HATCHMARK

A statistical tool to conduct an *a priori* analysis of the precision and bias of estimators that discriminate between hatchery- and wild-origin spawners using observed visible marks

Richard A. Hinrichsen and Rishi Sharma

October 18, 2013

SUMMARY

A reliable monitoring program to discriminate between hatchery- and wild-origin spawners is needed to determine the viability of listed populations, set harvest limits, and understand genetic risks. A web tool (located at www.onefishtwofish.net) demonstrates the effects of sample rate, visible marking (VM) fraction, and population size on the statistical precision of these spawners estimates. Precision[1] is depicted by standard errors and coefficients of variation. The assumptions are described in Table 1.

INTRODUCTION

Aside from the biological risks of hatchery-origin salmon spawning in the wild, the statistical difficulties of estimating wild-origin and hatchery-origin spawners presented by presence of non-VM hatchery fish have been recognized for over thirty years (Hankin 1982). Reliable estimation of the hatchery proportion of spawners for Columbia Basin spawning areas is needed for population viability analysis which is usually focused on the viability and trend of the wild-origin component of the population (McClure et al. 2003). It is broadly recognized that interbreeding of hatchery-origin with the wild-origin spawners may reduce the genetic fitness of subsequent generations (Waples 1991). Furthermore salmon runs that are boosted by hatchery-origin salmon may result in harvest rates that are greater than what the wild population can bear (Paulik et al. 1967; Wright 1981). These issues point to an important need to reliably enumerate wild- and hatchery-origin spawners. To allow distinction between wild-origin and hatchery-origin salmon in the Columbia Basin, the U.S. Congress presently requires the US Fish and Wildlife Service to VM all hatchery production intended for harvest. Although visible marking of hatchery releases is a widespread practice among hatchery operators in the Columbia River basin, non-VM procedures are sometimes substituted for or added to

---

[1] We use the term precision to describe the spread of the sampling distribution of an estimator. Precision increases as standard error and CV decrease.

VM. Currently not all hatchery fish are marked either visibly or nonvisibly and this presents statistical challenges to estimating hatchery-origin spawners (Hankin 1982).

Needed is an assessment of the accuracies of hatchery- and wild-origin spawners and how they depend on VM fraction and sample rate. We developed a statistical approach to evaluating the precision of estimates based on the size of the sampled population, the sample rate, and the VM fraction of the hatchery group. In this exercise, we assume that there are two groups of spawners sampled (wild- and hatchery-origin spawners) and the objective is to estimate the abundances of these two groups and the proportion of hatchery-origin spawners. Furthermore we assume that there is a single source hatchery contributing spawners to the spawning grounds. In the case where there are multiple source hatcheries, the tool HATCHMARK2, based on Hinrichsen et al. (2012), should be employed.

METHODS

To determine the effect of sample rate and VM fraction on the precision of estimates of wild-origin and hatchery-origin spawners and hatchery-origin spawners as a fraction of the total spawners, we developed method of moments estimators (MMEs). Variances for the estimators were evaluated using Classical theory and Monte Carlo simulations. Classical theory was used to derive variance formulas for these estimators. Monte Carlo variance does not rely on asymptotic theory and should be used when sample size low, making theoretical variances suspect.

The statistical methods were built on several assumptions that were needed to define the probability distributions of the hatchery- and wild-origin spawner counts (Table 1). These assumptions allowed us to treat as a Bernoulli trial whether a spawner was observed. The probability of "success" of each of these trials was equal to the sample rate. Furthermore, we could treat as a Bernoulli trail whether a hatchery fish had a mark. The probability of "success" of each of these trials was equal to the VM fraction. Together these assumptions allowed us to define the joint probability distribution of spawners to represent cell counts in a product of multinomial distributions. Hankin (1982) used similar assumptions in his estimation of hatchery-origin spawner proportions but he assumed a fixed sample size instead of a fixed sample rate. The R-code used to generate both the Monte Carlo estimates of precision and theoretical estimate of precision is given in Appendix A.

Table 1.—Assumptions[1]

A1. VM fraction and sample rate are known with certainty.
A2. All hatchery-origin spawners originate from hatcheries that use the same VM fraction.
A3. Every individual spawner has the same probability of being sampled.
A4. Every individual hatchery-origin spawner has the same probability of having a VM.
A5. Whether any individual is sampled has no effect on the probability that another individual is sampled.
A6. Whether any individual hatchery-origin spawners is observed to have a VM has no effect on the probability that another individual will have a VM.

