The Statistical Analysis of Insect Phenology

PAUL A. MURTAUGH,[1,2] SARAH C. EMERSON,[1] PETER B. MCEVOY,[3]
AND KIMBERLEY M. HIGGS[3]

Environ. Entomol. 41(2): 355–361 (2012); DOI: http://dx.doi.org/10.1603/EN11260

ABSTRACT We introduce two simple methods for the statistical comparison of the temporal pattern of life-cycle events between two populations. The methods are based on a translation of stage-frequency data into individual 'times in stage'. For example, if the stage-*k* individuals in a set of samples consist of three individuals counted at time t_1 and two counted at time t_2, the observed times in stage *k* would be $(t_1, t_1, t_1, t_2, t_2)$. Times in stage then can be compared between two populations by performing stage-specific *t*-tests or by testing for equality of regression lines of time versus stage between the two populations. Simulations show that our methods perform at close to the nominal level, have good power against a range of alternatives, and have much better operating characteristics than a widely-used phenology model from the literature.

KEY WORDS phenology, logistic, *t*-test, *Tyria jacobaeae*

Quantitative models of phenology are useful for describing the timing of life-cycle events and for comparing patterns of development between different populations, locations, and times. Insect phenology has been an important component of studies of pest management (Petitt et al. 1991, Candy 2003), interactions between insect species and their host plants (Volney and Cerezke 1992), spatial and temporal variation in development rate (Weber et al. 1999), and effects of climate change (Hodgson et al. 2011).

Many approaches have been suggested for modeling and analyzing phenology data, including logistic stochastic processes (Dennis et al. 1986), semi-Markov processes (Munholland and Kalbfleisch 1991), generalized linear models (Manel and Debouzie 1997), continuation ratios (Candy 2003), circular statistics (Morellato et al. 2010), and generalized additive models (Hodgson et al. 2011). Here we present new methods for the statistical comparison of phenology between populations, based on *t*-tests and simple linear regression. We discuss implementations of the methods that are appropriate for simple random sampling and cluster sampling of insects. In addition, we use simulation to compare the performance of our methods to that of the logistic phenology model of Dennis et al. (1986), an alternative approach that has been widely used and discussed.

We claim that our approach is superior for comparing phenology between different populations or locations. Other models (e.g., degree-day models for specific organisms) may be superior for other uses, such as predicting the timing of developmental events.

A motivating example used throughout the paper is the phenology of the cinnabar moth, *Tyria jacobaeae* L. (Lepidoptera: Arctiidae), in the Pacific Northwest region of the United States. This moth, which was introduced to North America as a biocontrol agent for ragwort (McEvoy et al. 1991), develops from the egg to adult through five larval stages and a pupal stage.

Notation and Data Structure. We are interested in investigating the development of an organism that has a life cycle with K prepupal stages. For the cinnabar moth we have $K = 6$, for egg, L1, L2, L3, L4, and L5. At each of J sampling occasions, counts are obtained for each development stage; n_{jk} is the observed count for sampling occasion j, $j = 1, \ldots, J$, and stage k, $k = 1, \ldots, K$. Let t_j be the time (calendar time or accumulated degree-days) associated with the j^{th} sample. Such data may be gathered for more than one population or location, and the goal is then to compare the timing of life-cycle events between different populations, and to perform tests of the null hypothesis that the phenological patterns are the same across populations.

The Logistic Phenology Model. Dennis et al. (1986) proposed a logistic phenology model based on the bud phenology model of Osawa et al. (1983). The amount of development up to time t is modeled as a stochastic process $S(t)$. The process $S(t)$ is modeled with a logistic distribution with mean t and variance $\frac{\pi^2}{3}\beta_0^2 t$, where β_0 is a parameter that determines the spread of the stochastic process at time t. Parameters $-\infty \leq \beta_1 \leq \ldots \leq \beta_{K-1} \leq \infty$ are then used to model development as follows:

[1] Department of Statistics, Oregon State University, Corvallis, OR 97331.

[2] Corresponding author, e-mail: murtaugh@stat.oregonstate.edu.

[3] Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331.

