SPAM

Version 3.5

Addendum to User's Guide for Version 3.2

Addendum to Special Publication No. 15

Joel H. Reynolds
Alaska Department of Fish and Game
Division of Commercial Fisheries
Gene Conservation Laboratory
333 Raspberry Road
Anchorage, Alaska 99518

7 May, 2002

<page left intentionally blank>

SPAM

(Statistics Program for Analyzing Mixtures)

Version 3.5

Addendum to User's Guide for Version 3.2

Addendum to Special Publication No. 15

Joel H. Reynolds
Alaska Department of Fish and Game
Division of Commercial Fisheries
Gene Conservation Laboratory
333 Raspberry Road
Anchorage, Alaska 99518

7 May, 2002

Citing the Software:

SPAM 3.2:

Debevec, E. M. , R. B. Gates, M. Masuda, J. Pella, J. Reynolds, L. W. Seeb. 2000. SPAM (Version 3.2): Statistics Program for Analyzing Mixtures. Journal of Heredity 91 (6): 509 – 510.

SPAM 3.5:

Alaska Department of Fish and Game. 2001. SPAM Version 3.5: Statistics Program for Analyzing Mixtures. Alaska Department of Fish and Game, Commercial Fisheries Division, Gene Conservation Lab. Available for download from http://www.cf.adfg.state.ak.us/geninfo/research/genetics/Software/SpamPage.htm.

Citing the Manual:

SPAM 3.2:

Alaska Department of Fish and Game. 2000. SPAM Version 3.2: User's Guide. Special Publication 15, Alaska Department of Fish and Game, Commercial Fisheries Division, Gene Conservation Lab, 333 Raspberry Road, Anchorage, Alaska, 99518. 61 pages. Available for download from http://www.cf.adfg.state.ak.us/geninfo/research/genetics/Software/SpamPage.htm.

SPAM 3.5:

Reynolds, J. H. 2001. SPAM Version 3.5: User's Guide Addendum. Addendum to Special Publication 15, Alaska Department of Fish and Game, Commercial Fisheries Division, Gene Conservation Lab, 333 Raspberry Road, Anchorage, Alaska, 99518. 63 pages. Available for download from http://www.cf.adfg.state.ak.us/geninfo/research/genetics/Software/SpamPage.htm.

Table of Contents

<page left intentionally blank>

Overview of Enhancements

This addendum describes the features appearing for the first time in SPAM version 3.5. For a full description of the program, including input and output files, please see the SPAM version 3.2 User's Guide, available online[1]. The following descriptions and example applications assume a familiarity with the material in the User's Guide (referred to below as 'UG:3.2').

SPAM 3.5 provides two new bootstrap confidence intervals and the ability to conduct likelihood ratio tests using Monte Carlo simulation (Reynolds and Templin, in review). General descriptions are given below, followed by detailed technical information, keywords for associated control file options, analysis flowcharts, and example input and output files. Example applications of each feature are presented, including a step-by-step guide for conducting likelihood ratio tests of mixture equality.

In addition to the new features, a number of changes have been made to the default behavior of the program. These changes are listed after the general descriptions.

— *Confidence Intervals* —

SPAM 3.2 provided two types of confidence interval estimates: ***likelihood-based*** confidence intervals, which use a simplified binomial model and rely on asymptotic results (UG:3.2 § Control File: *Options), and ***symmetric percentile*** bootstrap confidence intervals (UG:3.2 § Control File: *Parameters). SPAM 3.5 also provides ***nonsymmetric percentile*** bootstrap confidence intervals and ***studentized*** or, as they are sometimes referred to, ***bootstrap-t*** confidence intervals (Lunneborg 2000). These two bootstrap intervals are, respectively, more appropriate for skewed sampling distributions (such as mixture contribution estimates when the true mixture contribution increasingly differs from 0.50), and have more accurate coverage than standard symmetric percentile intervals (Davison and Hinkley 1997). However, neither method 'respects' parameter range boundaries, such as the restriction that contribution estimates, θ, are limited to $0 <= \theta <= 1$. It is an unfortunate complication that these better methods are only applicable away from the parameter range boundaries, for this leads to a hierarchy of interval methods for different segments of the parameter range:

1. $0.3 <= \theta <= 0.7$ (roughly) – In this range, the sampling distribution of the contribution estimate is fairly symmetric. The best available bootstrap confidence interval method is the studentized or bootstrap-t confidence interval.

[1] http://www.cf.adfg.state.ak.us/geninfo/research/genetics/Software/SpamPage.htm

2. $0.1 <= \theta < 0.3$ or $0.7 < \theta <= 0.9$ (roughly) – In this range, the sampling distribution of the contribution estimate tends to become skewed. The nonsymmetric percentile method is better able to cope with this skewness than the symmetric percentile method.

3. $0.0 <= \theta < 0.1$ or $0.9 < \theta <= 1.0$ (roughly) - As the true contribution becomes more extreme, the parameter boundary wreaks havoc with the nonsymmetric interval method (and the bootstrap-t method). The (default) symmetric percentile method should be used.

The bootstrap-t method is further restricted to situations when the baseline is assumed to be known with certainty. The method requires an estimate of the variance of each bootstrap replicate contribution estimate, and this is only available (via the infinitesimal jackknife) when the baseline population estimates are not resampled.

While the SPAM user can still simply rely on the default symmetric confidence intervals, the alternative methods are provided for their improved coverage should the user choose to select them.

— *Likelihood Ratio Tests* —

SPAM uses maximum likelihood estimation. That is, the numerical algorithms search for the mixture parameterization that maximizes the likelihood of the observed mixture sample (for a general introduction to likelihood estimation, see Edwards 1992). With SPAM 3.5, the user can now request that the program save, for each simulated mixture or resampled mixture, the maximum value of the likelihood produced by the search. With these likelihood values, and a few other new options described below, the user can conduct likelihood ratio tests of competing mixture models. While requiring some processing outside of SPAM, these features greatly expand the types of mixture analyses researchers can undertake using the software.

The Examples section (starting on page 17) demonstrates how to use these features to (i) reduce bias in the mixture estimates by reducing to a more parsimonious baseline of populations, and to (ii) compare independent mixture samples for equality of mixture contributions. A third example demonstrates how to use SPAM 3.5 (or SPAM 3.2) to conduct a power analysis of sample size selection. All of these analyses require multiple SPAM runs and processing of SPAM output files in a spreadsheet or data analysis package. Flowcharts are also given to aid the user in undertaking the analyses.

