Parameter Estimability in the Dennis-Kemp Phenology Model

A Thesis
Presented in Partial Fulfillment of the Requirements for the
Degree of Master of Science
with a
Major in Statistical Science
in the
College of Graduate Studies
University of Idaho
by
Michelle A. Londe

Major Professor: Brian Dennis, Ph.D.
Committee Members: Timothy Johnson, Ph.D.; William Price Ph.D.
Department Administrator: Chris Williams, Ph.D.

May 2018

Authorization to Submit Thesis

This thesis of Michelle A. Londe, submitted for the degree of Master of Science with a Major in Statistical Science and titled "Parameter Estimability in the Dennis-Kemp Phenology Model," has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: _____ Date: _____
Brian Dennis, Ph.D

Committee Members: _____ Date: _____
Timothy Johnson, Ph.D

_____ Date: _____
William Price, Ph.D

Department Administrator: _____ Date: _____
Chris Williams, Ph.D

Abstract

Dennis, Kemp, and Beckwith (1986) developed a model describing the stochastic, temperature-dependent development of insect life stages in the field. The model predicts the proportion of insects in a population that are in a given life stage at a given time. The model is based on a logistic probability distribution where maximum likelihood (ML) is used to compute parameter estimates. This paper explores the estimability of parameters within this model. Parametric bootstrap confidence intervals for parameters and functions of the parameters were studied for efficacy. Limitations due to low sample sizes were also studied. The ML parameter and function of the parameters bootstrap sampling distributions indicate can be described well with a large sample multivariate normal distribution. Confidence intervals also followed the prescribed coverage rate and are adequate with this model. Lower sample sizes had lower bootstrap confidence interval coverage rates, but converged quickly to 95% as sample size increased. Results from this paper should be of interest to insect-pest managers and researchers who model development of insects, plants, and animals. Examples of these techniques are presented using data from the western spruce budworm, *Choristoneura occidentalis*, rangeland grasshopper species *M. sanguinipes* and the almond tree *P. dulcis*.

Acknowledgements

I would first like to thank my advisor Dr. Brian Dennis. The door was always open whenever I had a question about my research, writing, or when I had moments of self-doubt. This work would not have been possible without his patience, guidance, and knowledge. I would like to extend my thanks to my committee members Dr. Tim Johnson and Dr. William Price for their support and advice throughout this process. I would also like to thank Dr. William Kemp withthe USDA Agricultural Research Services for his collaboration.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Introduction

The ability to forecast the occurrence of development stages of organisms is important for integrated pest management and the planning of agricultural events. For example, economic losses in agriculture or forestry are often attributable to specific development stages of given insect populations (Dennis et al 1986). Additionally, the pest organisms may respond differently to control measures applied to different development stages. Hence, understanding and predicting growth stage occurances is important.

Osawa et al. (1983) first presented a phenology model describing the development of balsam fir, *Abies balamea.* His model was unique in that it incorporated an environmental factor, temperature, into a model for development stages through time. The model describes the stochastic, temperature-dependent development of balsam fir buds stages in the field by predicting the proportion of fir buds in a population that are in a given life stage at specified times. Later, a simpler version of this phenology model was developed by Dennis, Kemp, and Beckwith in 1986 and applied to the western spruce budworm *Choristoneura occidentalis*. Other useful tools were developed related to this model that are of interest to practitioners such as: estimating the time to reach the maximum proportion of insects in a given growth stage, or finding the time at which a predetermined proportion of insects have reached a specified growth stage or beyond (Dennis and Kemp 1988). While other phenology models have been develped, including the Bayesian model introduced in Knape and De Valpine (2016) with shrimp brine data, they can be complex. The Dennis-Kemp model, however, provides a simpler alternative.

The D-K phenology model has been used in a variety of applications, including western spruce budworms (Kemp et al. 1986), rangeland grasshoppers (Kemp and Osanger 1986), glassy winged sharpshooter bugs (Castle et al. 2005), white spruce (Volney and Cerezke 1991), and snail passage in the gut of rainbow trout (Bruce et al. 2009).

The data required for the Dennis-Kemp model are an example of specialized count data that often occur in ecological problems. The phenology data set is represented in a $q \times r$ table for counts of the number of organisms in stage j $(j = 1, 2, \dots, r)$ recorded in samples taken at cumulative degree day times t_i $(i = 1, 2, \dots, q)$. Frequently, the

individuals in each row (i.e. each sampling time) are aggregated to within only a few of the possible growth stages. The table of counts typically contains many zeros, sometimes occurring in more than half the table cells, even when the numbers of individuals in each sample is quite large. As a result of the data sparsity, there is potential for the breakdown of traditional statistical procedures for parameter estimation and subsequent model evaluation. Among various parameter estimability problems that could arise are: (1) multiple local maxima in the likelihood function, leading to sensitivity of model-fitting calculations to initial parameter values, (2) ridge-shaped likelihood function, in which two or more parameters are difficult to estimate separately, (3) failure of the large-sample normal approximations for maximum likelihood (ML) estimates, such as described in Dennis and Kemp 1988, resulting in biased estimates and confidence intervals with poor coverage properties.

While the Dennis-Kemp model has been used in a variety of applications, the estimability of the parameters and functions of the parameters has not been closely studied. Additionally, the statistical properties of the estimates and the quality of the statistical inference procedures described in Dennis and Kemp (1988) have not been evaluated. The objective of this research is to evaluate the quality of the asymptotic confidence interval for parameters and functions of the parameters. The adequate sample size needed for satisfactory estimability was also evaluated.

Chapter 2 Model and Methods

2.1 Model Description

The Dennis-Kemp phenology model describes the temperature-dependent, stochastic development of organism life stages. While development is a continuous process where the organism accumulates development through time, the growth stages are recorded as visible categorizations made by observers. For modeling purposes, it is assumed that a given insect can be categorized into one of *r* mutually exclusive and sequential development stages.

Because the model represents a stochastic process, the probability distribution for the amount of development an organism has accumulated at time *t* changes over time. Let $X(t)$ be the amount of development accumulated at time *t*. In a previous phenology model, proposed by Osawa et al (1983), the $X(t)$ were assumed to follow a normal distribution. The normal distribution and the integral must be evaluated through numerical integration or other methods (Dennis et al 1986). An alternative distribution is based on the logistic curve with probability density function (PDF):

$$f(x,t) = \exp\left(\frac{x-t}{\sqrt{vt}}\right) \Big/ \left\{\sqrt{vt}\left[1+\exp\left(\frac{x-t}{\sqrt{vt}}\right)\right]^2\right\} \quad [1]$$

where v is a positive constant and a function of time. This distribution has a mean of *t* and variance $(\pi^2/3)vt$. Dennis et al 1986 modifed this distribution such that the mean and variance of the distribution increase linearly with *t.* The logistic cumulative distribution function (CDF) is given by:

$$\Pr[X(t) \le x] = \int_{-\infty}^{x} f(u,t)du = 1 \Big/ \left\{1+\exp\left[-\left(\frac{x-t}{\sqrt{vt}}\right)\right]\right\}. \quad [2]$$

This form of the logistic curve is easier to compute and analyze than the normal alternative of Osawa. Additionally, the logistic distribution has heavier tail probabilities than the normal, and thus is able to account for extra variability often present in biological data. Previous research has shown that the logistic distribution provides almost identical results to the normal probability function (Dennis et al 1986).

Let $X(t)$ be be a random variable representing the continuous level of development level for a randomly selected organism at time *t*, where *t* is measured in thermal time such as degree days (cumulative DD). Typically, the exact $X(t)$ value is unobservable. The

process, however, can be categorized into observable units. If the stochastic process begins with the insect in a given life stage at $t = 0$, as *t* progresses, the organism will sequentially pass through different life stages. Let $a_j, j = 1, \ldots, r-1$ be the amount of development in DD necessary for an organism to attain the $(j+1)^{st}$ stage. That is, the a_j values are the DD values separating the *r* stages:

Stage 1: $X(t) \leq a_1$

Stage 2: $a_1 < X(t) \leq a_2$

⋮

Stage $r-1$: $a_{r-2} < X(t) \leq a_{r-1}$

Stage r: $a_{r-1} < X(t)$.

Let random variable $Y(t)$ be the categorized growth stage of a randomly sampled member of the population at time *t*, where the possible values for $Y(t)$ are the discrete values {1, 2, …, *r*}. Given the correspondence between $X(t)$ and $Y(t)$, if $X(t)$ has not yet attained a_j, then $Y(t)$ has not yet advanced beyond stage *j*. From Eq. 2, the probability that $X(t)$ has not yet attained signpost a_j is the area of the logistic density curve below a_j at time *t*:

$$\Pr\left[X(t) \leq a_j\right] = \Pr[Y(t) \leq j] = \begin{cases} 0, j = 0\ (a_0 \equiv -\infty); \\ \left\{1 + \exp\left[-\left(a_j - t\right)/\sqrt{vt}\right]\right\}^{-1},\ j = 1,\ \ldots,\ r-1 \\ 1,\ j = r\ (a_r \equiv +\infty) \end{cases} \quad [3]$$

The event that $a_{j-1} < X(t) < a_j$ is equivalent to the event that the organism is in stage *j*, or $Y(t) = j$. For completeness, a_0 and a_r are respectively defined to be $-\infty$ and $+\infty$ so that the sum of the expected probabilities is equal to 1.

The expected probability that the organism is in stage *j* at time *t*, $\Pr(Y(t) = j) = p_j(t)$, is the area between a_{j-1} and a_j and is obtained from Eq. 3 as $\Pr[Y(t) \leq j] - \Pr[Y(t) \leq j-1]$. Following this, and accounting for end points, the model of Dennis et al (1986) takes $p_j(t)$ to be given by:

$$p_j(t) = \begin{cases} \left\{1 + \exp\left[-\left(a_j - t\right)/\sqrt{vt}\right]\right\}^{-1}, j = 1 \\ \left\{1 + \exp\left[-\left(a_j - t\right)/\sqrt{vt}\right]\right\}^{-1} - \left\{1 + \exp\left[-\left(a_{j-1} - t\right)/\sqrt{vt}\right]\right\}^{-1}, j = 2, \ldots, r-1 \\ 1 - \left\{1 + \exp\left[-\left(a_{r-1} - t\right)/\sqrt{vt}\right]\right\}^{-1}, j = r \end{cases} \quad [4]$$

The quantity a_j can be interpreted as the time *t* at which half the population is in stage *j* or below: $\Pr[Y(a_j) \leq j] = \Pr[Y(a_j) > j] = 1/2$. As the organisms progress through the different stages, there will be variability in the rates of progress, and thus variability in the amounts of progress through time. The parameter, *v*, measures this variability. The quantities $a_1, a_2, \ldots, a_{r-1}$ and v are unknown parameter values estimated from the data and can be written as the column vector, **θ**:

$\boldsymbol{\theta} = [a_1, a_2, \ldots, a_{r-1}, v]'$.