0046-225X/12/0355–0361$04.00/0 © 2012 Entomological Society of America

$$\begin{aligned} P(\text{Stage 1 at time } t) &= P(S(t) \le \beta_1)\\ P(\text{Stage 2 at time } t) &= P(\beta_1 \le S(t) \le \beta_2)\\ &\;\;\vdots\\ P(\text{Stage } K \text{ at time } t) &= P(\beta_{K-1} \le S(t)) \end{aligned} \qquad [1]$$

The model may be fit to the observed data using a maximum likelihood approach, conditioning on the number of individuals observed at each sampling occasion. Hypothesis tests may be performed to compare the resulting estimates of $(\beta_0, \beta_1, \ldots, \beta_{K-1})$ between two populations; details are given in *Appendix* 1 and in the original paper by Dennis et al. (1986).

Comparison of Observed Times in Stage. We present two simple approaches to comparing phenology between two populations that are based on the "average" times that the organism is observed in a given stage. A set of observed times for each of the K stages is constructed from the count data as follows. Let $N_K = \sum_{j=1}^{J} n_{jk}$ be the total number of individuals observed in stage k, and define

$$\mathbf{X}_k = (X_{k1}, X_{k2}, \ldots, X_{kN_k})$$

to be the set of observed times obtained by repeating time t_j a total of n_{jk} times for each $j = 1, \ldots, J$. The average time in stage k is then

$$\bar{\mathbf{X}}_k = \sum_{i=1}^{N_k} X_{ki}/N_k\,. \qquad [2]$$

Table 1. Hypothetical data

Sample	Time	Number observed	
		Stage 1	Stage 2
1	2.6	3	0
2	4.7	5	1
3	4.8	2	4
4	5.1	1	7

For instance, assume the data for a particular population are as in Table 1. Then the sets of observed times would be:

$$\mathbf{X}_1 = (2.6, 2.6, 2.6, 4.7, 4.7, 4.7, 4.7, 4.7, 4.8, 4.8, 5.1)$$

$$\mathbf{X}_2 = (4.7, 4.8, 4.8, 4.8, 4.8, 5.1, 5.1, 5.1, 5.1, 5.1, 5.1, 5.1).$$

We consider two ways of using the time-in-stage data to compare two populations.

1. The *t*-test approach. To compare stage k between two populations, Welch's (1951) two-sample t-test can be performed on $(\mathbf{X}_k)_1$ and $(\mathbf{X}_k)_2$, the observed times for stage k in populations 1 and 2, respectively. Let p_k be the resulting P value. The smaller the value of p_k, the stronger the evidence against the null hypothesis that the mean time in stage k is the same in the two populations.

A "global" test of the hypothesis that the two populations have identical phenology can be based on Fisher's meta-analysis method of combining P values (Fisher 1970). Calculate

$$Q = -2\sum_{k=1}^{K} \log(p_k),$$

where p_k is the P value obtained for stage k. Under the null hypothesis, assuming independence of the P values, Q has a χ^2_{2K} distribution.

There may be some weak dependence among the stage-specific P values. For example, if stage k tends to have a larger mean time (or accumulated degree days) in one population, that will tend to increase the sizes of the means for subsequent stages in that population. However, these effects will likely be small under the null hypothesis of identical distributions between the two populations.

2. The Linear Regression Approach. For each observation in the data set, form a triplet of values, (X_i, Y_i, Z_i), where X_i is the stage of individual i (coded as $1, 2, \ldots, K$), Y_i is the time (or degree days) at which individual i was encountered, and Z_i identifies which of the two populations individual i is from. Then fit two regression models, using ordinary least squares:

Model 1: $E(Y_i) = \beta_0 + \beta_1 X_i$
Model 2: $E(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i$

Model 2 fits a separate regression line for each population. To test the null hypothesis that the two lines are the same (i.e., $\beta_2 = \beta_3 = 0$), we can do an extra-sum-of-squares test comparing Models 1 and 2. Let RSS_j and df_j be the residual sum of squares and residual degrees of freedom, respectively, for Model j $(j = 1, 2)$. Compute

$$F^* = \frac{\text{RSS}_1 - \text{RSS}_2}{\text{df}_1 - \text{df}_2} \div \frac{\text{RSS}_2}{\text{df}_2}, \qquad [3]$$

and obtain a P value by comparing F^* to an F distribution with $(\text{df}_1 - \text{df}_2)$ numerator degrees of freedom and df_2 denominator degrees of freedom.