— *Changes to the Default Behavior* —

- The default number of bootstrap resamplings, (keyword `SIZE` in the `*Parameters` section of the control file, UG:3.2) has been changed from 100 to 1000 following current statistical recommendations (Lunneborg 2000).

- SPAM 3.5 will no longer report an error when only 1 bootstrap resample is requested to allow certain likelihood ratio tests. It is recommended that a minimum of 1000 bootstrap resamplings be conducted for confidence interval and standard error estimation.

- The log file (`*.log`) has been enhanced to record the convergence criterion used for each bootstrap resample: guaranteed percent maximum, estimate tolerance, likelihood tolerance, or maximum number of iterations in the search algorithm. The user is now able to judge the reliability of the resulting estimates from each bootstrap replicate or Monte Carlo simulation by coordinating the `*.log` file contents with the `*.rsm` or `*.rlk` files. Monte Carlo likelihood ratio tests should only employ likelihoods from runs that adequately converge. S-plus (MathSoft 1999) code to summarize the `*.log` file and coordinate convergence results with the simulation or resample parameter estimates in the `*.rsm` file is available upon request from the Gene Conservation Lab.

- The sequence in which simulated observations are generated has been modified from SPAM 3.2 in order to implement the likelihood ratio test feature. If one runs identical simulation control files, with identical random number seeds, under both SPAM 3.2 and SPAM 3.5, the results will be approximately identical but not exactly identical as the random number generator will be called in a different sequence. In general, this slight backward-incompatibility should not present any problem; if it does, please contact the Gene Conservation Lab via the SPAM webpage http://www.cf.adfg.state.ak.us/geninfo/research/genetics/Software/SpamPage.htm.

- For other changes and answers to frequently asked questions, check the SPAM webpage.

New Control File Options

For full documentation regarding required input files, including the general options available in the control file, see the SPAM version 3.2 User's Guide. A brief summary follows.

A control file contains eight required sections, each identified by an asterisk (*) in the first column followed by a section keyword. The keywords are:

1) *ESTIMATION or *SIMULATION
2) *OPTIONS
3) *PARAMETERS
4) *CHARACTERS
5) *POPULATIONS
6) *REGIONS
7) *FILES
8) *RUN

For proper execution of SPAM, each section must appear in the control file and in the order given here. The section keywords can occur anywhere in the line following the asterisk and can be upper- or lowercase. Only the first four characters of a keyword are required (except for RUN). Other words can be mixed with the keywords to allow for more readability; however, care should be taken not to mix keywords within a line. SPAM will parse the control file line by line until it reaches a recognizable section label keyword. Once a keyword is encountered, all subsequent lines belong to that section until the next keyword is encountered. This means that while lines within a section can be in any order, all lines belonging to one section should remain within that section.

Each section consists of a series of program control statements that specify the SPAM analysis. As with section labels, each control statement uses one or two keywords that are recognized by SPAM by their first four characters. See UG: 3.2 for a full discussion of options. While all control statements are shown, only the ones introduced in version 3.5, listed in **bold**, are described. Examples of the various control file sections are provided in shaded text boxes.

NOTE: To turn on and off each option, SPAM accepts T, F, TRUE, FALSE, YES, NO, ON and OFF. The parser is not case sensitive. The switch is always the first argument to the right of the colon. SPAM will use default values for any of the control

statements that are not specified in the control file. A section label is still required if all defaults are to be used.

* Estimation/Simulation

No new options.

* Options

```
* options selected for optimization
  use IRLS algorithm in optimal search       :  f
  print mixture file                         :  t
  print baseline relative frequencies        :  t
  print conditional genotype probabilities   :  t
  print conditional population probabilities :  t
  print bootstrap estimates                  :  t
  print iterations                           :  t
  print likelihoods of simulations/resamples :  t
  compute likelihood confidence intervals    :  t
  compute infinitesimal jacknife std.dev.    :  t
  compute studentized conf. intervals        :  t
  compute likelihood at external estimate    :  t
  resample mixture frequencies               :  t
  resample baseline                          :  t
```

This section is used to select performance and output options. The keywords for the various options are listed here.

Keyword(s)	*Default*	*Description*
IRLS	F	Use IRLS algorithm in optimal search
PRIN		Print...
BASE	F	baseline relative frequencies
MIXT	F	mixture file
GENO	F	conditional genotype probabilities
POPU or STOC	F	conditional population (stock) probabilities
BOOT	F	estimates from each bootstrap resample
ITER	F	MLE search iterations
LIKE	**F**	**likelihood for each simulation or resample**
COMP		Compute...
CONF	F	likelihood confidence intervals

JACK	F	infinitesimal jackknife standard deviations
STUD	**F**	**studentized bootstrap conf. intervals**
EXTE	**F**	**use externally provided m.l.e. in likelihood ratio**
RESA		Resample...
MIXT	F	mixture frequencies
BASE	F	baseline

— PRINT —

LIKELIHOOD

The keyword LIKELIHOOD is used when conducting likelihood ratio tests of different mixture models using Monte Carlo simulation. Two detailed examples are given in §Examples (page 17).

Setting the keyword LIKELIHOOD to one of {t, true, yes, on} will cause SPAM 3.5 to record the natural logarithm of the maximized likelihood for each simulated mixture. An output file (named *.rlk) is created whose first column records these values (in the notation of Reynolds and Templin (in review), $\ln(L(\hat{\Theta}* | \mathbf{X}*, \hat{\Phi}))$). The other columns of numbers are currently used for error checking (see §Examples).

— COMPUTE —

STUDENTIZED

The keyword STUDENTIZED will cause SPAM 3.5 to calculate studentized or bootstrap-t confidence intervals for each regional contribution estimate. These intervals will be listed in the *.bot file. Nonsymmetric percentile bootstrap confidence intervals are automatically listed in the *.bot file whenever bootstrap resampling is requested. Note that the different bootstrap intervals are appropriate in different settings (see page 1 for a description). The studentized intervals are the most accurate of the three bootstrap methods available, but can only be used when the baseline population estimates can be treated as known without error and when the sampling distribution is approximately symmetric (i.e., contribution proportions are not extreme).

EXTERNAL

If both keywords LIKELIHOOD and EXTERNAL are set to one of {t, true, yes, on}, and it is a simulation, then the second column of the *.rlk file will record the log-likelihood of the simulated data under the mixture parameterization given by the ESTIMATE keyword in the *Populations section of the control file (see UG:3.2 page25). This option is provided to allow development of bootstrap likelihood confidence intervals (see '§5.8 Multiparameter Confidence Intervals' in Davison and Hinkley 1997).