2.2 Parameter Estimation

The data for the model consist of a sample of organisms at fixed times t_1, t_2, …, t_q. At each time t_i, the number of organisms in stage *j* is recorded. The count in stage *j* at time *i* is denoted as y_{ij} assumed to be random variates from a multinomial distribution with $j = 1$ to r categories. The sample size at each time is $n_i = \sum_j y_{ij}$. The joint distribution of the random variables Y_{i1}, Y_{i2}, …, Y_{ir} is multinomial with sample size n_i and respective probabilities $p_1(t_i)$, $p_2(t_i)$, …, $p_r(t_i)$, where the probabilities are a result of Eq. 4, i.e. $[Y_{i1}, Y_{i2}, \ldots, Y_{ir}] \sim \text{multinomial}(n_i, p_1(t_i), p_2(t_i), \ldots, p_r(t_i))$.

Assuming independent sampling, the likelihood function is a product of the respective multinomial probabilities evaluated at the count values from the samples:

$$L(a_1, a_2, \ldots, a_{r-1}, v) = \prod_{i=1}^{q} \frac{n_i!}{y_{i1}! y_{i2}! \ldots y_{ir}!} [p_1(t_i)]^{y_{i1}} [p_2(t_i)]^{y_{i2}} \ldots [p_r(t_i)]^{y_{ir}}. \quad [5]$$

Log-likelihood is used more often for ease of calculation in parameter estimation. Then, the log-likelihood is:

$$\log L(a_1, a_2, \ldots, a_{r-1}, v) = \log C + \sum_{i=1}^{q} \sum_{j=1}^{r} y_{ij} \log p_j(t_i) \quad [6]$$

where $C = \prod n_i! / [(y_{i1}!)(y_{i2}!) \ldots (y_{i2}!)]$ is a combinatorial constant that does not contain parameter values. The maximum likelihood (ML) estimates of the unknown parameters are the values of a_1, a_2, …, a_{r-1}, and v which jointly maximize the log-likelihood function of the data. The column vector of ML estimates are denoted as: $\hat{\theta} = [\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_{r-1}, \hat{v}]'$.

The function can be maximized with numerical optimization, including the Nelder-Mead or Broyden-Fletcher-Goldfar-Shanno algorithms, within the optim() function in R (Venables and Ripley 2002) or the fminsearch() function in MATLAB (Lagarias et al.

1998). Alternatively, Jenrich and Moore (1975) showed that maximizing *L* or log *L* provides equivalent results to performing an iteratively reweighted least squares regression. In a nonlinear regression solution, the y_{ij} values would be used as observations on the dependent variable as a function of $n_i p_{ij}$ values, with weights $\left(n_j p_{ij}\right)^{-1}$ computed at each iteration.

For large samples (Dennis and Kemp 1988), ML estimates have many desirable statistical properties, among them a large-sample multivariate normal distribution:

$$\widehat{\boldsymbol{\theta}} \xrightarrow{d} \text{multivariate normal}(\boldsymbol{\theta},\ \sum(\boldsymbol{\theta})), \qquad [8]$$

where $\xrightarrow{d}$ denotes convergence in distribution as sample size becomes large, and $\sum(\boldsymbol{\theta})$ is an $r \times r$ variance-covariance matrix (the inverse of a matrix, the "information matrix," obtained as expected values of the matrix of negative second derivatives of the log-likelihood function) with elements that are functions of θ (Bishop et al. 1975).

2.3 Picking Initial Values

Initial values must be chosen in order to use the numerical optimization algorithms to calculate the parameter estimates. Careful selection of initial values will allow for the best results from the optimization and will help avoid false solutions from of local maxima in the likelihood function (Irvine 2011). The parameters have biological interpretations that aid in selecting the initial values. From Eq. 4, the parameter a_j is the value of *t* at which $Pr[Y(t) \leq j] = 1/2$, or the time at which half the population are in stage *j* or below. This time for stage *j* can be approximated by calculating the cumulative row proportions of the data matrix ($q \times r$ matrix of counts) then finding the row (denoted as row k) where the cumulative proportions for stage *j* changes from less than $1/2$ to greater than $1/2$. An initial value for a_j can then be taken as t_k.

Once all a_j are chosen, then the initial value of v, v_i, can be estimated from one of the samples as the sample variance of the frequency distribution of development times in that *i*th sample as:

$$v_i = \left(\frac{3}{\pi^2 t_i}\right) \sum_{j=1}^{r} y_{ij} \left(m_j - \bar{m}_i\right)^2 / (n_i) \ . \qquad [9]$$

Here, $m_1 = (a_j - a_{j-1})/2$ is the midpoint of the interval (a_{j-1}, a_j), with $m_j = a_1 - (a_2 - a_1)/2$ and $m_r = a_{r-1} + (a_{r-1} - a_{r-2})/2$, and $\bar{m}_i = \sum_{j=1}^{r} y_{ij}\, m_j / n_i$ being the weighted mean of the m_j values using the counts from sample *i*. The sample *i* chosen to calculate the initial *v* value should be one where the development stages are spread out. Another strategy is to calculate an initial *v* value for every row of data using Eq. 8 and use the average of the values as the initial value for numerical optimization.

2.4 Functions of the parameters

2.4.1 Peak Time

Peak time at each stage during each state occurs at the maximum of $p_j(t)$ for *j*=2, 3, …, *r*-1. The first and last stages are monotone functions of *t* that do not peak. Numerical maximization can be used on $p_j(t)$ to find peak time from Eq. 4.

2.4.2 Time at which 100ξ% of plants have attained stage j

Another function of the model parameters is the time in which a chosen proportion, say ξ, of organisms have attained the *j*th stage or beyond. The time τ at which 100ξ% of the organisms have attained the *j*th stage or beyond is the solution to

$$\xi = 1 - \left\{1 + \exp\left[-\left(a_{j-1} - \tau\right)/\sqrt{v\tau}\right]\right\}^{-1}. \qquad [10]$$

This equation can be solved for τ (Kemp and Dennis 1991):

$$\tau = a_{j-1} + \frac{v}{2}\left[\log\left(\frac{1-\xi}{\xi}\right)\right]^2 - \frac{1}{2}\log\left(\frac{1-\xi}{\xi}\right)\sqrt{v\left\{4a_{j-1} + v\left[\log\left(\frac{1-\xi}{\xi}\right)\right]^2\right\}}, \ \xi < 0.5 \qquad [11]$$

$$\tau = a_{j-1} + \frac{v}{2}\left[\log\left(\frac{1-\xi}{\xi}\right)\right]^2 + \frac{1}{2}\log\left(\frac{1-\xi}{\xi}\right)\sqrt{v\left\{4a_{j-1} + v\left[\log\left(\frac{1-\xi}{\xi}\right)\right]^2\right\}}, \ \xi > 0.5 \qquad [12]$$

When ξ=0.5, τ = a_{j-1} by definition of a_{j-1}. If one is interested in the last stage, then the time τ 100ξ% of the organisms have reached last stage is given by Eq. 11 or Eq. 12 with a_{j-1} replaced by a_{r-1}.

2.5 Simulation Methods

The bootstrap method is a computationally intensive method to assess the accuracy of statistical estimates through resampling the sample data. It is a procedure that uses the

data to estimate the model that gave rise to the data, followed by using the estimated model to simulate the sampling distribution of the parameter estimates. The method is useful for determining the shape, bias, and spread of the sampling distribution. Bootstrapping has two approaches: parametric and nonparametric sampling. The parametric bootstrap generates data from a parametric model $\hat{F}$ estimated from the sample data. Under the nonparametric bootstrap, data is not generated from a parametric model; it instead resamples from a nonparametric model estimate such as the empirical distribution function.

Bootstrapping also is a useful tool for constructing confidence intervals (Efron and Tibshirani 1993) and an alternative to complex or intractable analytical solutions. Bootstrap confidence intervals tend to have coverage rates close to the claimed $100(1-\alpha)\%$ rate (Pawitan 2001). The percentile interval is an effective technique to construct the bootstrap confidence interval. Let $\hat{G}_i$ be the cumulative distribution function of $\hat{\theta}_i$. Then the $1-\alpha$ percentile interval is defined by the $\alpha/2$ and $1-\alpha/2$ percentiles from $\hat{G}_i$:

$$\left[\hat{\theta}_{lo}, \hat{\theta}_{up}\right] = \left[\hat{G}_i^{-1}(\alpha/2), \hat{G}_i^{-1}(1-\alpha/2)\right] \quad [13]$$

which, by definition, can be rewritten as

$$\left[\hat{\theta}_{lo}, \hat{\theta}_{up}\right] = \left[\hat{\theta}_i^{(\alpha/2)}, \hat{\theta}_i^{(1-\alpha/2)}\right]. \quad [14]$$

If the distribution of $\hat{\theta}_i$ is approximately normal, then the percentile intervals and the confidence intervals derived from the standard normal distribution will produce similar results (Efron and Tishirani 1993).

A Monte Carlo simulation was used to evaluate the coverage rate of the bootstrap confidence interval. Monte Carlo simulations utilize simulated data, rather than empirical data to investigate the behavior of a statistic. The main principle behind the simulation is that the behavior and distribution of the statistic can be studied by the empirical process of drawing random samples (Mooney 1997). The basic Monte Carlo procedure, described by Mooney 1997, is to (1) specifiy ‘psuedo-population’ to generate random samples of data, (2) sample data from the pseudo-population, (3) calculate the statistics from the data computed in Step 2, (4) repeat Step 2 and 3 for a given number of trials, and (5) construct the Monte Carlo estimate of the sampling distributions for the statistics calculated in Step 3.

The Monte Carlo procedure can provide insight into the statistic's sampling distribution and assess the quality of inferential methods. The procedure can evaluate the quality of the bootstrap confidence interval and its coverage probability for parameters and functions of the parameters. The observed claimed rate should converge to the true $1-\alpha$ coverage probability if the objective is a $100(1-\alpha)\%$ confidence interval. Comparison of the claimed rate to the observed rate provides a quantitative validation of the bootstrap method to compute the confidence intervals (Van den Boogaard and Hall 2004).