[1] Assumptions may need to be modified to more accurately depict a particular spawner monitoring program. For example, one important assumption is that sample rate is known with certainty. To model the effects of an uncertain sample rate a probability distribution would need to be specified and included in the simulations and equations.

Joint probability distribution.—We first derive an expression for the joint probability distribution of sampled VM and non-VM hatchery-origin spawners and wild-origin spawners. Let N_{HOS} represent the number of hatchery-origin spawners and let N_{NOS} represent the number of wild-origin spawners. The number of VM and non-VM fish sampled may be modeled using an underlying multnomial distribution with the following cells (associated cell counts are given in parentheses): sampled and VM hatchery spawners (x_1), sampled and non-VM hatchery spawners (x_2), sampled wild-origin spawners (x_3), and the unsampled hatchery-origin spawner population ($N_{HOS} - x_1 - x_2$), and the unsampled wild-origin spawner population $\left(N_{NOS} - x_3\right)$. Not all of these cell counts can be observed, but these are the random variables that underlie the observed spawners and are used to derive the variance formulas. The joint probability distribution of these random variables is

$$f(x_1, x_2, x_3; \lambda, \theta) = \frac{N_{HOS}!}{x_1! x_2! (N_{HOS} - x_1 - x_2)!} \left(\lambda\theta\right)^{x_1} \left(\theta(1-\lambda)\right)^{x_2} (1-\theta)^{N_{HOS} - x_1 - x_2} \tag{1}$$
$$\times \frac{N_{NOS}!}{x_3! (N_{NOS} - x_3)!} \theta^{x_3} (1-\theta)^{N_{NOS} - x_3}$$

which is a product of trinomial probability distribution with three cell counts x_1, x_2, and $N_{HOS} - x_1 - x_2$, and a binomial distribution with two cell counts x_3 and $N_{NOS} - x_3$. The parameter λ represents the fraction of hatchery population that is marked, and θ represents the fraction of the total spawners that is sampled. This formulation differs from Hankin (1982) who used a trinomial distribution alone as the basis for estimating the proportion of hatchery-origin spawners and also assumed a fixed sample size rather than

a sample size that follows a binomial distribution with a fixed sample rate ("success") probability).

Estimators.—Spawner estimators were derived using the method of moments, which is sometimes used to estimate the *n* parameter in a binomial distribution (Blumenthal and Dahiya 1981). Using the method of moments, the non-VM and VM spawner observations (E_M and E_U, respectively) were set equal to their expected values as determined by the joint probability distribution in Equation 1 as follows:

$$E_M = x_1 = N_{HOS}\lambda\theta \quad \text{and}$$

(2)

$$E_U = x_2 + x_3 = N_{HOS}\theta(1-\lambda) + N_{NOS}\theta$$

(3)

Solving these equations for the spawners yielded the following MMEs:

$$\hat{N}_{HOS} = \frac{E_M}{\lambda\theta} \quad \text{and}$$

(4)

$$\hat{N}_{NOS} = \frac{E_U}{\theta} + \frac{E_M}{\theta} - \frac{E_M}{\lambda\theta},$$

(5)

where $\hat{N}_{HOS}$ denotes hatchery-origin spawners estimator and $\hat{N}_{NOS}$ denotes wild-origin spawners estimator. These spawner estimators were then used to define an estimator of the proportion of hatchery-origin spawners:

$$\hat{p}_{HOS} = \frac{\hat{N}_{HOS}}{\hat{N}_{HOS} + \hat{N}_{NOS}}.$$

(6)

Analytic expressions for variance.—Using the known variances and covariances of multinomial distribution cell counts we derived the variance for the spawner estimators:

$$\text{var}(\hat{N}_{HOS}) = \text{var}\left(\frac{x_1}{\lambda\theta}\right) = \frac{\text{var}(x_1)}{(\lambda\theta)^2} = \frac{N_{HOS}\lambda\theta(1-\lambda\theta)}{(\lambda\theta)^2} = \frac{N_{HOS}(1-\lambda\theta)}{\lambda\theta} \quad \text{and}$$

(7)

$$\text{var}\left(\hat{N}_{NOS}\right) = \text{var}\left[\frac{\lambda(x_2 + x_3) - (1-\lambda)x_1}{\lambda\theta}\right]$$