* Parameters

The control parameters specify the number of populations and characters in the analysis, upper limit parameters, tolerances to control the optimization search, and features to synchronize and partition simulated mixture samples across runs of SPAM 3.5 (for use in Monte Carlo simulation of likelihood ratios for testing mixture equality).

```
* control parameters
  number of populations in analysis  :  14
  number of characters in analysis   :   9
  maximum number of genotypes        : 200
  maximum number of classes          :  20
  maximum # of iterations            : 300
  maximum number of missing loci     :  4
  estimate tolerance                 : .1E-3
  likelihood tolerance               :  1.0e-10
  genotype tolerance                 :  1.0e-6
  algorithm switch tolerance         :  0.01
  GPA                                :  90
  number of resamplings              :  100
  simulation sample size             :  100

  number of null observations after  :  0
  confidence intervals               :  90
  random seed                        :  -718805
```

The keywords for the command are listed here:

Keyword(s)	*Default*	*Description*
NUMB		Number of…
POPU or STOC	-	populations (stocks) in the analysis
CHAR	-	characters in the analysis
RESA	100	bootstrap resamplings
BEFO	**0**	**observations to simulate, but not use, before simulating mixture sample of interest.**
AFTE	**0**	**observations to simulate, but not use, after simulating mixture sample of interest.**
MAXI		Maximum number of…
GENO	100	genotypes
CLAS	1	classes

ITER	100	iterations
MISS	0	missing (unscored) loci in mixture
TOLE		Tolerances for…
ESTI	10^{-4}	estimates
LIKE or FUNC	10^{-10}	likelihood (function)
GENO	10^{-10}	genotype probability
ALGO or SWIT	10^{-2}	algorithm switch (CG to IRLS)
GUAR, PERC, or GPA	90	Guaranteed percent achievement of the maximal likelihood (GPA)
CONF	90	Confidence interval size (percent)
SIZE	100	Simulation sample size
SEED	*From CPU clock*	Random number generator seed

— *NUMBER* —

BEFORE and *AFTER*

The BEFORE and AFTER keywords allow one to re-simulate a subset of a larger simulated mixture sample. This allows simulated mixtures to be partitioned into (random) subsamples in a manner that is coordinated across multiple SPAM runs. This feature allows one to approximate, using Monte Carlo simulation, the null reference distribution for testing mixture equality with likelihood ratios. It is best explained by example.

Assume we want to compare three mixture samples, {A, B, C}, of sizes 100, 200, and 300, respectively. To approximate the null reference distribution, we need to be able to simulate a single mixture sample of size 600 (=100+200+300) then randomly partition it into three smaller mixture samples, {A', B', C'}, of sizes 100, 200, and 300. This gives three mixture samples that all come from a common mixture. We want to fit each of the smaller mixture samples independently, calculating their maximum likelihoods, and then combine them back into a single sample and fit that, calculating its maximum likelihood. The first calculation gives the likelihood of the simulation under the alternative model where the three samples come from three possibly different mixtures. The second calculation gives the likelihood of the simulation under the null model where all three samples come from a common mixture. The calculations are detailed in §Examples (page 17); the key here is to see the need for a coordinated method of partitioning randomly generated mixture samples.

This would be simple if we actually had a file listing the 600 simulated character vectors (for example, genotypes). However, we want to have SPAM 3.5 do as much of this as possible internally, reducing the need to create, then edit, then re-read external mixture files. If one simply calls SPAM three times, simulating mixture samples of size 100, then 200, then 300, you will NOT get exactly the same 600 character vectors simulated in the original call because the random number generator won't be synchronized properly. Rather, the random number generator requires that you re-simulate all 600 character vectors each time, but only use the first 100 for fitting A', then only use character vectors 101 – 300 for fitting B', then only use character vectors 301 – 600 for fitting C'. SPAM 3.5 can do this using the keywords BEFORE and AFTER:

(i) If the *Parameters section of *.ctl contains

```
Number of Resamples   :1000
BEFORE                :0
AFTER                 :0
SIZE                  :600
SEED                  :10000
```

SPAM 3.5 will simulate, and fit, 1000 mixtures of 600 randomly generated observations.

(ii) If the *Parameters section of *.ctl contains

```
Number of Resamples   :1000
BEFORE                :0
AFTER                 :500
SIZE                  :100
SEED                  :10000
```

SPAM 3.5 will simulate, and fit, 1000 mixtures of 100 randomly generated observations, where each mixture is simulated by randomly generating 600 observations and taking the first 100 (600 = BEFORE + SIZE + AFTER = 0 + 100 + 500).

(iii) If the *Parameters section of *.ctl contains

```
Number of Resamples   :1000
BEFORE                :100
AFTER                 :300
SIZE                  :200
SEED                  :10000
```

SPAM 3.5 will simulate, and fit, 1000 mixtures of 200 randomly generated observations, where each mixture is simulated by randomly generating 600 observations and taking the 101^{st} – 300^{th} observations (600 = BEFORE + SIZE + AFTER = 100 + 200 + 300).

(iv) If the *Parameters section of *.ctl contains

```
Number of Resamples   :1000
BEFORE                :300
```

```
AFTER                :0
SIZE                 :300
SEED                 :10000
```

SPAM 3.5 will simulate, and fit, 1000 mixtures of 300 randomly generated observations, where each mixture is simulated by randomly generating 600 observations and taking the 301st – 600th observations (600 = BEFORE + SIZE + AFTER = 300 + 300 + 0).

Note that each set of simulations, (i) – (iv), requires its own control file (*.ctl) and will produce its own set of output files. Also, note that each of these control files must be using the same random number SEED in order to synchronize the random number generator across the calls to SPAM. If the *Options keyword LIKELIHOOD is toggled on, then each set of the simulations will produce a file (*.rlk) listing the maximum log-likelihood of each simulated mixture. These can be processed outside of SPAM to generate 1000 random observations from the null reference distribution of the likelihood ratio. See §Examples (page 17). In general, generating the null reference distribution to test the equality of M mixture samples will require M + 1 *.ctl files and calls to SPAM.

* Characters

No changes. See UG:3.2 for full command description.

* Populations

The information in the *Populations section defines the identification number, population names, baseline files, and regional aggregation of populations. This section also allows the user to input an initial estimate of the mixture when estimating, or identify the mixture to simulate when simulating. See UG:3.2 for full command description.

```
 * populations used in analysis
 [id #] [population]         [file]        [lev1] [Estimate]
   1    Warm Springs        : warm.frq     :  2       1
   2    Rapid               : rapid.frq    :  4       2
   3    Kooskia             : kooskia.frq  :  4       3
   4    Round Butte         : round.frq    :  5       1
   5    Carson              : carson.frq   :  3       1
   6    Eagle               : eagle.frq    :  1       2
```

The first line after the *Populations keyword defines the order in which the population attributes will be provided. The attribute labels are listed here.