A simple example of the use of these approaches is presented, along with R code, in *Appendix* 2.

Both of these approaches can be extended easily to compare more than two populations. The t-test method generalizes to Welch's contrast test, which does not assume equal variance between the groups being compared (Welch 1951). At each stage, Welch's contrast test is used to test that all of the groups have the same mean observation time for that stage; then Fisher's combined P value is computed in exactly the same way as for two populations. The regression method is extended easily by incorporating a factor variable that indicates which population each observation comes from. A full model containing stage, the population factor, and all of the interaction terms between stage and population is then compared to a reduced model that includes only stage, generating an *F*-statistic as in Equation 3. For both of these approaches, a significant result would imply that it is unlikely that all of the populations have exactly the same phenology pattern. Identifying which populations differ would require post-hoc tests and possible adjustments for multiple comparisons.

The Case of Cluster Sampling. As presented above, the logistic phenology model and the two approaches based on time in stage all assume that we have simple random samples of the animals in two populations. In practice, it is more likely that some form of cluster sampling is used to gather data. For example, Kemp et al. (1986) enumerated budworms on individual branches removed from randomly selected trees. If cluster sampling is used, it is essential that the analysis method account for the dependence in the data that this induces.

Suppose we randomly select *P* plants from each population and then enumerate all of the animals on each plant. If X_{kps} is the time (or degree days) associated with the s^{th} individual of stage *k* on plant *p*, and N_{kp} is the total number of stage-*k* individuals on plant *p*, then our data for stage *k* look like:

$$\mathbf{X}_k = [(X_{k11}, X_{k12}, \ldots, X_{k1N_{k1}}), (X_{k21}, X_{k22}, \ldots, X_{k2N_{k2}}), \ldots, (X_{kP1}, X_{kP2}, \ldots, X_{kPN_{kP}})].$$

The mean for stage *k* in a particular site is

$$\bar{\mathrm{X}}_{\mathrm{k}} = \frac{1}{\sum_{\mathrm{p}=1}^{\mathrm{P}} \mathrm{N}_{\mathrm{kp}}} \cdot \left(\sum_{\mathrm{p}=1}^{\mathrm{P}} \sum_{\mathrm{s}=1}^{\mathrm{N}_{\mathrm{kp}}} \mathrm{X}_{\mathrm{kps}} \right), \qquad [4]$$

and the stage-*k* difference between locations is

$$D_k = \bar{X}_k(\text{location } 1) - \bar{X}_k(\text{location } 2). \qquad [5]$$

Because of the clustering of observations on plants, the two methods based on time in stage, as described in the preceding section, underestimate the variance of the observed times. Simulations show that this can result in grossly inflated rejection rates when the null hypothesis is true. We assume the same problem affects the logistic phenology model, but we have not explored this. We developed bootstrapping approaches to properly estimate the variances for our two methods when cluster sampling is used; these are described in *Appendix* 3.

Comparison of the Methods. We used simulation to compare the performances of the methods under the assumption of random sampling. The following procedure was performed 500 times for each set of simulations. We chose random starting times in a 2-wk window for two sets of 5,000 eggs. For each individual, we randomly selected durations of six stages (egg through L5) from a normal distribution having a mean of 10 d and standard deviation of 2 d. The individuals were allowed to grow according to these stage durations, over a 10-wk period. At weekly intervals, random samples of size equal to 1% of the starting population were taken (with replacement) from the two sets of insects, and the numbers of individuals in the different stages were recorded. We then compared the two sets of sampled individuals using the logistic phenology model and the two approaches based on time in stage.

