$

The standard error of the bias using equation (2.2) is given by,

$$\sqrt{\frac{\sum_{b=1}^{1000} (\hat{G}^{*}_{STb} - 0.027145)^2}{999}} = 0.0054915$$

The ratio of the estimated bias to the standard error is $\dfrac{0.0073745}{0.0054915} \approx 1.34$ which is fairly large.

4. Conclusions

In this paper we have described how the bootstrap method can be used to estimate the fixation index parameter and perform associated inferences for the Taylor fish data. The tests seem to indicate that the G_{ST} parameter is significantly different from zero. Accordingly the following conclusions can be drawn:

It appears that the Taylor fish stocks are genetically different which indicates that they do not travel significantly. They just stay in their own area or only a small number is travelling to other areas. These movements are not significant. When the fish stocks are different, if one stock is depleted due to over fishing it will not be replenished by other stocks. This is an important finding for the Fisheries managers responsible for conserving fish stocks.

In the analysis we have taken only 1000 resamples, which is the minimum need according to Efron and Tibshirani(1986). However, we expect that the estimated G_{ST} will be similar for larger number of resamples. This and the removal of bias needs further investigation.

5. References

Efron, B., and R.J. Tibshirani, *An introduction to Bootstrap*, Chapman and Hall, New York,1993.

Efron, B., Bootstrap methods: Another look at the Jackknife,*Annals of Statistics*,(7),1-26,1971.

Efron, B., and R.J. Tibshirani, Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy, *Statistical Science*,(1), 54-77,1986.

Hall, P., and S.R. Wilson, The consultant's forum: Two Guidelines for bootstrap hypothesis testing, *Biometrics*, (47), 757-762, 1991.

Mooney, C.Z., and R.D. Duval, *Bootstrapping a nonparametric approach to statistical inference*, Sage publication, Newbury Park, New York, 93.

Nei, M., Analysis of gene diversity in subdivided populations, *Proceedings of the national academy of sciences, U.S.A.*, (70),3321-3323,1973a.

Nei, M., F-Statistics and analysis of gene diversity in subdivided populations, *Annals of Human Genetics* (41), 225-233,1977.

Ryman, N., and F. Utter, *Population Genetics and Fishing Management*, University of Washington Press, Seattle, 1987

Wright, S., Systems mating,*Genetics*(28), 111-178,1921.

Wright, S., Variability within and among natural populations, *Evolution and the Genetics of Populations* (4), University of Chicago Press, Chicago, (1978)