Label	*Default*	*Attribute*
[ID #]	*see text*	Identification number for the population
[POPULATION]	*see text*	Population name
[FILE]	*see text*	Baseline file
[LEV1]	-	Level 1 regional identifier
[LEV2]	-	Level 2 regional identifier
[LEV3]	-	Level 3 regional identifier
[ESTIMATE]	*see text*	Initial estimates (estimation) or relative population contribution to mixture (simulation)

When performing an estimation analysis, the ESTIMATE identifier is used to set the initial contribution estimates of the mixture (in general, one should not enter starting values of zero). If it is not provided, the starting values in the MLE search are $1/p$, where p is the number of populations defined under the *Parameters command. It is useful to try various starting values to verify that the same contribution estimates are obtained, providing evidence that the true maximum likelihood is found and not just a local maximum.

For a simulation analysis, the ESTIMATE identifier defines the true mixture that is generated stochastically using the baseline frequencies. Values for ESTIMATE do not have to sum to one, and can be on any convenient scale. For example, the listing above (shaded box) would simulate samples from a mixture with the following contributions: Warm Springs - 10%, Rapid – 20%, Kooskia – 30%, Round Butte – 10%, Carson – 10%, Eagle – 20%.

* Regions

No changes. See UG:3.2 for full description.

* Files

No changes. See UG:3.2 for full description.

* Run

No changes. See UG:3.2 for full description

Input Files

No changes. See UG:3.2 for full description.

Output Files

All results from a SPAM analysis are printed to a collection of ASCII text files that can be viewed through the SPAM environment or separately with any text editor. The set of files created depends on the analysis requested in the control file. All files, except the resampled estimate files, are formatted for convenient viewing and printing. Every SPAM analysis will produce a log file (`*.log`) and either an estimation (`*.est`) or a simulation file (`*.sim`), depending on the type of analysis run. Only changes in content or new files are discussed below. See UG:3.2 for a full description of the other output files created by SPAM 3.5.

Log (`*.log`)

Every SPAM analysis generates a log file containing a list of the steps undertaken and any errors encountered. This file must be reviewed to make sure the estimation procedure converged properly. The log file uses the same path and root filename as the control file since the log file is initiated before the control file is parsed. See UG:3.2 for a full description.

NEW - This file now lists the convergence criterion employed for each bootstrap resample or Monte Carlo simulation, as well as the GPA estimate ('guaranteed percent achieved' – the ratio of the final likelihood where the algorithm stopped divided by an estimate of the upper bound on the unknown maximum value of the likelihood, see Pella et al. 1996). If either Monte Carlo simulation or bootstrap resampling is being used, it is imperative that the log file be reviewed. One should check to make sure each replicate estimation attained convergence and that the same convergence criterion was used each time (preferably GPA as all the other stopping criteria are relative convergence criteria). Recent empirical investigations have shown that varying the stopping criterion employed by the E-M algorithm can lead to different maximum likelihood estimates (Seidel et al. 2000). This reinforces the need for users to check, and if necessary screen, the Monte Carlo simulation or bootstrap replication results. See the Frequently Asked Questions section of the SPAM webpage[2] for discussion of convergence issues and interpretation of this file.

[2] http://www.cf.adfg.state.ak.us/geninfo/research/genetics/Software/SpamPage.htm

Estimation (*.est)

See UG:3.2 for a full description.

log-likelihood (line 4 of the file): the value listed is the *support function* evaluated at the observed (or simulated) mixture data. That is $\ln(L(\hat{\Theta} \mid \mathbf{X}, \hat{\Phi})) = \sum_{i=1}^{n} \ln(\Pr(x_i \mid \hat{\Theta}, \hat{\Phi})) = \sum_{i=1}^{n} \ln\left[\left\{\sum_{j=1}^{J} \Pr_j(x_i \mid \hat{\theta}_j, \hat{\phi}_j)\right\}\right]$ where X is the mixture sample, $\{x_i\}$, $\hat{\Theta}$ is the maximum likelihood estimate of the unknown mixture proportions conditional on the observed baseline character frequency distributions, $\hat{\Phi}$, and $\Pr(x_i)$ is the probability of observing a characteristic vector x_i while $\Pr_j(x_i)$ is the probability of observing a characteristic vector x_i from Population j.

Bootstrap Output file (*.bot)

Created when the *Options keywords RESAMPLE BASE or RESAMPLE MIX are selected. In addition to the results discussed in UG:3.2, the *.bot file will also automatically list the nonsymmetric percentile bootstrap confidence interval for each region. PLEASE review the discussion on page 1 for limits as to when this method is trustworthy.

If the *Options keyword COMPUTE STUDENTIZED is toggled on, then *.bot will also list the bootstrap-t confidence interval for each region. PLEASE review the discussion on page 1 for limits as to when this method is trustworthy. An error will be reported if this option is requested in conjunction with bootstrap resampling of the baseline.

Likelihood Output file (*.rlk) Appendix 1

Created when the *Options keyword PRINT LIKELIHOOD is toggled on, this file contains five columns for each simulated or bootstrap replicate mixture sample. See the §Examples section for usage of this information.

Column one: the value of the support function evaluated on the current mixture sample at the maximum likelihood mixture estimate, $\ln(L(\hat{\Theta}* \mid \mathbf{X*}, \hat{\Phi}))$.

Column two: the value of the support function evaluated on the current mixture sample at either (i) the original mixture sample's mixture parameter estimate (default), or (ii) the mixture given by the Estimate keyword in the *Populations section (when *Options keyword EXTERNAL is toggled on), $\ln(L(\hat{\Theta}_{\mathbf{obs-or-external}} \mid \mathbf{X*}, \hat{\Phi}))$.

Columns three, four, and five record the value of the random number seed when the current simulation or estimation round enters, respectively, (a) the baseline resampling function, (b) the mixture simulation or resampling function, and (c) completes the mixture simulation or resampling. These values are provided for error checking during Monte Carlo simulation of the null reference distribution for likelihood ratio tests. For example, if one is testing equality of M mixture samples (see §Examples), then each of the M+1 `*.rlk` files created should have the same last three columns. If these columns differ across files, then there is an error in the `BEFORE`, `AFTER`, or `SIZE` settings.

Note that if one is not resampling the baseline, then the value in column three will always be zero.