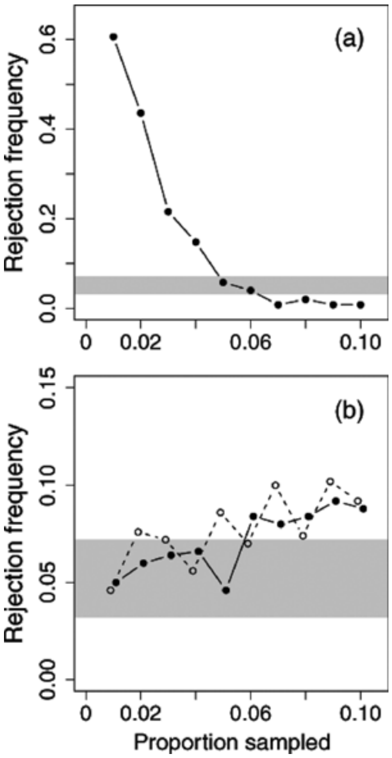

Fig. 1. Rejection rates of (a) the logistic phenology model, and (b) the new methods based on *t*-tests (solid line) and linear regression (dashed line), applied to data simulated under the null hypothesis. The horizontal axis gives the proportion of the original cohort (of 5,000 animals) sampled for each combination of location and date. See the text for other details of the simulations. Each point represents 500 simulations. Points falling within the gray regions are statistically indistinguishable from 0.05 (at the 0.05 level).

First, we consider the performances of the methods when the null hypothesis is true, i.e., when there are truly no differences in phenology between populations. As shown in Fig. 1, the rejection rates for the logistic phenology model are strongly dependent on the proportion of the original cohort (of 5,000 animals) that is sampled, and they can stray widely from the nominal 0.05: for the smallest samples in the figure, the rejection frequency was 0.61, and for the largest samples, it was just 0.008. This behavior alone suggests that the logistic model is not a useful tool for comparing phenology patterns between populations.

The two methods based on time in stage have close to the nominal level for all sample sizes, though there is a tendency for rejection frequencies to increase as sample size increases (Fig. 1b). This appears to be due to dependence that is introduced when a sizeable fraction of the original cohort is sampled.

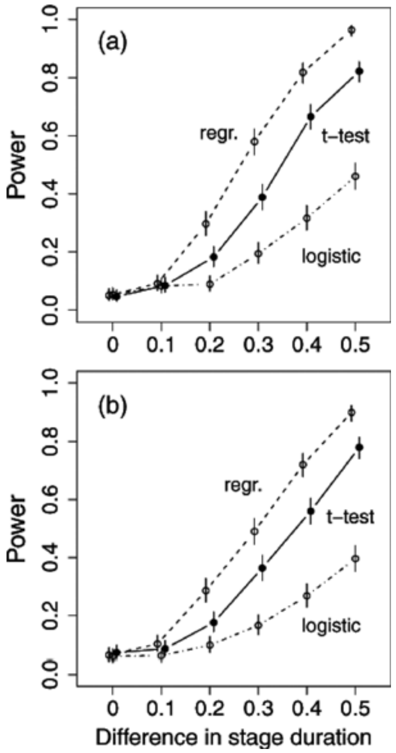

Fig. 2. Rejection rates of the *t*-test and regression-based methods and the logistic phenology model under a range of alternatives. The rates for the logistic model are calibrated to give a 0.05 rejection probability under the null hypothesis (cf. Figure 1a). In (a), the baseline stage durations are 10 d for all six stages; in (b), the durations are 12, 11, 10, 9, 8 and 7 d, for egg through L5. The horizontal axis shows the difference in mean stage durations between populations, assumed the same for all six stages. Each point represents 500 simulations; vertical lines are 95% confidence intervals.

Fig. 2 shows power curves for the two methods based on time in stage, for increasing differences in mean stage duration between populations. In addition, we show calibrated power for the logistic method, i.e., we use a cutoff for statistical significance (4.5×10^{-7}) that would yield a 5% level under the null hypothesis for this sample size (cf. Figure 1a). The method based on linear regression is more powerful than the method based on *t*-tests for the alternatives used for Fig. 2, and for other alternatives considered in *Appendix* 4. Surprisingly, the regression method works well even when the true relationship between time and stage is nonlinear (Fig. 2b, *Appendix* 4). The logistic method, even after calibration, is inferior to the other two methods.