(8)

$$= \frac{1}{(\lambda\theta)^2}\left\{\lambda^2\left[\text{var}(x_2) + \text{var}(x_3)\right] + (1-\lambda)^2\,\text{var}(x_1) - \lambda(1-\lambda)\,\text{cov}(x_1, x_2)\right\}$$

$$= N_{NOS}\frac{(1-\theta)}{\theta} + N_{HOS}\frac{(1-\lambda)}{\lambda\theta}.$$

The above derivation of $\text{var}\left(\hat{N}_{NOS}\right)$ relied on the fact that $\text{cov}(x_2, x_3) = 0$ and $\text{cov}(x_1, x_3) = 0$ and $\text{cov}(x_1, x_2) = -N_{HOS}\lambda\theta^2(1-\lambda)$ (using the fact that for a multinomial distribution, the covariance for two cell counts equals minus the product of the number of trials and the two cell probabilities). The covariance of the MMEs is given by

$$\text{cov}\left(\hat{N}_{HOS}, \hat{N}_{NOS}\right) = \text{cov}\left[\frac{x_1}{\lambda\theta}, \frac{\lambda(x_2 + x_3) - (1-\lambda)x_1}{\lambda\theta}\right]$$

(9)

$$= -\frac{N_{HOS}(1-\lambda)}{\lambda\theta}.$$

Notice that when the abundance of hatchery-origin spawners is not zero, the covariance term is zero only when the marking fraction is one.

Treating $\hat{p}_{HOS}$ as a function of the spawner estimates $\hat{N}_{HOS}$ and $\hat{N}_{NOS}$ and expanding the function in a Taylor series up to the first order term about the true number of spawners yielded

$$\hat{p}_{HOS} - p_{HOS} \cong \left(\nabla p_{HOS}\right)'\begin{bmatrix}\hat{N}_{HOS} - N_{HOS}\\\hat{N}_{NOS} - N_{NOS}\end{bmatrix},$$

(10)

where $\nabla p_{HOS} = \begin{bmatrix}N_{NOS} & -N_{HOS}\end{bmatrix}'/N^2$ is the gradient (2-vector of partial derivatives of p_{HOS} with respect to N_{HOS} and N_{NOS}.

Taking the variance of both sides of this equation yielded the following formula:

$$\text{var}\left(\hat{p}_{HOS}\right) \cong \left(\nabla p_{HOS}\right)'\,\text{var}\begin{bmatrix}\hat{N}_{HOS}\\\hat{N}_{NOS}\end{bmatrix}\nabla p_{HOS}$$

(11)

$$= \frac{N_{HOS}}{N^4 \lambda \theta} \left[N_{NOS}^2 (1 - \lambda\theta) + N_{HOS} N_{NOS} (2 - \lambda - \theta\lambda) + N_{HOS}^2 (1 - \lambda) \right].$$

This variance formula for $\text{var}(\hat{p}_{HOS})$ is approximate because the higher order terms of the Taylor series expansion were ignored. Such an approximation is best when the population size (N) is large. For small population sizes, Monte Carlo simulation is used to estimate $\text{var}(\hat{p}_{HOS})$. Armed with these formulas for variance, the standard error (square root of the variance) and CV (standard error divided by parameter value) are easily calculated.

That the estimators of N_{HOS} and N_{NOS} are unbiased[2] may be readily seen. Using the fact that the mean of a cell count is equal to the likelihood of a "success" in that cell times the number of trials, we can write

$$E(\hat{N}_{HOS}) = E\left(\frac{x_1}{\lambda\theta} \right) = \frac{E(x_1)}{(\lambda\theta)} = \frac{N_{HOS}\lambda\theta}{(\lambda\theta)} = N_{HOS} \quad \text{and} \tag{12}$$

$$E(\hat{N}_{NOS}) = E\left[\frac{\lambda(x_2 + x_3) - (1 - \lambda)x_1}{\lambda\theta} \right] \tag{13}$$

$$= \frac{1}{(\lambda\theta)} \left[\lambda(E(x_2) + E(x_3)) - (1 - \lambda)E(x_1) \right]$$

$$= \frac{1}{(\lambda\theta)} \left[\lambda(N_{HOS}\theta(1 - \lambda) + N_{NOS}\theta) - (1 - \lambda)N_{HOS}\theta\lambda \right]$$

$$= \left[(N_{HOS}(1 - \lambda) + N_{NOS}) - (1 - \lambda)N_{HOS} \right] = N_{NOS} .$$

The Taylor series approximation to $\hat{p}_{HOS}$ given in equation (10) suggests that $\hat{p}_{HOS}$ is approximately unbiased because the estimators $\hat{N}_{HOS}$ and $\hat{N}_{NOS}$ are unbiased. Monte Carlo simulations using the methods in Appendix A indicated low bias of $\hat{p}_{HOS}$. This was demonstrated comparing the Monte Carlo replications of $\hat{p}_{HOS}$ with the known value of p_{HOS}.