A process to evaluate the coverage probability of the parametric bootstrap confidence interval with the phenology data and methods is as follows:

1. Select a set of 'true' parameter values for the parametric model to serve as the reference. Here, the ML estimates of $\boldsymbol{\theta} = [\, a_1,\ a_2, \ldots, a_{r-1}, v]'$, found from applying the D-K model to an observed data set, and the empirical sample times will serve as the references. The later constructed confidence intervals will indicate if the parameter reference values are contained within the interval or not. Additionally, multiple sample sizes $n_1, n_2, \ldots, n_q$ can be selected for evaluation.
2. Simulate $B_1 = 1000$ product-multinomial data sets assuming the parameters estimates θ, written as $\boldsymbol{x}^* = (\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots, \boldsymbol{x}_{1000}^*)$ from the reference D-K model. Fit the D-K model to each of the simulated product-multinomial data sets using ML estimation, resulting in 1000 sets of ML parameter estimates $\widehat{\boldsymbol{\theta}}_i = [\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_{r-1}, \hat{v}]'$. In this estimation series, the original sample sizes and sampling times should be used.
3. Simulate $B_2 = 1000$ product-multinomial data sets, written as $\boldsymbol{z}_i^* = (\boldsymbol{z}_1^*, \boldsymbol{z}_2^*, \ldots, \boldsymbol{z}_{1000}^*)$ from each set of the parameter estimates $\widehat{\boldsymbol{\theta}}_i = [\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_{r-1}, \hat{v}]'$, using reference sample sizes and times. Re-fit the D-K model to each bootstrap data set in $\boldsymbol{z}_i^*$, obtaining 1000 bootstrap ML estimates for each $\widehat{\boldsymbol{\theta}}_i$ parameter vector. Construct a $100(1-\alpha)\%$ bootstrap confidence interval from the 1000 bootstrap ML estimates for each parameter. If evaluating the bootstrap confidence interval for functions of the parameters, then the peak times and τ should be evaluated for each $\boldsymbol{z}_i^*$.

4. The above process has yielded 1000 confidence intervals for each $\widehat{\boldsymbol{\theta}}_i$. Obtain the proportion of the confidence intervals for each parameter that contains the 'true' reference parameter value. The proportion obtained is the estimated coverage rate for the bootstrap confidence interval.

2.6 Sample size needed for adequate results

Estimability and the bootstrap confidence interval were also evaluated for varying sample sizes. The sample sizes were gradually increased until the coverage rate was approximately 95% for all parameter confidence intervals. The empirical $p_j(t)$ were used to construct the table of counts at each t_i with the varying sample sizes. Sample sizes evaluated included $n_i = 10, 20, \dots, 100$. For each sampling time, n_i was kept the same.

2.7 Data

Data from three different sources were used in this paper as sources of reference values for parameters. Each of the data sets are product-multinomial count datasets, explained in a previous section, with samples taken at several DD times. The data are phenology count data. The number of organisms in each stage $j = (1, 2, \dots, r)$ are collected at each time t_i. However, the data sets contain mostly zeros at each time point. Each data set is a sparse table involving a different species with varying number of stages and sampling times t_i. This will allow us to study parameter estimability and assess the model in a variety of settings and applications encountered in the literature. The following count data sets can be found in Appendix 1.

The first dataset used is the western spruce budworm, *Choristoneura occidentalis*, is described in Kemp et al. (1986). This represents the development of a western spruce budworm population. The seven development stages are five instars (instar II-instar VI) pupa, and adult. Samples were drawn at 12 different DD times.

The second dataset is described in Kemp and Osanger (1986) and represents the development of rangeland grasshoppers in Montana, specifically the *Melanoplus sanguinipes*. The stages of interest were five instars (instar I-instar V) and adult stage. Samples were drawn at 15 DD times.

Finally, phenology data for the almond tree, *Prunus dulcis*, was extracted from the Blue Diamond Growers website. The stages contained in the model are dormant bud, green tip, pink bud, popcorn, bloom, petal fall, and jacket and are collected at 25 different DD times. However, the Blue Diamond Grower data did not specify the sample size n_i at each DD time the data was collected and only provided the percentage of trees in each stage. For analysis purposes, it was assumed $n_i = 100$ for each time t_i.

Chapter 3 Results

3.1 *C. occidentalis*

Using ML parameter estimates (Table 3-1), the model-estimated proportion of insects in development stage j, $p_j(t)$, was plotted as a function of t in DD. The actual data was compared with the model predictions for instars II-VI, pupae, and adults (Fig. 3-1). The estimated model conformed well to the actual observations at each life stage.

3.1.1 Parameter Bootstrap Confidence Interval Evaluation

The percentage of confidence intervals that captured the parameter estimates from the fitted model are approximately 95%, as per the claimed rate (Table 3-2). The exception is the quantity *v*. The coverage for *v* was consistently lower than 95%. Histograms of the bootstrap samples for the ML estimations were produced with the sample's corresponding normal curve (Fig. 3-2). The histograms conform to the normal curve and do not deviate far from the curve's estimates and are notably normal. The normal distribution was created using the sample mean and standard deviation calculated from the bootstrap sample.

3.1.2 Functions of the Parameters Bootstrap Confidence Interval Evaluation

Estimates for peak time were found (Table 3-3). The coverage rates for peak time are all near 95% (Table 3-4). Histograms of the bootstrap sample for the peak times were produced with the sample's corresponding normal curve, which calculated using the mean and standard deviation of the estimated bootstrap sample (Fig. 3-3). The histograms follow the shape of the normal curve well for each peak time.

Estimates for τ were found (Table 3-5). The value for τ was arbitrarily chosen to be $\xi = .1$ in stage 6, or the pupa stage. The coverage rates are approximately 95% (Table 3-5). Histograms of the bootstrap sample for the τ estimation with the sample's corresponding normal curve indicate normality in the sampling distribution (Figure 3-4).

3.2 *M. sanguinipes*

Using ML parameter estimates (Table 3-1), the model-estimated proportion of insects in development stage j, $p_j(t)$, was plotted as a function of t in DD. The actual data was

compared with the model predictions for instars I-VI and adult (Fig 3-1). Model estimates conformed to the actual observations at each life stage. Stages instar II, instar III, and instar IV had notably lower proportions of insects than other stages.

3.2.1 Parameter Bootstrap Confidence Interval Evaluation

The percentage of confidence intervals that captured the parameter estimates from the fitted model are approximately 95%, as per the claimed rate (Table 3-2). The exception is the quantity *v.* The coverage probability for *v* lower than 95%. Histograms of the bootstrap samples for the ML estimations were produced with the sample's corresponding normal curve (Fig. 3-5). The histograms conform to the normal curve and do not deviate far from the curve's estimates and are notably normal. The normal distribution was created using the sample mean and standard deviation calculated from the bootstrap sample.

3.2.2 Functions of the Parameters Bootstrap Confidence Interval Evaluation

Estimates for peak time were found (Table 3-3). The coverage rates for peak time are all near 95% (Table 3-5). Histograms of the bootstrap sample for the ML estimations were produced with the sample's corresponding normal curve, which calculated using the mean and standard deviation of the bootstrap sample (Fig. 3-6). The histograms follow the shape of the normal curve well for each peak time.

Estimates for τ were found (Table 3-5). The value for τ was arbitrarily chosen to be $\xi = .75$ in stage 3, or the instar IV stage. The coverage rates are approximately 95% (Table 3-5). Histograms of the bootstrap sampling distribution for the τ with the sample's corresponding normal curve indicate normality in the sampling distribution (Figure 3-7).

3.3 *P. dulcis*

Using ML parameter estimates (Table 3-1), the model-estimated proportion of trees in development stage j, $p_j(t)$, was plotted as a function of t in DD. The actual data was compared with the model predictions for dormant bud, green tip, pink bud, popcorn, bloom, petal fall, and jacket (Fig 3-1). Model estimates conformed to the actual

observations at each life stage. Stages 4 and 5 had notably lower proportions of trees than the other stages.

3.3.1 Parameter Bootstrap Confidence Interval Evaluation

The percentage of confidence intervals that captured the parameter estimates from the fitted model are approximately 95%, as per the claimed rate (Table 3-2). The exception is the quantity *v.* The coverage probability for *v* was consistently lower than 95%. Histograms of the bootstrap samples for the ML estimations were produced with the sample's corresponding normal curve (Fig. 3-8). The histograms conform to the normal curve and do not deviate far from the curve's estimates and are notably normal. However, the bootstrap sampling distribution for a_1 did have low extreme values. The normal distribution was created using the sample mean and standard deviation calculated from the bootstrap sample.

3.3.2 Functions of the Parameters Bootstrap Confidence Interval Evaluation

Estimates for peak time were found (Table 3-3). The coverage rates for peak time are all near 95% (Table 3-3). Histograms of the bootstrap sample for the ML estimations were produced with the sample's corresponding normal curve, which calculated using the mean and standard deviation of the bootstrap sample (Fig. 3-9). The histograms follow the shape of the normal curve for each peak time.

Estimates for τ were found (Table 3-5). The value for τ was arbitrarily chosen to be $\xi = .10$ in stage 5, or the bloom stage. The coverage rates are approximately 95% (Table 3-5). Histograms of the bootstrap sample for the τ estimation with the sample's corresponding normal curve indicate normality in the sampling distribution (Figure 3-10).

3.4 Sample Size Evaluation

The *C. occidentalis* data was used to evaluate the bootstrap confidence interval at differing sample sizes (Table 3-6). The coverage rates coverged quickly to approximately 95% for stage signposts $a_1, a_2, \ldots, a_6$ at $n_i = 20$. The coverage rate for v steadily increases until approximately $n_i = 50$. However, the coverage probably does not converge to 95%

with the tested sample sizes; it consistently revolves around 92-93% for sample sizes of $n_i = 50$ or greater.