Application to the Cinnabar Moth Data. We compared the phenology of the cinnabar moth between two sites in western Oregon: one in the Coast Range and the other in the Willamette Valley. (These two sites were chosen arbitrarily from four that were sampled by PBM and KMH in 2010.) On each of 11 (Coast Range) or 12 (Willamette Valley) sampling dates, thirty plants were randomly selected from 50 previously marked plants, and all of the eggs and larvae on each plant were enumerated. Pooled over all sampling dates, 12,021 individuals were counted from the Coast Range site, and 5,768 individuals were counted from the Willamette Valley site. (Because sampling was nondestructive, it is possible that some individuals were counted on more than one date.)

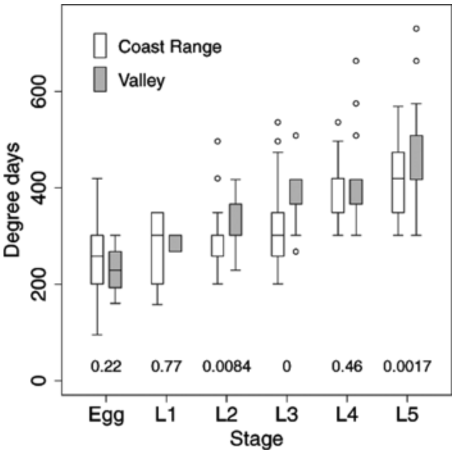

Fig. 3. Boxplots of observed degree-days by stage for two locations, based on the cinnabar moth data. The numbers across the bottom of the plot are stage-specific *P* values from the *t*-test approach (the *P* value for L3 is 5.2×10^{-8}). The global *P* value from the *t*-test approach and the single *P* value from the regression-based approach are both highly significant (see text). All *P* values are based on bootstrap-corrected variance estimates, as described in *Appendix* 3.

We used the two methods based on time in stage to compare phenology between these two locations. Because of the cluster sampling used in the gathering of data, it was essential to use the bootstrapping estimation of standard errors described in *Appendix* 3. Simulations show that, if we were to use the unadjusted methods that are appropriate for random sampling, we would greatly overestimate the evidence for differences between locations.

Fig. 3 shows a graphical summary of the observed times in stage for the two sites. Based on the *t*-test approach, the stage-specific *P* values are statistically significant (<0.05) for three stages and nonsignificant for three stages. The global *P* value from the *t*-test approach is 1.5×10^{-8}, and the single *P* value from the regression-based approach is 5.9×10^{-8}, suggesting real differences in phenology between these two sites.

To illustrate the importance of adjusting standard errors for the cluster sampling that was used in data collection, we redid the stage-specific comparisons summarized in Fig. 3, ignoring the clustering of insects on plants. The *P* values from the resulting *t*-tests (with the adjusted *P* values from Fig. 3 in parentheses) are: egg, 4.3×10^{-26} (0.22); L1, 0.049 (0.77); L2, 1.5×10^{-70} (0.0084); L3, 7.6×10^{-154} (5.2×10^{-8}); L4, 0.24 (0.46); and L5, 1.4×10^{-15} (0.0017).

Discussion

The two tests based on observed times in stage operate at close to the nominal level (Fig. 1) and have good power against a range of alternatives (Fig. 2, *Appendix* 4). The logistic phenology model, however, has a level that can differ markedly from the nominal value (Fig. 1), and, even after calibration, it is less powerful than the two other approaches (Fig. 2).

The interpretation of the parameters in the logistic phenology model is difficult. Heuristically, the relationships in (1) imply that "Individual i will be in stage k at time t if the random variable $S_i(t)$ is in the k^{th} of the intervals defined by $-\infty \leq \beta_1 \leq \ldots \leq \beta_{K-1} \leq \infty$." The parameter β_0 is related to the variance of the stochastic process that determines the intervals in (1): smaller values of β_0 indicate that the stages do not overlap each other very much in the population, whereas larger values of β_0 indicate that at any given time there may be individuals in several different stages.

The models based on observed time in stage are more easily understood. The estimates $\bar{X}_k$ (Equation 2) provide a simple summary of the times at which stage k is observed in a particular population. A t-test can be used to compare these summaries between populations for each stage, and the stage-specific results can be combined into a global P value for testing whether the two populations have the same phenology. Alternatively, one can fit linear regressions of observed times versus stage for each population, and then compare the lines between populations. The regression-based method was somewhat more powerful against the alternatives that we considered, but the t-test approach has the advantage that it provides stage-specific P values.