Examples

The three example analyses illustrate how to use SPAM 3.5 to: (i) conduct a Monte Carlo likelihood ratio test of whether a smaller population baseline is sufficient for explaining the observed mixture sample ('baseline reduction'), (ii) to test whether two or more independently collected mixture samples came from the same underlying mixture, or (iii) to conduct a priori mixture sample size power analyses.

Flowcharts for each analysis are given at the end of this section.

Reducing to a more parsimonious collection of baseline populations

Increasingly, mixed stock analyses are conducted using very large baselines. For example, recent analysis of illegal highseas sockeye salmon (*Oncorhynchus nerka*) harvests used an international sockeye baseline of 170 populations (Wilmot et al. 2000) and analysis of chum salmon (*O. keta*) harvests along a major migratory pathway used an international baseline of over 250 populations (Seeb and Crane 1999). These analyses estimate mixture contributions using conditional maximum likelihood estimation, the method implemented in SPAM. That is, all potentially contributing populations are assumed to be in the baseline and every population in the baseline is assumed to be a potential contributor. Conditioning on such extensive baselines can exacerbate concerns of bias and precision in the mixture estimates (see Reynolds et al. 1 in preparation).

Likelihood ratio tests can be used to find a more parsimonious baseline that is sufficient to have generated the observed mixture sample. Reducing to a more parsimonious baseline can reduce the potential bias from overestimating rare or absent populations, and hence underestimating major contributors (see discussion in Reynolds et al.1 in preparation). Two different models are fit and their likelihoods compared: the general (alternative) model in which the full baseline is used in the fitting, and the null model in which the populations proposed to not contribute are dropped from the baseline before fitting (forcing their contribution to zero). This is basically a 'variable or model selection' problem, and therefore subject to the same concerns that occur when fitting regression models (Ryan 1997). For example, issues of multiple comparisons if a large sequence of testing is employed to reduce the baseline, and masking of small significant effects if large numbers of parameters are dropped simultaneously.

We illustrate the method with an analysis of chinook salmon (*O. tshawytscha*) troll fishery harvest in Southeast Alaska (Reynolds et al. 1 in preparation). The analysis uses the coastwide baseline of allozyme data from 254 populations ranging from

California through Alaska and including two populations from far eastern Russia (Teel et al. 1999). Ostensibly, we conduct a sequence of model fitting procedures where the models are defined by the populations included in the baseline. This topic is discussed in further detail in Reynolds et. al. 1 (in preparation).

There are two stages of calculations: (1) calculate the observed likelihood ratio, and then (2) approximate the likelihood ratio distribution when the null model is true. The results from both stages are used in estimating a p-value for testing the null hypothesis. The method assumes that the mixture sample is a simple random sample, that all contributing populations are included in the baseline in each fit, and that the characteristics being observed are in equilibrium within each population – i.e., loci are in Hardy-Weinberg equilibrium. The statistical assumptions are discussed in more detail in Reynolds et. al. 1 (in preparation).

Partition the complete set of baseline populations into two subsets: subset A, which is expected to contribute to the mixture, and subset B that is not expected to contribute to the mixture. We wish to test the null hypothesis that the contribution from every population in subset B is zero. In practice, subset B will be defined using the results from an initial fit using the full baseline.

Calculating the Observed Likelihood Ratio

The observed likelihood ratio requires the likelihood of the observed mixture sample under both the null model, using only the subset A populations in the baseline, and the general model, using the full set of populations in the baseline. This information will require two estimation runs of SPAM, each with its own control file.

(i) Use SPAM to estimate the mixture using the full baseline. Be sure the control file options are set to record bootstrap confidence intervals. In the resulting `*.est` file, record the value listed as *log-likelihood* (line 4); this is the *support* for the observations under the general model. See first example `*.ctl` file in Appendix 1 (page 48) and first example `*.est` file in Appendix 2 (page 50). Under this general model, the example observations have a support of –2806.76.

(ii) Select the subset B populations, those that don't appear to contribute to the mixture. If regions were defined in the `*Population` section of the `*.ctl` in (i), review the bootstrap confidence intervals for each region's contribution (in the file `*.bot`). Otherwise, review the population contribution estimates and standard errors (in the file `*.est`). Regions whose confidence intervals contain zero, or any population whose estimated contribution is zero or whose estimated contribution is within 1-2 standard errors of zero, should be assigned to subset B. Adjusting the confidence level to account for issues of simultaneous inference and multiple testing (see Reynolds and Templin, in

review) will increase the number of populations assigned to subset B.

(iii) Estimate the mixture using the baseline without the subset B populations. A simple way to do this is to comment out the appropriate lines in the *Populations section of the *.ctl file. Place either a backward slash (\) or a forward slash (/) as the first character in each line that refers to a subset B population and change the number of populations in the *Parameters section. For example, the Region 1 populations are commented out in the following excerpt from the *Populations section of the control file.

```
* populations used in analysis
 [id #]  [population]           [file]        [lev1]
/1 Mokelumne & Nimbus H. fall :  kMoaNHa.frq  :   1
/2       Merced Hatchery fall :  kMerHat.frq  :   1
/3      Feather Hatchery fall :  kFeaHat.frq  :   1
/4    Feather Hatchery spring : kFeaHat1.frq  :   1
/5      Coleman Hatchery fall :  kColeHa.frq  :   1
/6 Upper Sacramento R. winter :  kUSacra.frq  :   1
7          Mattole River fall :  kMattoR.frq  :   2
8        Van Duzen River fall :  kVDuzRi.frq  :   2
9           Salmon Creek fall :  kSalmCr.frq  :   2
10         Redwood Creek fall :  kRedwoC.frq  :   2
...
```

Population id numbers do not need to be revised. However, if regions are dropped from the baseline, then the region id numbers (assigned under [lev1], etc.) need to be reassigned both in the *Population section and in the *Regions section. See the second example control file in Appendix 1 (page 48) and second *.est file in Appendix 2 (page 50).
Under the null model, the example observations have a support of –2815.86.

(iv) SPAM gives us the natural logarithm of the likelihood, not the likelihood. Rather than exponentiating, we can conduct all the calculations on the logarithm scale. The value of interest, the logarithm of the likelihood ratio, is equal to the support under the general model (from step (i)) minus the support under the null model (from step (ii)): -2806.76 –(-2815.86) = 9.10.

Calculating the Null Reference Distribution

Assessing the significance of the observed likelihood ratio requires that we know the distribution of likelihood ratios expected to be observed when the null hypothesis is true. That is, we need to know the distribution of likelihood ratios expected when none of the populations in subset B contribute to the mixture. A conditional approximation to this null reference distribution can be constructed using Monte Carlo simulations in SPAM 3.5.