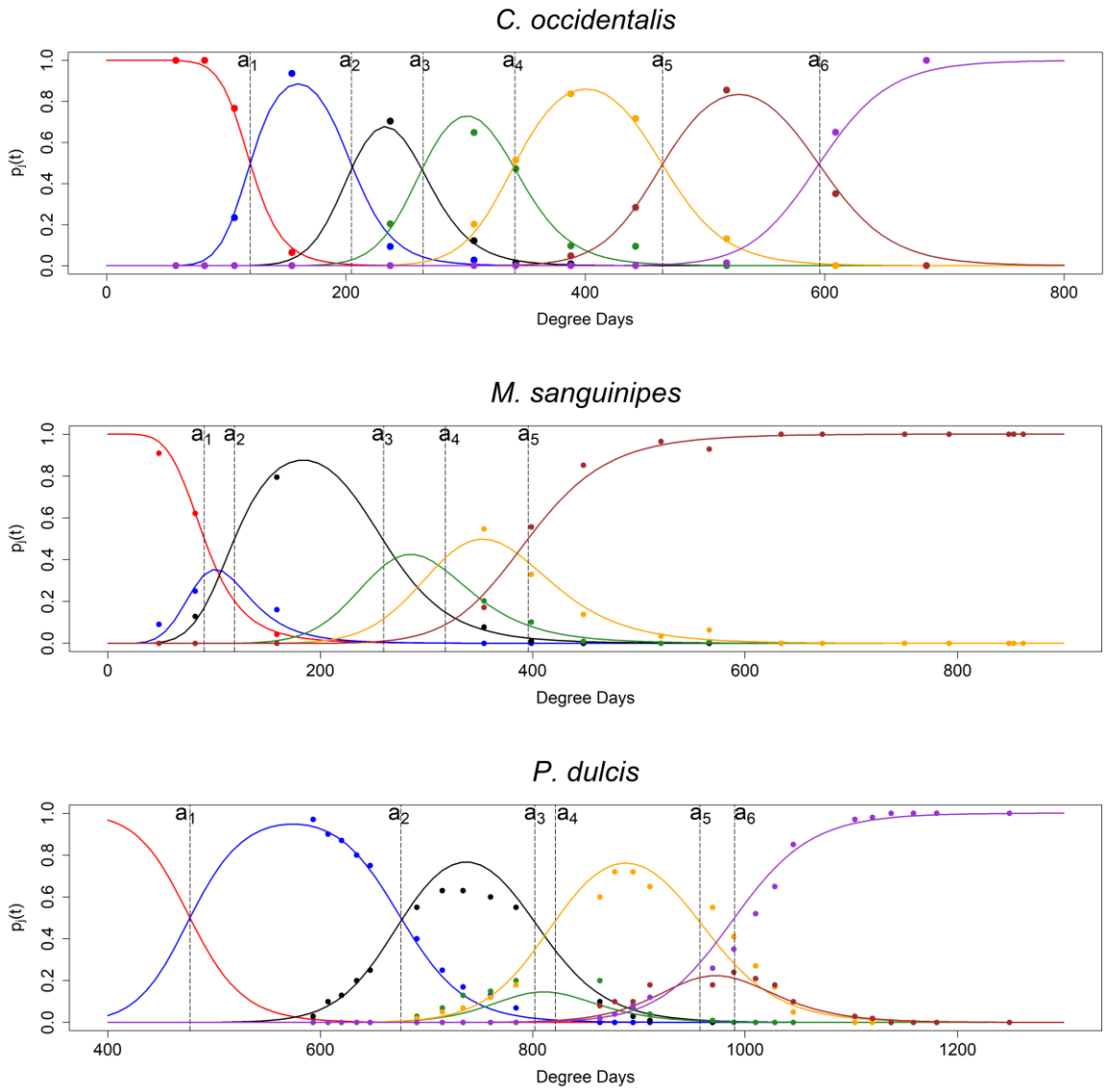

Figure 3- 1: Comparison of raw data (plotted points) with the fitted Dennis-Kemp model for the population in each life stage as a function of cumulative DD, along with phenological ML estimates.

Table 3- 1: Maximum likelihood parameter estimates for each species population (a_j in DD, v in growth/DD).

	a_1	a_2	a_3	a_4	a_5	a_6	v
C. occidentalis	120.04	204.67	264.59	341.29	464.78	595.71	1.41
M. sanguinipes	90.88	119.25	259.77	314.94	395.82	-	3.57
P. dulcis	477.34	676.09	801.93	821.16	957.44	989.95	1.31

Note: A '-' indicates that stage is not measured for that population.

Table 3- 2: Coverage percentages for ML parameter estimate bootstrap confidence intervals.

	a_1	a_2	a_3	a_4	a_5	a_6	v
C. occidentalis	94.4%	93.5%	96.1%	94.4%	96.1%	94.3%	92.6%
M. sanguinipes	95.3%	94.7%	95.9%	95.5%	94.2%	-	92.2%
P. dulcis	91.9%	94.4%	94.5%	93.9%	94.0%	94.5%	93.6%

Note: A '-' indicates that stage is not measured for that population and results were not calculated.

Table 3- 3: Estimates of peak times in each stage for each species in degree days.

	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6
C. occidentalis	160.11	232.91	301.15	400.76	528.06
M. sanguinipes	101.38	183.95	285.06	352.98	-
P. dulcis	574.22	737.29	810.23	887.58	972.37

Note: A '-' indicates that stage is not measured for that population.

Table 3- 4: Coverage percentages for peak time in each stage.

	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6
C. occidentalis	94.0%	94.6%	95.8%	95.8%	94.7%
M. sanguinipes	95.0%	94.6%	95.3%	95.8%	-
P. dulcis	92.3%	94.5%	94.2%	94.6%	94.7%

Note: A '-' indicates that stage is not measured for that population and results were not calculated.

Table 3- 5: τ estimates in degree days and corresponding coverage percentages for each species.

	τ	Coverage Probabiltiy
C. occidentalis*	535.30	94.0%
M. sanguinipes**	73.13	93.9%
P. dulcis***	733.91	93.9%

* τ was measured at $\xi = 0.1$ in stage 6, or pupa.
** τ was measured at $\xi = 0.75$ in stage 3, or instar III.
*** τ was measured at $\xi = 0.1$ in stage 5, or bloom.

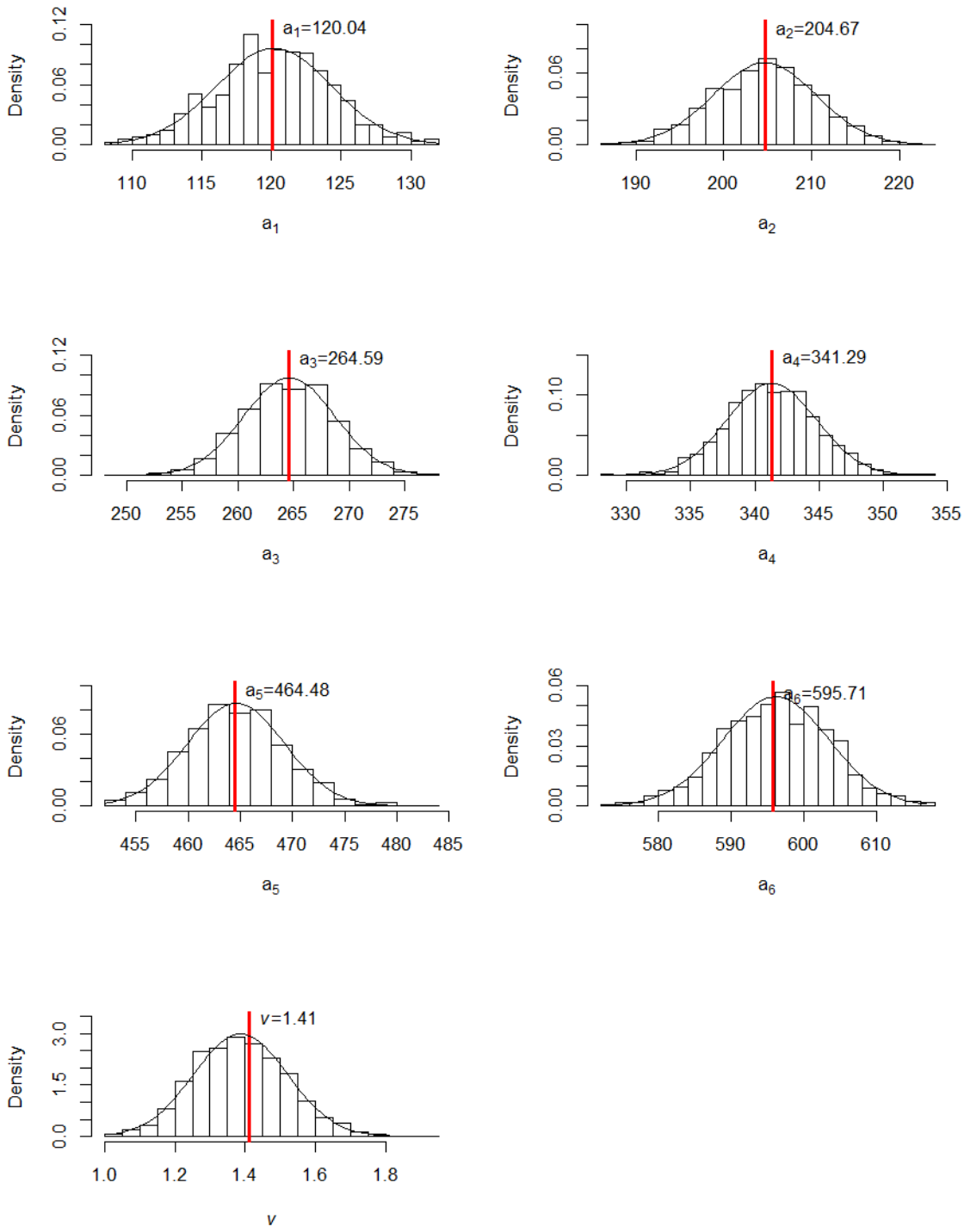

Figure 3- 2: Histograms of *C. occidentalis* parameter bootstrap distributions with a normal distribution constructed using the distribution's mean and standard deviation. ML estimates are added as reference.