Monte Carlo simulations.—To evaluate the precision of these estimators without relying on asymptotic theory, Monte Carlo simulation was used (Appendix A). The simulations proceeded by first assuming some true values of the hatchery-origin and

---

[2] The term bias is used here in its statistical sense and refers to the expected value of the estimator minus the true value of the estimated parameter. Other forms of bias may be important when the assumptions of the model are not met. For example, if the assumed marking fraction is inaccurate perhaps due to straying (Quinn and Fresh 1984, Quinn et al. 1991, Pascual and Quinn 1994).

wild-origin spawners. The estimation process was then simulated again and again as draws from the joint probability distribution given in equation (1) to yield many Monte Carlo replications of the estimates. The variance of each estimator was then estimated as sample variance of its Monte Carlo replications. Because at low sample sizes, bias may be an issue for the estimator $\hat{p}_{HOS}$, we estimated Monte Carlo relative bias as the difference between the mean of the Monte Carlo replications minus the known value of p_{HOS}, all divided by the known value of p_{HOS}.

Application

To demonstrate the importance of design variables on the precision of the estimate of p_{HOS}, we run a sensitivity analysis showing how the CV of the estimate changes with VM fraction, λ, and sample rate, θ. Such an analysis can demonstrate what values of VM fraction and sample rate will deliver a targeted level of precision. In this analysis, we vary the VM fraction over the values $\lambda = 0.1, 0.2, ..., 0.9$ and the values of the sample rate over $\theta = 0.1, 0.2, ..., 0.9$ and calculate the corresponding value of CV for each VM fraction-sample rate pair drawn from these values, giving a total of 81 CV values. The remaining input values for the statistical procedure are $N_{HOS} = 100$ and $N_{NOS} = 900$, which gives $p_{HOS} = 0.10$. We then construct of contour plot of the CVs to show how CV varies with VM fraction and sample rate. We also construct a plot of CV versus the expected number of VM recoveries $Np_{HOS}\theta\lambda$ to reveal this important relationship.