The power and calibration of all of these methods will depend on the sampling frequency. More frequent sampling will increase the (calibrated) power, but will also slightly increase the probability of a Type I error, because the correlation of the observations will increase with the chance that an individual is sampled more than once. In practice, sampling frequency is likely constrained by time and budgetary considerations. We therefore recommend sampling as often as is feasible, with the understanding that the improvement in power must be balanced against the risk of obtaining a highly correlated sample if some individuals are encountered repeatedly.

In summary, the methods based on observed time in stage provide a simple way of comparing phenology between sites that is easily implemented and interpreted. In addition, simulations show that these methods have much better operating characteristics than the logistic phenology model, a more complicated approach that has been used in past studies of phenology.

Acknowledgments

We thank two anonymous referees for constructive criticisms that helped improve the manuscript.

References Cited

Candy, S. G. 2003. Predicting time to peak occurrence of insect life-stages using regression models calibrated from stage-frequency data and ancillary stage-mortality data. Agric. For. Entomol. 5: 43–49.

Dennis, B., W. P. Kemp, and R. C. Beckwith. 1986. Stochastic model of insect phenology: estimation and testing. Environ. Entomol. 15: 540–546.

Fisher, R. A. 1970. Statistical methods for research workers. Hafner Publishing Company, Darien, CT.

Hodgson, J. A., C. D. Thomas, T. H. Oliver, B. J. Anderson, T. M. Brereton, and E. E. Crone. 2011. Predicting insect phenology across space and time. Glob. Change Biol. 17: 1289–1300.

Kemp, W. P., B. Dennis, and R. C. Beckwith. 1986. Stochastic phenology model for the western spruce budworm (Lepidoptera: Tortricidae). Environ. Entomol. 15: 547–554.

Manel, S., and D. Debouzie. 1997. Modeling insect development time of two or more larval stages in the field under variable temperatures. Environ. Entomol. 26: 163–169.

McEvoy, P. B., C. Cox, and E. Coombs. 1991. Successful biological control of ragwort, *Senecio jacobaea*, by introduced insects in Oregon. Ecol. Appl. 1: 430–442.

Morellato, L.P.C., L. F. Alberti, and I. L. Hudson. 2010. Applications of circular statistics in plant phenology: a case studies approach, pp. 339–359. *In* I. L. Hudson and M. R. Keatley (eds.), Phenological research: methods for environmental and climate change analysis. Springer, Dordrecht, The Netherlands.

Munholland, P. L., and J. D. Kalbfleisch. 1991. A semi-Markov model for insect life-history data. Biometrics 47: 1117–1126.

Osawa, A., C. A. Shoemaker, and J. R. Stedinger. 1983. A stochastic model of balsam fir bud phenology utilizing maximum likelihood parameter estimation. For. Sci. 21: 478–490.

Petitt, F. L., J. C. Allen, and C. S. Barfield. 1991. Degree-day model for vegetable leafminer (Diptera, Agromyzidae) phenology. Environ. Entomol. 20: 1134–1140.

R Development Core Team. 2011. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (http://www.R-project.org).

Volney, W.J.A., and H. F. Cerezke. 1992. The phenology of white spruce and the spruce budworm in northern Alberta. Can. J. For. Res. 22: 198–205.

Weber, J. D., W.J.A. Volney, and J. R. Spence. 1999. Intrinsic development rate of spruce budworm (Lepidoptera: Tortricidae) across a gradient of latitude. Environ. Entomol. 28: 224–232.

Welch, B. L. 1951. On the comparison of several mean values: an alternative approach. Biometrika 38: 330–336.

Received 6 October 2011; accepted 27 January 2012.

Appendix

Appendix 1: **Hypothesis Testing with the Logistic model.** As described by Dennis et al. (1986), the maximum-likelihood estimates of the model parameters have an asymptotically normal distribution. Letting $\hat{\beta}_{km}$, $m = 1$ or 2, be the estimate of β_k for population m, single-parameter tests are based on the statistic

$$z_k = \frac{\hat{\beta}_{k1} - \hat{\beta}_{k2}}{\sqrt{\hat{\sigma}^2_{k1} + \hat{\sigma}^2_{k2}}}$$

for $k = 0, \ldots, K - 1$, where $\hat{\sigma}_{k1}$ and $\hat{\sigma}_{k2}$ are the estimated standard errors of the estimates $\hat{\beta}_{k1}$ and $\hat{\beta}_{k2}$, respectively. Asymptotically, z_k has a standard normal distribution under the null hypothesis, so two-sided P values may be obtained as $p_k = 2 \min(\Phi(z_k), 1 - \Phi(z_k))$.