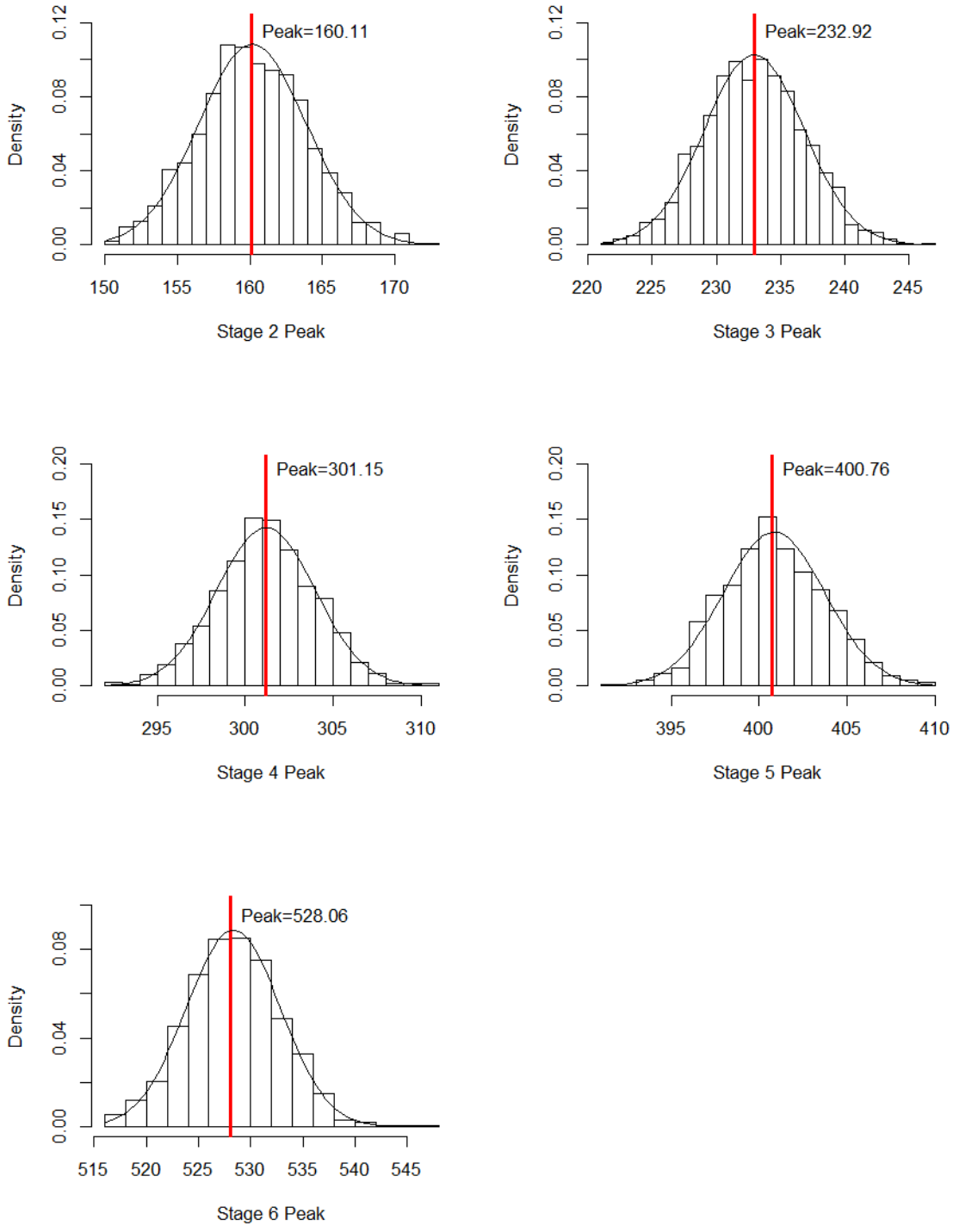

Figure 3- 3: Histograms of *C. occidentalis* peak time bootstrap distributions with a normal distribution constructed using the distribution's mean and standard deviation.

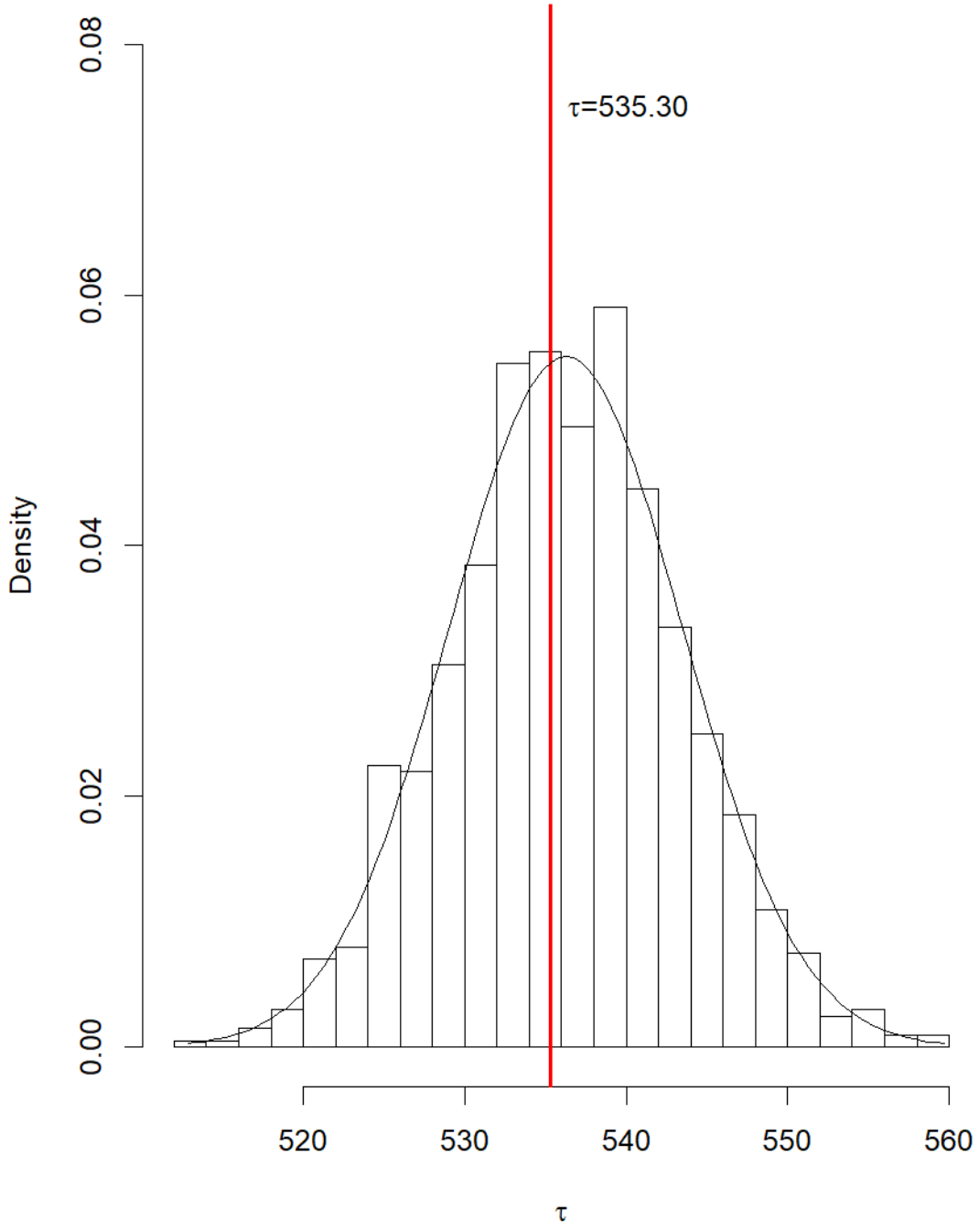

Figure 3- 4: Histograms of *C. occidentalis* τ bootstrap distribution with a normal distribution constructed using the distribution's mean and standard deviation.

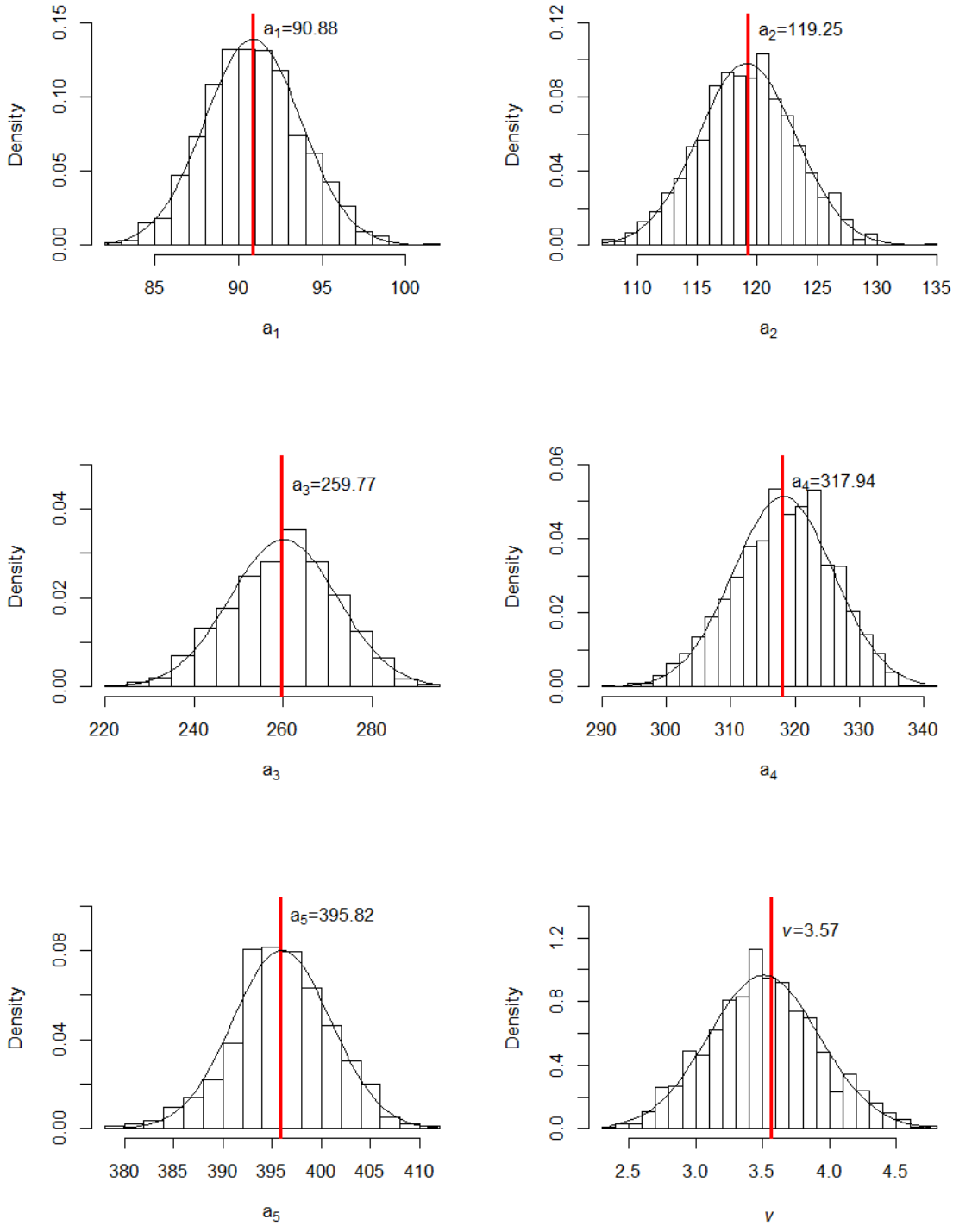

Figure 3- 5: Histograms of *M. sanguinipes* parameter bootstrap distributions with a normal distribution constructed using the distribution's mean and standard deviation. ML estimates are added as a reference.