A contour plot of the CV values derived from the sensitivity analysis is given in Figure 1. There are two important observations to make about this plot. First, the contour plot reveals symmetry about the line $\theta = \lambda$: exchanging the values of VM fraction and sample rate does not change the value of the CV. Second, diminishing returns is evident in the CV. When values of θ and λ are small, increasing either yields are large increase in CV. When values of θ and λ are large, increasing either θ or λ yields a small increase in CV. The reason for symmetry in the contour plot becomes apparent in the plot of CV versus expected number of mark recoveries (Figure 2). Note that CV decreases with expected number of mark recoveries, and since expected number of marked recoveries depends only on the product of VM fraction and sample rate, then these values can be exchanged without changing the expected number of marked recoveries. Figure 2 also reveals diminishing returns. When expected number of marked recoveries is small, increasing expected marked recoveries yields big gains in precision. However, when expected number of marked recoveries is large, increasing expected marked recoveries yields very little improvement in precision.
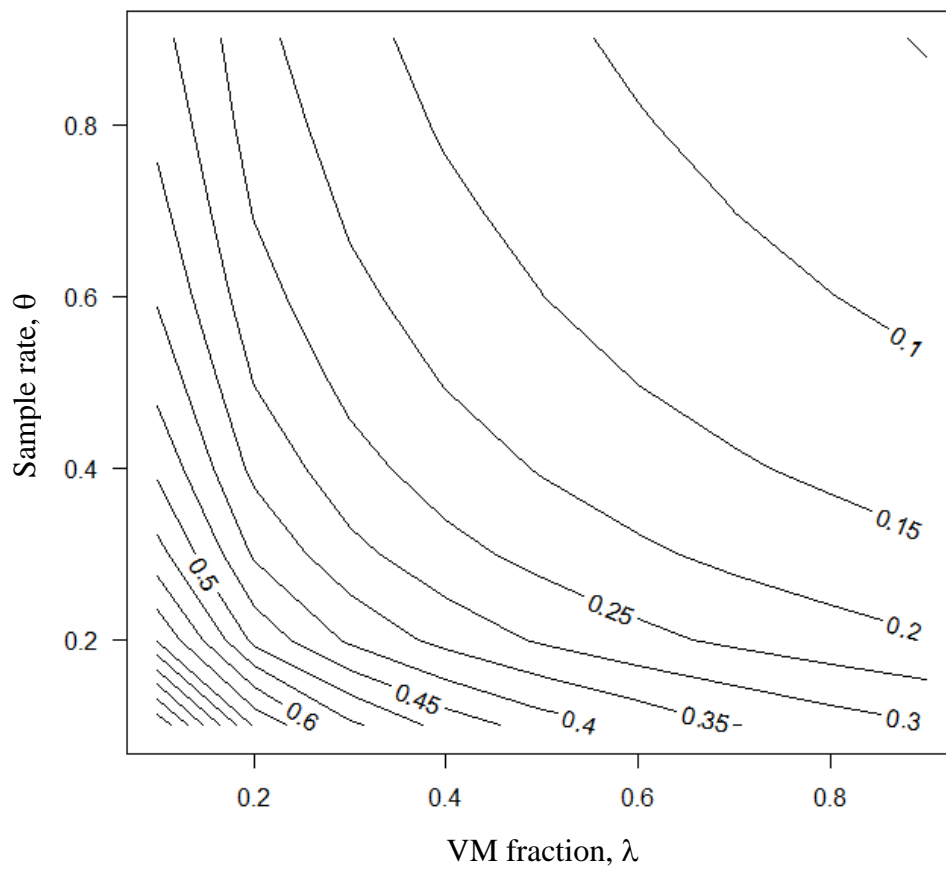
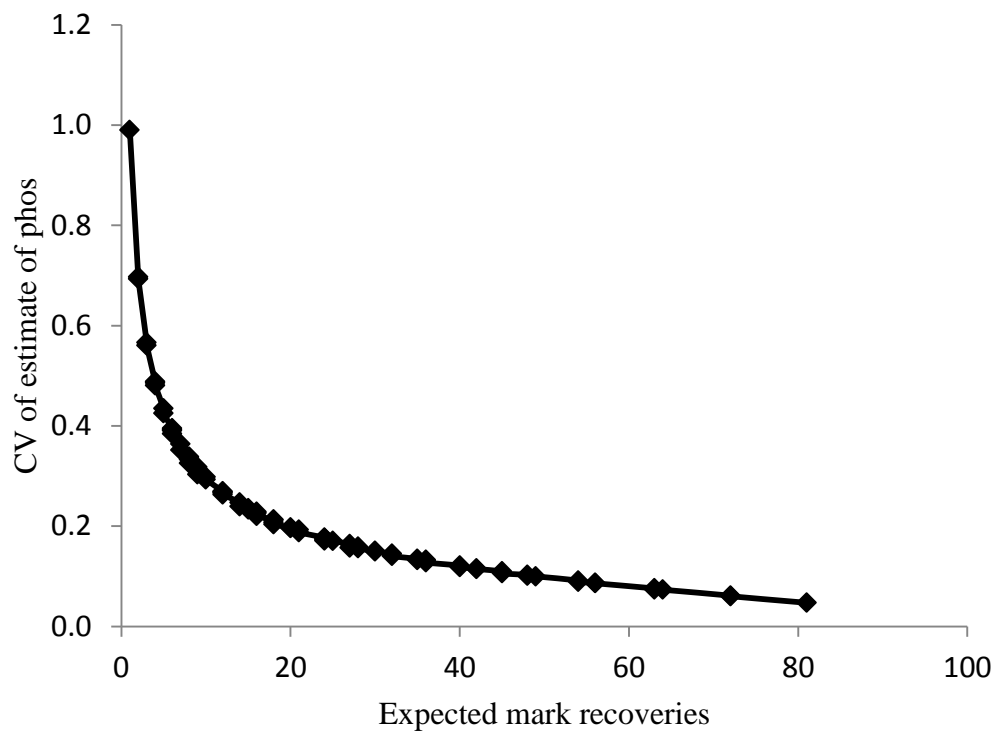Figure 1.—Contour plot of CV vs. VM fraction and sample rate.

Figure 2.—Plot of CV vs. expected number of mark recoveries.