The null hypothesis makes no claim regarding the value of the contribution from any population in subset A. However, values must be assigned to these contributions in order to conduct Monte Carlo simulations. We form a conditional

test by fixing the subset A contributions to their estimates under the null model (McLachlan and Peel, 2000). Simulation studies have shown some tendency for this approach to overestimate the upper tails of the null reference distribution, and hence overestimate the P-value (McLachlan and Peel 2000, p. 200). That is, the test may tend not to reject.

The *.est file from step (iii) above contains the mixture estimate under the null model. What we need to know is the distribution when one takes, in this case, a sample of size 328 from this mixture, calculates the likelihood of this sample under both the null and general models, then calculate the ratio of the likelihoods (actually, the log of the likelihood ratio). We can approximate this null reference distribution by having SPAM 3.5 repeatedly simulate such samples and then calculate their likelihoods.

The null reference distribution can be approximated using just two separate, but synchronized, *simulation* runs of SPAM 3.5 if the baseline population allele frequencies can be treated as known without error. This approach is illustrated below.

(i) Create a *.ctl file to simulate R samples of 328 observations from the full baseline using the mixture given in the *.est file from step (iii) above. Use the ESTIMATE label in the *Populations section to enter the estimated mixture contribution for each baseline population. Set GPA to a high value, such as 98 (98%), the maximum number of iterations in each search to be high (~1000), and set the other convergence tolerances to be very low. As we are interested in the likelihood values themselves, we need to make sure the final estimates are very near the (unknown) true maximum likelihood value. See Appendix 3 (page 51) for example *.ctl files.

R should be on the order of 1000 – 5000 (Davison and Hinkley, 1997). Set PRINT LIKELIHOOD :True in the *Options section so that the *.rlk output file is produced. Running this *.ctl file in SPAM 3.5 will create a *.rlk file whose first column lists, for each simulated mixture, the support for the observations under the general model (full baseline).
The first few lines of this *.rlk file were:

```
-2811.0      -2846.5              0   -4533880     438528
-2855.9      -2880.4              0     438528     678120
-2722.1      -2751.6              0     678120      11677
-2931.3      -2953.3              0      11677     298199
```

The *.log file must be reviewed to make sure each simulation converged according to the 'GPA' criterion. If a simulation does not converge by GPA, note which simulation it was (in the sequence) and proceed.

(ii) Now create a *.ctl file to re-simulate the R samples of 328 observations using the mixture given in the *.est file from step (iii) above, but fit using the reduced baseline (subset A populations only). This is easily done by taking the *.ctl file from step (i) and commenting out the subset B populations as in step (ii) in the *Observed Likelihood Ratio* description above. The seed used should be identical to the one just used in step (i), as should be the GPA setting and all of the other convergence tolerance values.

Running this *.ctl file in SPAM 3.5 will create a *.rlk file whose first column lists, for each simulated mixture, the support for the observations under the null model (reduced baseline).
The first few lines of this *.rlk file were:

```
-2817.3      -2846.5              0   -4533880     438528
-2865.0      -2880.4              0     438528     678120
-2727.8      -2751.6              0     678120      11677
-2937.2      -2953.3              0      11677     298199
```

The last three columns in the *.rlk files from this and the former step should be identical. If they are not, the random number generator is not synchronized between the runs and the simulations are not identical – check that the random number seed is the same in both *.ctl files.

Check the *.log file to make sure each simulation converged by the GPA criterion. If a simulated mixture did not converge by GPA in under either the full baseline model (step (i)) or reduced baseline model (step (ii)), it should be removed from both sets of likelihood results. I.e., only adequately converged simulations should be used in approximating the null reference distributions.

(iii) Subtract the support for the first set of simulated values under the null model (that is, the first value in the first column of the *.rlk file from step (ii)) from the support for the first set of simulated observations under the general model (the first value in the first column of the *.rlk file from step (i)). This gives one random observation from the null reference distribution. Subtract the support for the second set of simulated values under the null model (that is, the second value in the first column of the *.rlk file from step (ii)) from the support for the second set of simulated observations under the general model (step (i)). This gives a second random observation from the null reference distribution. Repeat this process to get R random observations from the null reference distribution.

For example, the log of the likelihood ratio testing the performance of the reduced baseline against that of the full baseline is
-2811.0 - -2817.3 = 6.3 for the first Monte Carlo simulation,
-2855.9 - -2865.0 = 9.1 for the second Monte Carlo simulation, etc.

These values should all be >=0. If they aren't, either the simulations are not being properly synchronized between step (i) and step (ii) or the simulations are not all converging properly. Check each of the *.ctl files to make sure the same random number seed, GPA value, and other convergence tolerance values were used. Also check both *.log files to make sure all the simulations converged using the GPA criterion. If some did not, they should be discarded as they might not have stopped sufficiently close to the maximum likelihood value.

Most runs will stop very near, but not exactly at, the maximum of the likelihood surface (the value of GPA determines how near). If a mixture is just as likely under the full baseline model as under the reduced baseline model, then very small differences in the attained GPA's at the final estimate of the likelihood maximum under the two models may produce very small numerical differences in the likelihood values. This may produce very small negative log likelihood ratios that in 'fact' are zeros. They should be considered zeros in the null reference distribution.

(iv) Count how many of the R random log likelihood ratios are equal to or larger than the observed log likelihood ratio; call this Q. The approximate Monte Carlo *P value* for the likelihood ratio test is (1 + Q) / (1 + R) (Davison and Hinkley 1997), which in this case is (1+155)/(1+1000) = 0.16. See Figure 1 for the approximate null reference distribution.