Dennis et al. (1986) present a global test based on the asymptotic multivariate normal distribution of the vector $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{k-1})$. Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ be the vectors for populations 1 and 2, respectively, and let $\mathbf{S}_1$ and $\mathbf{S}_2$ be the corresponding estimated covariance matrices. Then a global test statistic given by

$$T = (\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2)^{\mathrm{T}}(\mathbf{S}_1 + \mathbf{S}_2)^{-1}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2)$$

has a χ^2_K asymptotic distribution under the null hypothesis. Therefore, a p-value for the global test of the hypothesis that the two populations have identical phenology across all stages may be obtained by comparing the statistic T to a χ^2_K distribution.

Appendix 2: **A Simple Example with R Code.** Suppose we have random samples of larvae from two locations, yielding Table A1. Applied to these data, the *t*-test method gives stage-specific P values (egg through L5) of 0.013, 0.82, 0.26, 0.57, 0.26 and 0.22, and a global P value of 0.099. The regression-based method gives a P value of 0.16. R code (R Development Core Team 2011) that can be used to implement our methods is given below.

Appendix 1 Table 1. Hypothetical phenology data from two sites

Site	Day	Egg	L1	L2	L3	L4	L5
1	194	20	10	0	0	0	0
1	201	3	5	10	1	0	0
1	208	0	0	10	8	2	0
1	215	0	0	0	10	15	1
1	222	0	0	0	0	5	10
2	194	15	12	0	0	0	0
2	201	12	5	8	0	0	0
2	208	0	0	5	15	0	0
2	215	0	0	4	9	10	2
2	222	0	0	0	0	5	1

R code:

```
phen.test <- function(input="example.
 txt", method=1, prnt=T) {

d <- read.table(input, header=T)
sites <- sort(unique(d$Site))
n.site <- 2

n.stage <- 6
counts <- cbind(d$Egg, d$L1, d$L2, d$L3,
 d$L4, d$L5)

pvals <- rep(NA, n.stage)
day <- site <- stage <- NULL
for (i in 1:n.stage) {
for (j in 1:n.site) {
ind <- d$Site==sites[j]
day.new <- rep(d$Day[ind], counts
 [ind,i])
day <- c(day, day.new)
site <- c(site, rep(sites[j],
 length(day.new)))
stage <- c(stage, rep(i, length(day.
 new)))
}
}

if (method==1) {
for (i in 1:n.stage) {
ind1 <- site==sites[1] & stage==i
ind2 <- site==sites[2] & stage==i
tmp <- t.test(day[ind1], day[ind2], var.
 equal=F)
pvals[i] <- tmp$p.value
}
combQ <- −2 * sum(log(pvals))
Pval <- 1 - pchisq(combQ, df=2*length
 (pvals))
}

if (method==2) {
mod1 <- lm(day ~ stage)
mod2 <- lm(day ~ stage*factor(site))
tmp <- anova(mod1, mod2)
Pval <- tmp$Pr[2]
}

if (prnt==T) {
cat(paste("\nStage-specific p-values:
 \n"))
print(signif(pvals,4))
cat(paste("\nGlobal P-value =",signif
 (Pval,4),"\n"))
}

invisible(list(pvals=pvals, Pval=Pval))
}
```

Appendix 3: **Bootstrap Estimation of Variances Under Cluster Sampling.** A single bootstrap sample is obtained as follows. For each combination of location and sampling date, select with replacement a sample of plants having the same size as the original sample for that location and date. Enumerate the individuals on each plant, keeping track of the stage and time (or degree days) for each individual. The subsequent steps differ between analysis methods.

1. **The t-test Approach.** When the b^{th} bootstrap sample is complete, compute $(D_k)_b$ as in Equations 4 and 5, for $k = 1, \ldots, K$.