ACKNOWLEDGEMENTS

REFERENCES

Blumenthal, S. and R.C. Dahiya. 1981. Estimating the binomial parameter n. Journal of the American Statistical Association 76 (No. 376):903-909.

Hankin, D.G. 1982. Estimating escapement of pacific salmon: marking practices to discriminate wild and hatchery fish. Transactions of the American Fisheries Society 111:286-298.

Hinrichsen, R.A., R. Sharma, T.R. Fisher. 2012. Precision and accuracy of estimators of the proportion of hatchery-origin spawners. Transactions of the American Fisheries Society 142:437-454.

McClure, M.M., E.E. Holmes, B.L. Sanderson, and C.E. Jordan. 2003. A large-scale multispecial status assessment: anadromous salmonids in the Columbia River basin. Ecological Applications 13(4):964-989.

Pascual, M.A. and T.P. Quinn. 1994. Geographical patterns of straying if fall chinook salmon (Oncorhynchus tschawytcha) from Columbia River (USA) hatcheries. Aquaculture and Fisheries Management 25 (Supplement 2):17-30

Paulik, G.J., A.S. Hourston, and P.A. Larkin. 1967. Exploitation of multiple stocks by a common fishery. Journal of the Fisheries Research Board of Canada 24:2527-2537.

Quinn, T.P., and K. Fresh. 1984. Homing and straying in chinook salmon (Oncorhynchus tschawytcha) from Cowlitz River Hatchery, Washington. Canadian Journal of Fisheries and Aquatic Sciences 41:1078-1082.

Quinn, T.P., R.S. Nemeth, and D.O. McIsaac. 1991. Pattersn of homing and straying by fall chinook salmon in the lower Columbia River. Transactions of the American Fisheries Society 120:150-156.

Waples, R.S. 1991. Genetic interactions between hatchery and wild salmonids: lessons from the Pacific Northwest. Canadian Journal of Fisheries and Aquatic Sciences 48(Supplement1):124-133.

Wright, S. 1981 Contemporary Pacific salmon fisheries management. North American Journal of Fisheries Management 1:29-40.