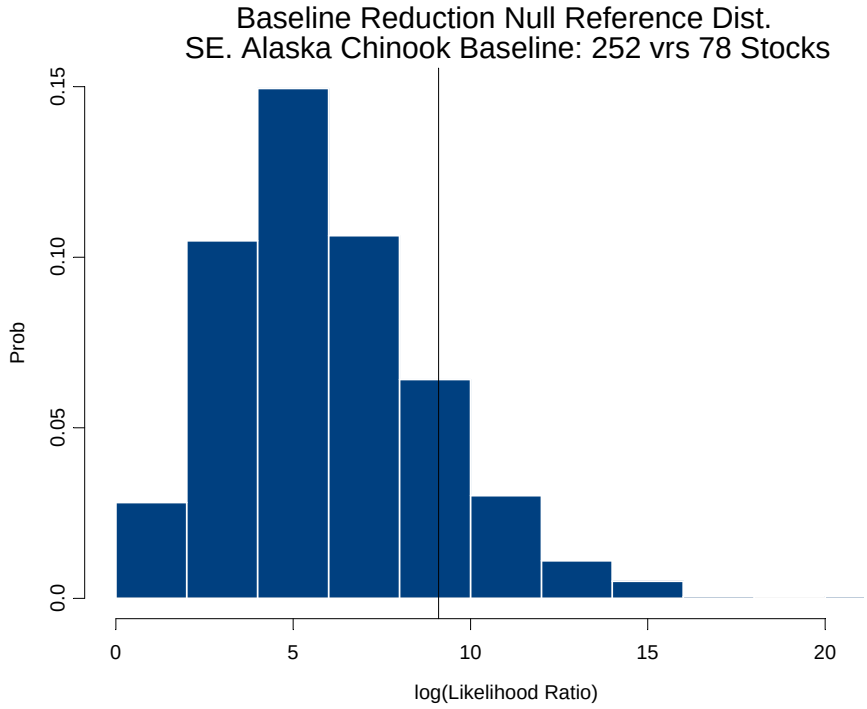


Figure 1. Approximate null reference distribution for the conditional likelihood ratio test of *Ho: Only subset A populations contribute to the observed mixture (78 stocks in 7 regions)* versus *Ha: subset A and subset B populations contribute to the observed mixture (258 stocks in 28 regions).* The null reference distribution is approximated by 1000 Monte Carlo simulations conducted in SPAM 3.5. A vertical dashed line marks the observed likelihood ratio. The test suggests that a baseline of only 78 stocks in 7 regions is sufficient to explain the observed mixture sample. The mixture sample was taken from opening 1 of the summer commercial chinook troll harvest in Southeast Alaska in 1999 (Crane et al. 2001).

The test fails to reject the baseline reduction from twenty-eight regions (252 populations) to seven regions (78 populations). When the mixture is fit using the smaller baseline, the estimated contribution from the Southern Southeast Alaskan region shifts significantly (Figure 2). As the reduced baseline is sufficient to generate the observed mixture sample, this shift may represent bias from overestimating absent stocks when fitting to the full baseline. Alternatively, it may represent actual low-frequency contributions from populations that are in excluded regions but which are sufficiently similar to the populations in the Southern Southeast Alaska region that their contribution has been absorbed into the estimated contribution of the Southern Southeast Alaska region. Arguing against this last interpretation is the fact that the regions are defined based on genetic similarity (Teel et al. 1999) and that the proposed 'excluded' populations differed sufficiently to be assigned to regions other than Southern Southeast Alaska. This lends weight to the former interpretation.

It is important to recognize this limitation: the test is only applicable if the reporting regions are defined based on genetic differences among populations. If they are defined otherwise, there is no basis for deciding whether a baseline reduction is overcoming bias or simply providing a more parsimonious description of the mixture by re-assigning contributions among genetically similar regions.

This method can detect small contributions from populations whose allele frequencies differ sufficiently from the rest of the baseline. The extreme case would be small contributions from a population with private alleles (Reynolds et al. 2 in preparation). The power to detect the population will be a function of the contribution rate, the mixture sample size, and the divergence of the population's allele frequencies relative to the rest of the baseline. If there are individuals in the

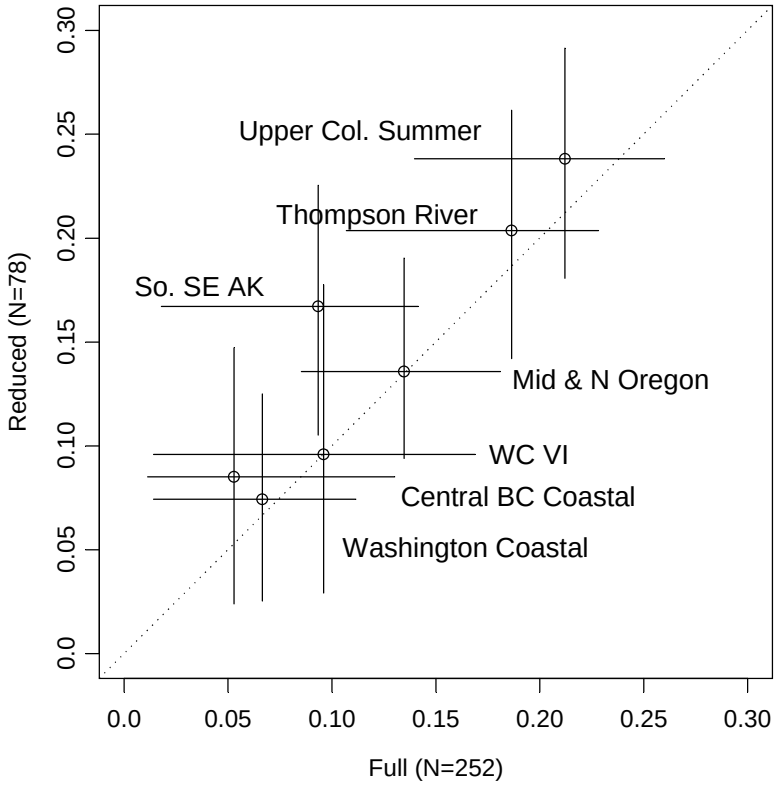


Figure 2. Plot of contribution estimate and 90% symmetric bootstrap confidence interval (1000 resamplings) under both the full baseline model (horizontal) and the reduced baseline model (vertical) for the seven regions in the reduced baseline model (Washington Coastal, Central British Columbia Coastal, West Coast Vancouver Island, Mid & Northern Oregon Coastal, Southern Southeast Alaska, Thompson River, Upper Columbia Summer and Fall and Snake Fall). The dashed line marks the 1:1 line; estimates that fall near this line change very little when the baseline is reduced. Note that this line is not contained in the confidence interval estimates for the Southern Southeast Alaska region ("So. SE AK"), suggesting a significant shift in contribution estimate depending on whether one uses the full baseline or the reduced baseline model. It is likely that this is bias from using such a large baseline.

mixture sample that have alleles unique to single population in the baseline (private alleles), and that population is dropped in a baseline reduction test, SPAM will assign a contribution to the 'unknown population' category in the reduced baseline run. This indicates that one or more populations in subset B truly contribute to the mixture and CANNOT be dropped from the baseline.

Testing equality of two mixture samples

Assume samples are available from two or more mixtures of the same baseline populations. A natural question is whether the mixture contributions are equal across the samples. Reynolds and Templin (in review) describe an investigation where commercial harvest deliveries from a fishery were sampled at two different processors to test whether just sampling one processor would introduce bias in the mixture estimate. That is, if the two mixture samples give equivalent mixture estimates then it is sufficient to sample deliveries at a single processor in estimating the total commercial harvest mixture. If the samples give different mixture estimates, then sampling just one processor will give a biased result.