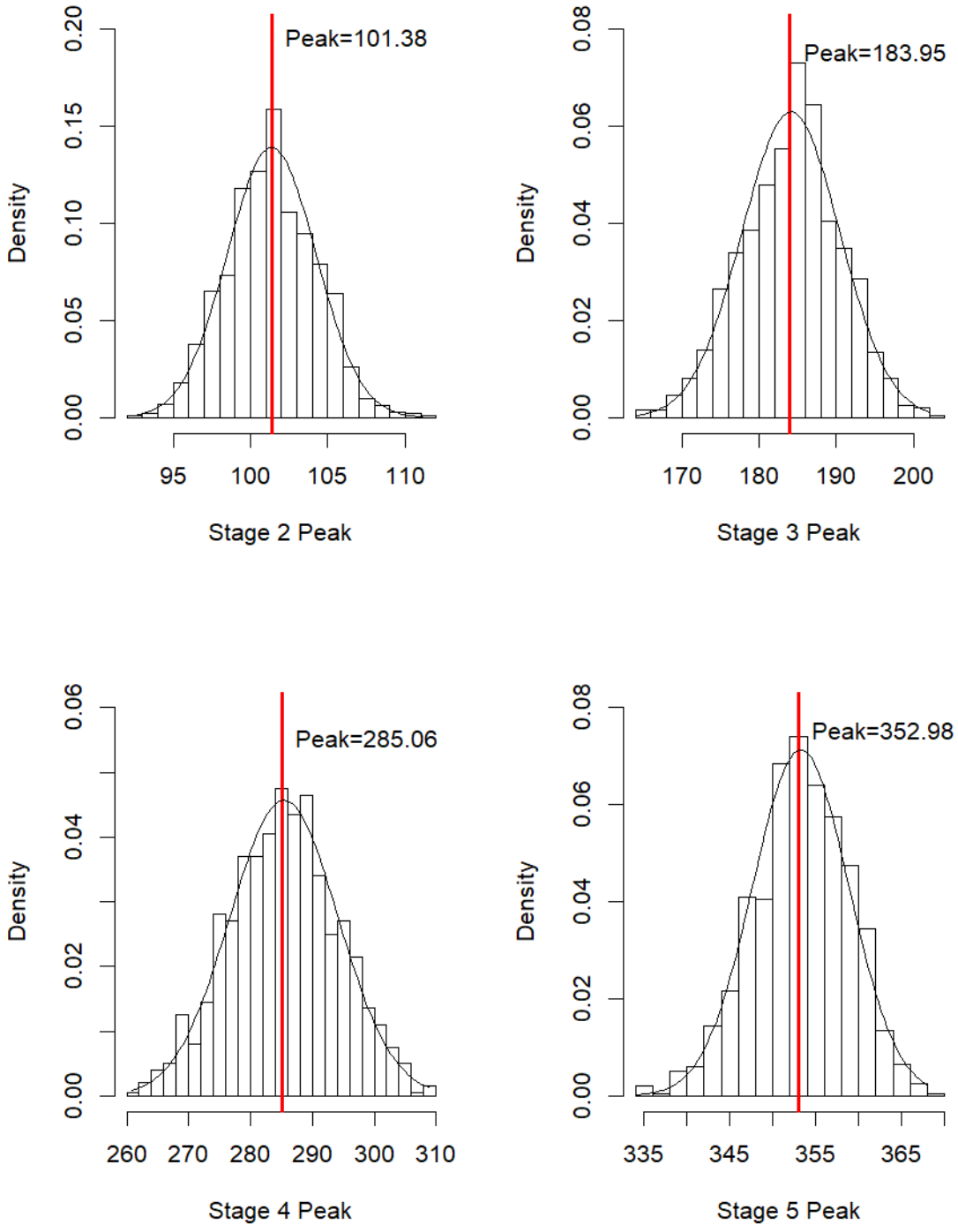

Figure 3- 6: Histograms of *M. sanguinipes* peak time bootstrap distributions with a normal distribution constructed using the distribution's mean and standard deviation.

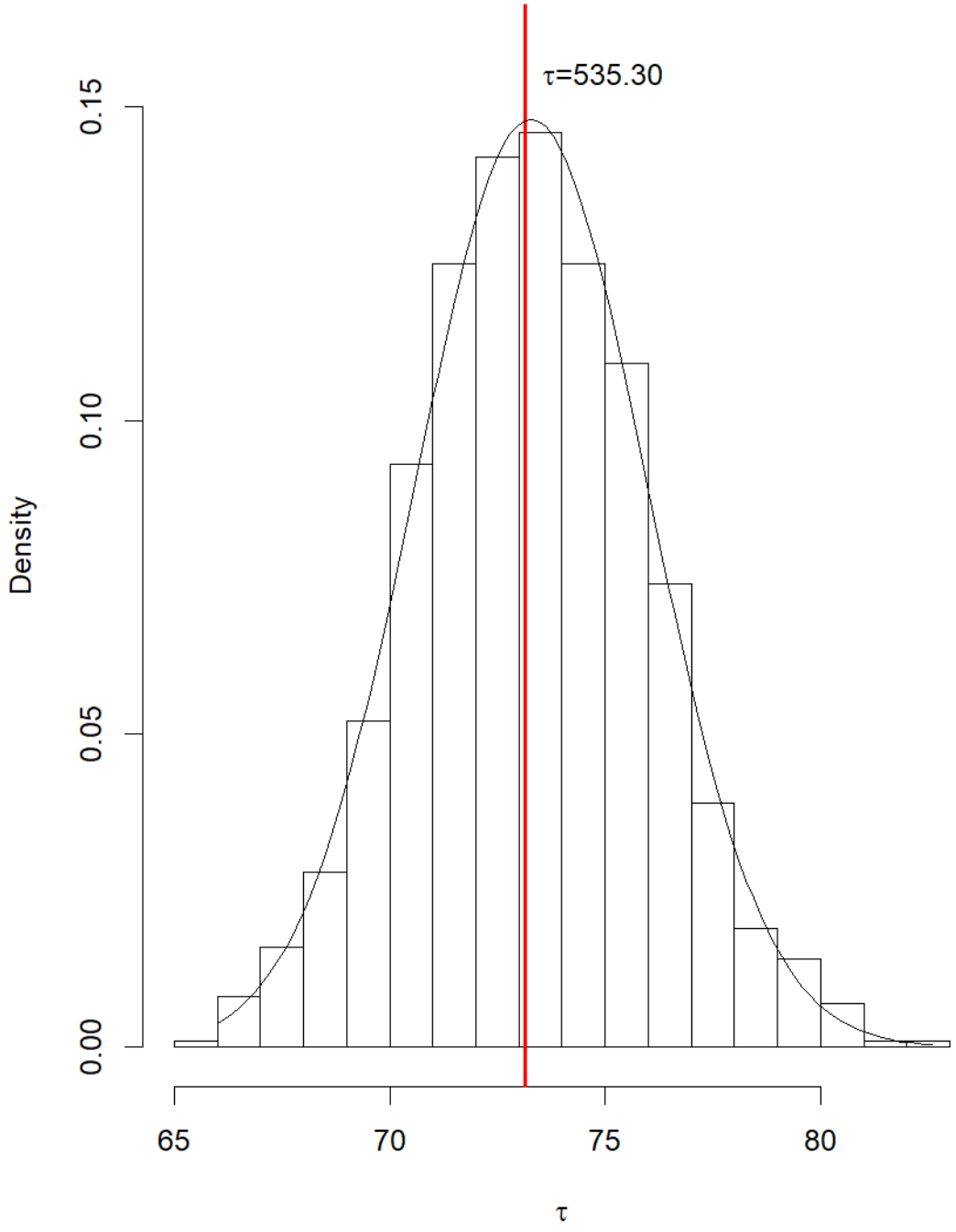

Figure 3- 7: Histogram of *M. sanguinipes* τ bootstrap distribution with a normal distribution constructed using the distribution's mean and standard deviation.