APPENDIX A: R-CODE

Table A.1.—R-code used to estimate hatchery-origin and wild-origin spawners.

```
#Program to calculate properties of spawner estimates
#when not all hatchery-origin spawners are visibly marked (VM)
#based on a known sampling rate and VM fraction
#AUTHOR: Richard A. Hinrichsen
#CONTACT: rich@hinrichsenenvironmental.com
#DATE MODIFIED: 10/18/2013

#Variables and parameters used in the analysis
#inputs
#MONTE = logical variable when TRUE, Monte Carlo simulations are used
#Nsims = total number of bootstrap replications
#Nhos = true number of hatchery-origin spawners
#Nnos = true number of wild-origin spawners
#theta = sample rate
#lambda = VM fraction
#Em = number of sampled spawners that are VM
#Eu = number sampled spawners that are non-VM
#
#
#intermediate variables
#phos = fraction of spawnerst that is of hatchery origin
#Ehatchsampled = Replications of number of hatchery-origin spawners that are sampled
#Enatsampled = Replications of number of wild-origin spawners that are sampled
#Em = Relications of number of VM spawners
#Eu = Replications of number of non-VM spawners
#Nhoshat = Replications estimate of Nhos
#Nnoshat = Replications of estimate of Nnos
#
#output variables
#phos (true value) calculated from Nhos and Nnos
#SE.***** = standard error (SE)
#CV.*** = Coefficient of variation
#BIAS.phoshat (relative bias of the estimator phoshat)


main<-function(MONTE=FALSE,Nsims=NA,Nhos=100,Nnos=100,theta=.25,lambda=.75){
check.inputs(MONTE,Nsims,Nhos,Nnos,theta,lambda)
 if(MONTE){res<-phos.estimates1(Nsims=Nsims,Nhos=Nhos,Nnos=Nnos,theta=theta,lambda=lambda)}
 else{res<-phos.estimates2(Nhos=Nhos,Nnos=Nnos,theta=theta,lambda=lambda)}
return(res)
}

#uses Monte Carlo simulation
phos.estimates1<-function(Nsims=10000,Nhos=100,Nnos=100,theta=0.25,lambda=0.75)
{
phos<-Nhos/(Nhos+Nnos)
#generate synthetic data sets
Ehatchsampled <-rbinom(Nsims,size=Nhos,prob=theta)
Enatsampled <-rbinom(Nsims,size=Nnos,prob=theta)
Em<-rep(NA,Nsims)
for(ii in 1:Nsims){
```

```
Em[ii]<-rbinom(1,size=Ehatchsampled[ii],prob=lambda)
}
Eu<-Ehatchsampled-Em+Enatsampled

#Replications of estimates
Nhoshat<-Em*(1/theta)*(1/lambda)
Nnoshat<-Eu*(1/theta)+Em*(1/theta)-Em*(1/lambda)*(1/theta)
phoshat<-Nhoshat/(Nhoshat+Nnoshat)
#properties of estimators
SE.Nhoshat<-sqrt(var(Nhoshat,na.rm=T))
SE.Nnoshat<-sqrt(var(Nnoshat,na.rm=T))
SE.phoshat<-sqrt(var(phoshat,na.rm=T))
CV.Nhoshat<-SE.Nhoshat/Nhos
CV.Nnoshat<-SE.Nnoshat/Nnos
CV.phoshat<-SE.phoshat/phos
BIAS.phoshat<-(mean(phoshat, na.rm=T)-phos)/phos
myres<-list(MONTE=TRUE,Nsims=Nsims,Nhos=Nhos,Nnos=Nnos,theta=theta,lambda=lambda,phos=phos,
SE.Nhoshat=SE.Nhoshat,CV.Nhoshat=CV.Nhoshat,
SE.Nnoshat=SE.Nnoshat,CV.Nnoshat=CV.Nnoshat,
SE.phoshat=SE.phoshat,CV.phoshat=CV.phoshat, BIAS.phoshat=BIAS.phoshat)
return(myres)
}
#Uses theoretical variance calculations
#the estimate of standard error of phos is unreliable when Nhos and Nnos are small
#because it is based on a first order Taylor series expansion about the spawner estimates.
#standard errors for spawner estimates are exact
phos.estimates2<-function(Nhos=100,Nnos=100,theta=0.25,lambda=0.75)
{
N<-Nhos+Nnos
phos<-Nhos/(Nhos+Nnos)
#properties of estimators
var.Nhoshat<-Nhos*(1-lambda*theta)/(lambda*theta)
var.Nnoshat<-Nnos*(1-theta)/theta+Nhos*(1-lambda)/(theta*lambda)
var.phoshat<-(Nhos/(lambda*theta*N^4))*(Nnos*Nnos*(1-lambda*theta)+Nhos*Nnos*(2-lambda-
lambda*theta)+Nhos*Nhos*(1-lambda))
SE.Nhoshat<-sqrt(var.Nhoshat)
SE.Nnoshat<-sqrt(var.Nnoshat)
SE.phoshat<-sqrt(var.phoshat)
CV.Nhoshat<-SE.Nhoshat/Nhos
CV.Nnoshat<-SE.Nnoshat/Nnos
CV.phoshat<-SE.phoshat/phos
myres<-list(MONTE=FALSE,Nsims=NA,Nhos=Nhos,Nnos=Nnos,theta=theta,lambda=lambda,phos=phos,
SE.Nhoshat=SE.Nhoshat,CV.Nhoshat=CV.Nhoshat,
SE.Nnoshat=SE.Nnoshat,CV.Nnoshat=CV.Nnoshat,
SE.phoshat=SE.phoshat,CV.phoshat=CV.phoshat, BIAS.phoshat=NA)
return(myres)
}

#make sure inputs make sense
check.inputs<-function(MONTE,Nsims,Nhos,Nnos,theta,lambda){
if(!is.logical(MONTE)){stop("MONTE must be a logical value")}
if(MONTE){
 if(!is.numeric(Nsims))stop("In Monte Carlo model, Nsims must be an integer")
 if(!(Nsims==round(Nsims)))stop("In Monte Carlo mode, Nsims must be an integer")
}
if(Nhos!=round(Nhos))stop("Nhos must be an integer")
```

```
if(Nhos<0)stop("Nhos must be nonnegative")
if(Nnos!=round(Nnos))stop("Nnos must be an integer")
if(Nnos<0)stop("Nnos must be nonnegative")
if(theta>1)stop("Sample rate, theta, must be less than or equal to one")
if(theta<=0)stop("Sample rate, theta, must be greater than or equal to zero")
if(lambda>1)stop("VM fraction, lambda, must be less than or equal to one")
if(lambda<=0)stop("VM fraction, lambda, must be greater than or equal to zero")
return(NULL)
}
```