From the B bootstrap samples, compute for each k

$$\bar{D}_k = \sum_{b=1}^{B} (D_k)_b / B, \quad \text{and}$$

$$z_k = \frac{\bar{D}_k}{\sqrt{\sum_{b=1}^{B} [(D_k)_b - \bar{D}_k]^2 / (B-1)}}.$$

The two-sided P value for stage k is then $p_k = 2 \min(\phi(z_k), 1 - \phi(z_k))$, where $\phi(\cdot)$ is the cumulative distribution function of a standard normal random variable.

2. The Regression Approach. When the b^{th} bootstrap sample is complete, regress observed time vs. stage (coded as 1, 2, . . ., K) *separately* for the two populations. Let $(\hat{\beta}_{0\mathrm{j}})_b$ and $(\hat{\beta}_{1\mathrm{j}})_b$ be the estimated intercept and slope, respectively, for sample j ($j = 1, 2$) in bootstrap sample b. Calculate

$$(d_0)_b = (\hat{\beta}_{01})_b - (\hat{\beta}_{02})_b$$

$$(d_1)_b = (\hat{\beta}_{11})_b - (\hat{\beta}_{12})_b.$$

Let $\mathbf{d}_0 = [(d_0)_1, (d_0)_2, \ldots (d_0)_{\mathrm{B}}]'$, $\mathbf{d}_1 = [(d_1)_1, (d_1)_2, \ldots (d_1)_B]'$, $\bar{d}_0 = \Sigma_{\mathrm{i}=1}^{\mathrm{B}} (\mathrm{d}_0)_{\mathrm{i}}/\mathrm{B}$, and $\bar{d}_1 = \Sigma_{i=1}^{B} (d_1)_i / B$. Compute

$$Q = (\bar{d}_0 \quad \bar{d}_1) \begin{pmatrix} \widehat{\mathrm{Var}}\,(\mathbf{d}_0) & \widehat{\mathrm{Cov}}\,(\mathbf{d}_0, \mathbf{d}_1) \\ \widehat{\mathrm{Cov}}\,(\mathbf{d}_0, \mathbf{d}_1) & \widehat{\mathrm{Var}}\,(\mathbf{d}_1) \end{pmatrix} \begin{pmatrix} \bar{d}_0 \\ \bar{d}_1 \end{pmatrix},$$

Where $\widehat{\mathrm{Var}}\,(\cdot)$and $\widehat{\mathrm{Cov}}\,(\cdot)$ indicate sample variance and covariance, respectively. A p-value for testing the hypothesis that the two regression lines are the same can then be obtained as $P(\chi_2^2 > Q)$, where χ_2^2 is a chi-square random variable with 2 degrees of freedom.

***Appendix* 4: Further Simulation Results.** Here we present comparisons of the performances of the t-test and regression-based approaches in some additional circumstances. Simulations of two populations were performed as described in the text, under *Comparison of the Methods.*

Nonmonotonic Stage Durations. One population had stage durations, in days, of (12, 8, 12, 8, 12, 8), and the other population had durations of (12 + δ, 8 + δ, 12 + δ, 8 + δ, 12 + δ, 8 + δ), with δ ranging from 0 to 0.5 d. The following table shows the proportions of 500 simulations in which the null hypothesis of identical phenology was rejected.

Appendix 4 Table 1. Rejection frequencies

Method	Value of δ					
	0	0.1	0.2	0.3	0.4	0.5
t-test	0.042	0.080	0.178	0.348	0.580	0.798
Regression	0.062	0.096	0.290	0.522	0.808	0.966

Populations Differing With Respect to Only One Stage. One population had stage durations, in days, of (10, 10, 10, 10, 10, 10), and the other population had durations of (10, 10, 10, 10 + δ, 10, 10), with δ ranging from 0 to 2.5 d. The following table shows the proportions of 500 simulations in which the null hypothesis of identical phenology was rejected.

Appendix 4 Table 2. Rejection frequencies

Method	Value of δ					
	0	0.1	0.2	0.3	0.4	0.5
t-test	0.040	0.090	0.156	0.264	0.430	0.660
Regression	0.082	0.104	0.256	0.430	0.684	0.884