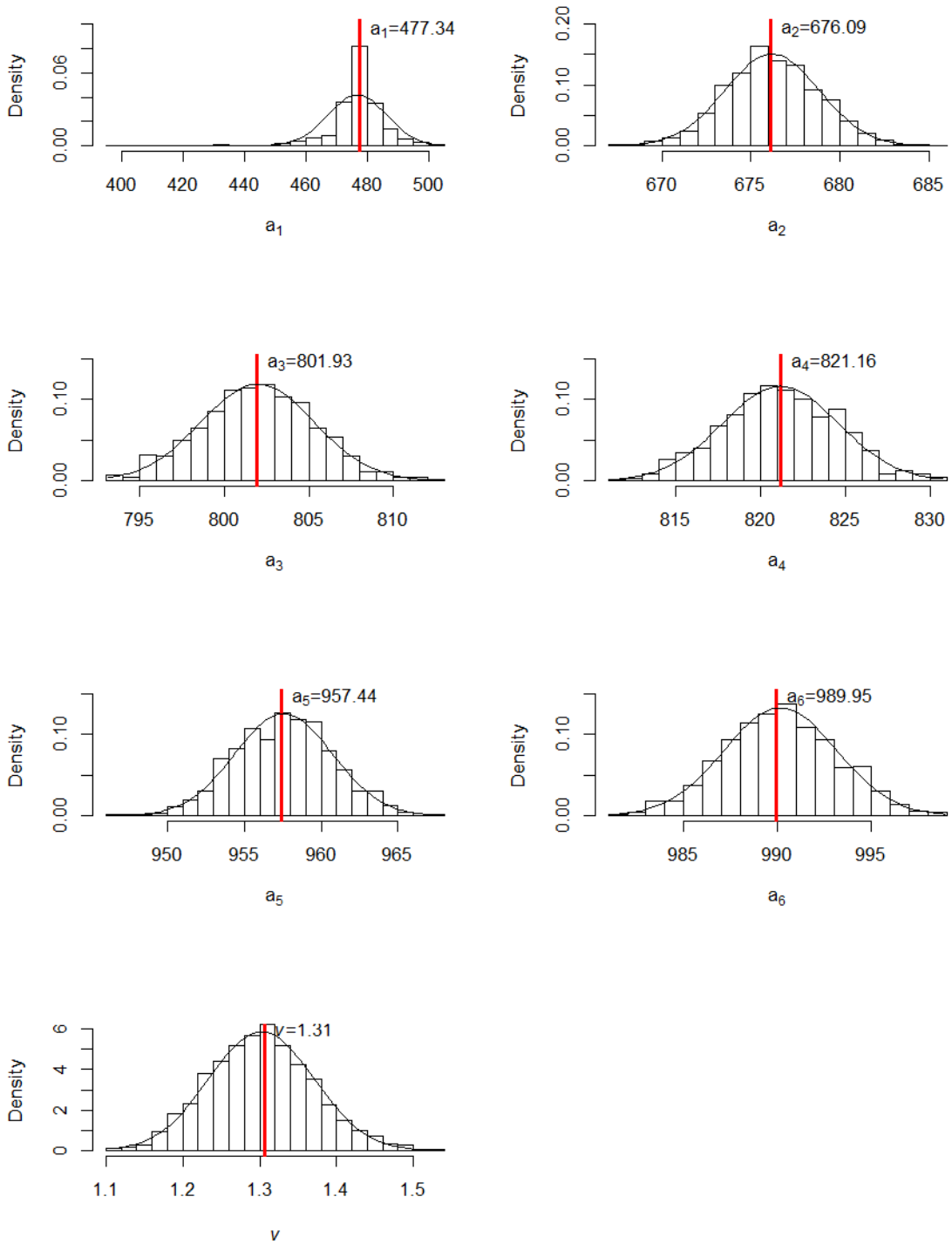

Figure 3- 8: Histograms of *P. dulcis* parameter bootstrap distributions with a normal distribution constructed using the distribution's mean and standard deviation. ML estimates are added as a reference.

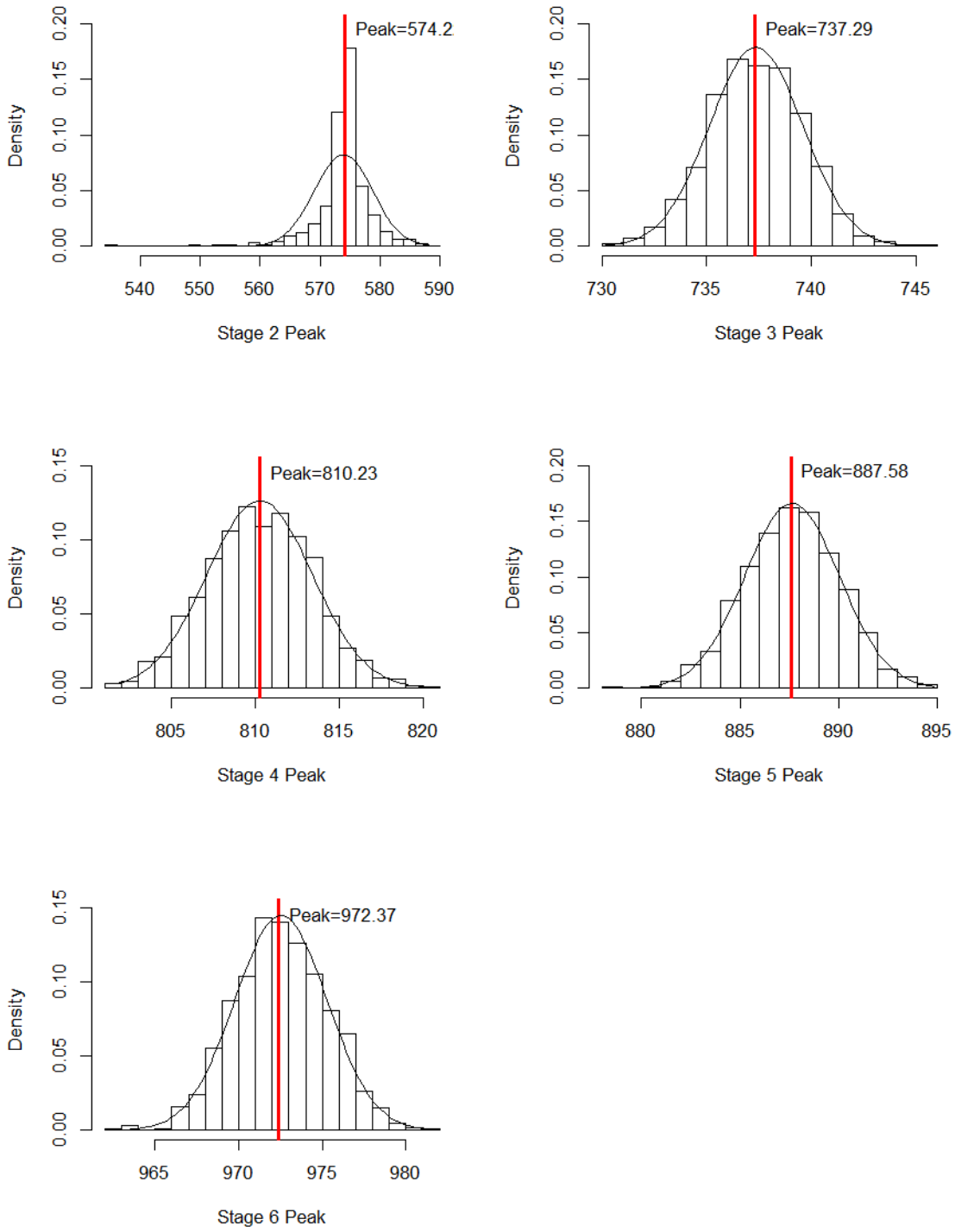

Figure 3- 9: Histograms of *P. dulcis* peak time bootstrap distributions with a normal distribution constructed using the distribution's mean and standard deviation.

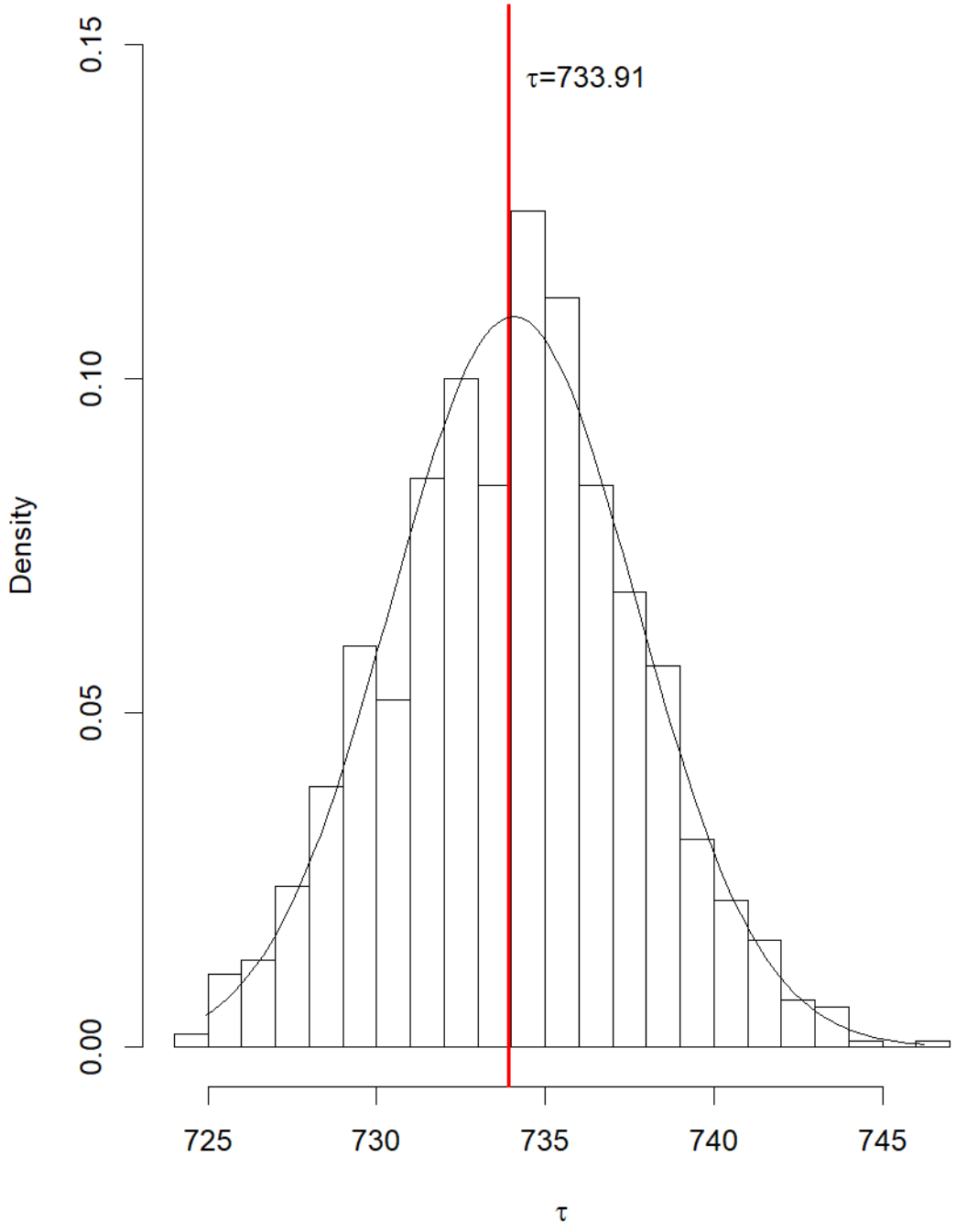

Figure 3- 10: Histogram of *P. dulcis* τ bootstrap distribution with a normal distribution constructed using the distribution's mean and standard deviation.

Table 3- 6: Coverage percentages for *C. occidentalis* at varying sample sizes.

	a_1	a_2	a_3	a_4	a_5	a_6	v
$n_i = 10$	93.9%	95.7%	92.5%	93.4%	93.5%	93.3%	80.1%
$n_i = 20$	94.2%	96.0%	95.9%	94.6%	94.9%	93.8%	86.1%
$n_i = 30$	95.0%	95.2%	94.5%	95.3%	95.3%	94.9%	90.1%
$n_i = 40$	95.5%	95.6%	94.4%	94.7%	94.2%	95.7%	90.1%
$n_i = 50$	94.1%	95.5%	94.5%	94.9%	93.5%	94.4%	92.0%
$n_i = 60$	95.0%	94.2%	94.9%	93.5%	94.3%	93.5%	91.7%
$n_i = 70$	96.7%	94.9%	94.8%	93.2%	95.3%	94.1%	94.2%
$n_i = 80$	94.0%	93.9%	92.9%	95.1%	95.3%	92.8%	92.7%
$n_i = 90$	95.4%	95.6%	93.9%	96.4%	93.5%	93.5%	93.3%
$n_i = 100$	94.5%	96.0%	94.5%	94.8%	96.0%	94.7%	92.4%

Chapter 4 Discussion

4.1 Discussion

The Dennis-Kemp phenology model used here worked well with sparse datasets. The model and the maximum likelihood procedures accurately describe the time-dependent life stage data of insects and trees. The parameters and functions of the parameters have excellent estimability. The histograms of the bootstrap sampling distributions indicate that the ML parameter vector estimates can be described well with a large sample multivariate normal distribution described in Dennis and Kemp 1988. The coverage properties of the bootstrap confidence intervals were also acceptable. The observed coverage probability was mostly consistent with the claimed 95% rate. However, the estimation of v is more variable versus the other parameters. No evidence was found of multiple modes during ML estimation. Likely, the multiple modes, when they occur, are an artifact of the Nelder-Mead algorithm rather than an estimability issue arising from the data, discussed by Irvine (2011). Potentially, if the data table is sufficiently sparse that certain development stages are representaly poorly or not at all, multimodality or other estimability problems could arise.

The bootstrap confidence interval was also evaluated with different sample sizes to find the size *n* needed for each sample at time *i* to get adequate coverage rates with the *C. occidentalis* data. The ML estimates of $a_1,\ a_2,\ \ldots, a_{r-1}$ converge quickly to normality and the 95% coverage rate. However, *v* had lower coverage rates at smaller sample sizes. A sample size of approximately $n_i = 70$ is suggested for reliable results. Collecting fewer organisms at sampling time t_i will reduce the estimability of the parameters.

Further sampling problems could arise regarding how often samples should be collected. Collecting fewer samples will result in limited information to accurately predict model parameters and construct confidence intervals. Additionally, estimability issues may aoccur as the algorithms may get stuck in local minima or maxima. It could also be more difficult to find ideal initial values. As such, they may be chosen poorly and would likely result in unsatisfactory ML estimates. The number of times that samples are collected should be designed such that they record sufficient information on the changes over time in all the development stages. If a sufficient number of sampling times cannot be collected, researchers could potentially consider pooling adjacent life stages into single operational stages for analysis.

This issue was found with the *C. occidentalis* data with varying values of sampling times t_i. The same stages were kept. The ML parameter estimates calculated from half the sampling times, $q = 6$, resulted in estimations that biologically did not make sense (Table 4-1). The $q = 8$ sampling times resulted in ML estimates that could be used to evaluate estimability. However, the ML estimates were using the $q = 8$ sampling times were consistently far lower than the ML estimates from the original data set. Further work is needed in order to determine an appropriate number of sampling times for adequate estimability.

Table 4- 1: ML estimates for varying amounts of sampling times *t* for comparison for *C. occidentalis,* where *q* denotes the number of sampling times.

	a_1	a_2	a_3	a_4	a_5	a_6	v
$q = 6$	66.4	104.45	325117.54	0	202.53	272.41	0.68
$q = 8$	81.81	94.27	128.74	188.96	308.55	345.57	1.68

The Dennis-Kemp model does not explicitly incude several factors that may affect phenology. The model does not have terms to account for mortality differences in the stages or incorporate heterogeneity in development rates. Other covariates that may be included in the model might include temperature, location, light, or any other variables hypothesized to affect phenology. Further work is needed to develop and incorporate the variables into the phenology model.

The Dennis-Kemp phonology model is recommended for use in areas such as ecology, rangeland management, agriculture or other applications where the specialized count data might occur. The parameters in the Dennis-Kemp phenology model are well-estimated with the sample sizes and sparse data tables encountered in published literature. The Dennis-Kemp model is a simpler model than many alternatives in the literature, and yet it provides satisfactory results and has decent estimability.

References

Bishop, Y. M., S. E. Fienberg & P. W. Holland, 1975. Discrete multivariate analysis: theory and practice. Massachusetts Institute of Technology, Cambridge, Mass.

Blue Diamond Growers. 2017. In the field. http://bluediamondnuts.com/applications/in-the-field/index.cfm?navid=101.

Bruce, R. L., C. M. Moffitt, and B. Dennis. 2009. Survival and passage of ingested New Zealand mudsnails through the intestinal tract of rainbow trout. North American Journal of Aquaculture 71:287-301.

Castle, S. J., S. E. Naranjo, J. L. Bi, F. J. Byrne, & N. C. Toscano. 2005. Phenology and demography of *Homalodisca coagulate* (Hemiptera: Cicadellidae) in southern

California citrus and implications for management. Bulletin of Entomological Research 95(6):621-634.

Chihara, L. & T. Hesterberg. 2011. Mathematical Statistics with Resampling and R. John Wiley & Sons, Inc. Hoboken, New Jersey.

Dennis, B., & W. P. Kemp 1988. Further statistical inference methods for a stochastic model of insect phenology. Environmental Entomology 17(5):887-893.

Dennis, B., W. P. Kemp, & R. C. Beckwith.1986. Stochastic model of insect phenology: estimation and testing. Environmental Entomology 15:540-546.

Efron, B., & R. J. Tibshirani. 1993. An introduction to the bootstrap. Chapman & Hall, New York, NY.

Irvine, P. M. 2011. Using a logistic phenology model with improved degree-day accumulators to forecast emergence of pest grasshoppers. University of Lethbridge Institutional Repository.

Jenrich, R. I. & R. H. Moore. 1975. Maximum likelihood estimation by means of nonlinear least squares, pp 57-65). In Proceedings, Statistical Computing Section, American Statistical Society.

Lagarias, J. C., J. A. Reeds, M. H. Wright, and P. E. Wright. 1998. “Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions.” SIAM Journal of Optimization 9:112–147.

Kemp, W.P. & J.A. Onsager. 1986. Rangeland grasshoppers (Orthoptera: Acrididae): modeling phenology of natural populations of six species. Environ. Entomol. 15:924–930.

Kemp, W. P., B. Dennis & R. C. Beckwith. 1986. Stochastic phenology model for the western spruce budworm (Lepidoptera: Tortricidae). Environ. Entomology. 15: 541-548.

Knape, J & P. de Valpine. 2016. Monte Carlo estimation of stage structured development from cohort data. Ecology 97(4):992-1002.

Mooney, C. Z. 1997. Monte Carlo Simulation. SAGE Publications, Thousand Oaks, California.

Osawa, A., C. A. Shoemaker & J. R. Stedinger. 1983. A stochastic model of basalm fir bud phenology utilizing maximum likelihood parameter estimation. Forestry Science 21: 478-490.

Pawitan, Y. 2001. In all likelihood: statistical modelling and inferences using likelihood. Oxford University Press, Oxford, UK.

Van den Boogard, H. F. R. & M. J. Hall. 2004. The constriction of confidence intervals for frequency analysis using resampling techniques: a supplementary note. Hydrology and Earth Systems Sciences Discussions, European Geosciences Union 8(6):1174-1178).

Venables, W. N. & B. D. Ripley. 2002. Modern applied statistics with S (Statistics and computing). 4th Edition. New York: Springer-Verlag.

Volney, W. J. A., and H. F. Cerezke. 1991. The phenology of white spruce and the spruce budworm in northern Alberta. Can. J. For. Res. 22:198-205.

Appendix 1: Original Phenology Data

Table A1- 1: Original data set with counts for *C. occidentalis* with stages and cumulative DD times recorded.

Time (DD)/Stage	Instar II	Instar III	Instar IV	Instar V	Instar VI	Pupa	Adult
58	16	0	0	0	0	0	0
82	10	0	0	0	0	0	0
107	23	7	0	0	0	0	0
155	3	44	0	0	0	0	0
237	0	6	45	13	0	0	0
307	0	2	9	48	15	0	0
342	0	0	1	34	37	0	0
388	0	0	1	10	87	5	0
442	0	0	0	7	53	21	0
518	0	0	0	0	20	65	1
609	0	0	0	0	0	14	26
685	0	0	0	0	0	0	42

Table A1- 2: Original data set with counts for *M. sanguinipes* with stages and cumulative DD times recorded.

Time (DD)/Stage	Instar I	Instar II	Instar III	Instar IV	Instar V	Adults
48.13	10	1	0	0	0	0
82.38	82	33	17	0	0	0
159.09	3	11	54	0	0	0
354.08	0	0	5	13	35	11
398.65	0	0	1	9	29	49
447.79	0	0	0	1	13	80
520.88	0	0	0	0	3	83
566.16	0	0	0	1	8	116
633.94	0	0	0	0	0	70
672.78	0	0	0	0	0	143
750.1	0	0	0	0	0	63
792.09	0	0	0	0	0	83
848.28	0	0	0	0	0	8
852.93	0	0	0	0	0	18
861.73	0	0	0	0	0	27

Table A1- 3: Original data set for *P. dulcis* with stages and cumulative DD times recorded. It is unknown if the cells are filled with counts or probabilities.

Time (DD)/ Stage	Dormant Bud	Green Tip	Pink Bud	Popcorn	Bloom	Petal Fall	Jacket
593.24	0	97	3	0	0	0	0
607.24	0	90	10	0	0	0	0
619.97	0	87	13	0	0	0	0
634.18	0	80	20	0	0	0	0
647.08	0	75	25	0	0	0	0
690.87	0	40	55	3	2	0	0
714.87	0	25	63	7	5	0	0
734.37	0	17	63	13	7	0	0
760.37	0	13	60	15	12	0	0
784.37	0	7	55	20	18	0	0
863.37	0	0	10	20	60	8	2
877.37	0	0	4	10	72	10	4
894.37	0	0	3	7	72	10	8
910.37	0	0	1	4	65	18	12
969.37	0	0	0	1	55	18	26
989.37	0	0	0	0	41	24	35
1009.87	0	0	0	0	27	21	52
1027.87	0	0	0	0	17	18	65
1045.37	0	0	0	0	5	10	85
1103.37	0	0	0	0	0	3	97
1119.87	0	0	0	0	0	2	98
1137.37	0	0	0	0	0	0	100
1158.37	0	0	0	0	0	0	100
1180.37	0	0	0	0	0	0	100
1248.87	0	0	0	0	0	0	100