Researchers have tested equality of mixtures by (i) calculating confidence intervals for each contribution from each mixture and (ii) comparing the respective confidence intervals across mixture samples. Contributions whose confidence intervals failed to overlap were taken to be statistically different (for example, Wilmot et al. 1998; Ruzzante et al. 2000). This approach has many statistical problems, the most important being that it inflates both the Type I and Type II error rates. See Reynolds and Templin (in review) for a full discussion.

Alternatively, one can use likelihood ratios to directly test equality of M different mixture samples. The likelihood ratio compares the likelihood of the observed samples under two different models: (a) the null model, where each of the M samples comes from a common mixture, and (b) the alternative model where each of the M samples comes from a (possibly) different mixture. The observed likelihood ratio is obtained by fitting both models then forming the ratio of their likelihoods. This observed likelihood ratio is then compared to the reference distribution of likelihood ratios expected under the null hypothesis to calculate a *P*-value.

We present an example from Reynolds and Templin (in review); interested readers should refer to their paper for details. Two samples of commercially harvested sockeye salmon (*O. nerka*) were taken from a fishery in Upper Cook Inlet, Alaska. The samples were from deliveries, made on the same day (21 July 1997), to two different processors on the Kenai River: Ward's Cove and Salamatof Seafood Inc. The question of interest is whether the samples appear to be from the same mixture. There are 398 observations in the Ward's Cove sample and 394 observations in the Salamatof Seafood, Inc., sample. The baseline consists of 44 populations (see Seeb et al. 2001 for details).

Background on the species, the fishery, and the motivating management problem are given in the application section of Reynolds and Templin (in review). Details of the calculations required to test mixture equality are given below. The necessary control files are given in the appendices.

There are two stages of calculations: (1) calculate the observed likelihood ratio, and (2) approximate the likelihood ratio distribution when the null model is true. The results from both stages are used in estimating a p-value for the test of mixture equality. Briefly, the method assumes that all mixture samples are simple random samples that were gathered independently, that all contributing populations are included in the baseline, and that the characteristics being observed are in equilibrium within each population. See Reynolds and Templin (in review) for a more thorough discussion of the statistical assumptions.

Calculating the Observed Likelihood Ratio

We wish to test the equality of two mixture samples. Calculating the observed likelihood ratio requires the likelihood of the observations both under the null model, where the samples come from the same mixture, and under the general model, where they come from potentially different mixtures. The necessary information will require three estimation runs of SPAM, each with its own mixture file and control file. In general, calculating the observed likelihood ratio to test the equality of M different mixture samples requires M+1 *estimation* runs of SPAM. The required information is in the `*.est` output file produced by each run.

(i) Use SPAM to estimate the mixture in each sample separately. In the example, this requires two runs of SPAM, each with a different mixture file and `*.ctl` file – one for the Ward's Cove sample and one for the Salamatof Seafoods, Inc. sample. Record the value listed as *log-likelihood* (line 4) from each of the resulting `*.est` files. See Appendix 4 (page 53) for sample `*.ctl` files and Appendix 5 (page 54) for the resulting `*.est` files.

(ii) Sum the log-likelihoods. This value, known as the *support*, is proportional to the maximum log-likelihood of the observations under the general model (equation 2 in Reynolds and Templin, in review).

(iii) Create a single mixture file containing all the observations (in this case, 398 + 394 = 792 observations), and an associated `*.ctl` file.

(iv) Use SPAM to estimate the mixture of the combined sample. This is an estimate, under the null model, of the common mixture from which all samples are drawn. Record the value listed as *log-likelihood* (line 4) in the resulting `*.est` file. This is the support of observations under the null model.

(v) SPAM gives us the logarithm of the likelihood, not the likelihood. Rather than exponentiating, we can conduct all the calculations on the natural logarithm scale. The value of interest, the logarithm of the likelihood ratio, is

equal to the support under the general model (from step (ii)) minus the support under the null model (from step (iv)). This value will be >= 0 since the Likelihood(data | general model) >= Likelihood(data | null model). Larger values are evidence against the null hypothesis of equality of all M mixtures. This value is the logarithm of equation 3 in Reynolds and Templin (in review).

From the `*.est` files excerpted in Appendix 5, the observed log (likelihood ratio) = (support | general model) – (support | null model) = -4530.58 – (-4543.46) = 12.88.

All of the required calculations up to this point are supported by SPAM 3.2. The `*.ctl` files in steps (i) and (iv) should use identical values for GPA and the other convergence tolerance criteria.

Calculating the Null Reference Distribution

Assessing the significance of the observed likelihood ratio requires knowing the distribution of likelihood ratios expected under the null hypothesis of mixture equality. If one reasonably expects all populations in the baseline to contribute to all of the mixture samples under comparison (so the expected values of the mixture components are all nonzero, or only a small percentage are zero), then standard likelihood ratio theory holds. In this case you can approximate the null reference distribution by the χ^2 distribution with (J-1)*(M-1) degrees of freedom (J = number of baseline populations, M = number of mixture samples being compared) (Stuart et al., 1999). Otherwise, one can use SPAM 3.5 to approximate the null reference distribution through Monte Carlo simulation (Reynolds and Templin, in review) as shown below.

The `*.est` file from step (iv) above contains the mixture estimate under the null model (see Appendix 4). What we need to know is the distribution when one takes, in this case, two samples of sizes 398 and 394 from this common mixture and calculates the likelihood ratio (or log-likelihood ratio) of the null and general models. Having SPAM 3.5 repeatedly simulate such samples and the necessary likelihoods lets us approximate the null reference distribution. In general, M mixture samples will require M + 1 separate *simulation* runs of SPAM 3.5. Baseline resampling should be turned off so that the simulations produce samples from the same null reference distribution. If the baseline resampling is used, then the resulting sampled likelihood-ratios will each be drawn from a slightly different null distribution. This will lead to overdispersion in the null reference distribution and may cause over-estimation of the p-value.

(i) Create a `*.ctl` file to simulate R samples of 398 + 394 = 792 observations from the mixture given in the `*.est` file from step (iv) above. R should be on the order of 1000 – 5000 (Davison and Hinkley, 1997). Set `PRINT LIKELIHOOD :True` in the `*Options` section so that the `*.rlk` output file

is produced. Use the ESTIMATE label in the *Regions section to enter the estimated mixture contribution from each baseline population (from the *.est file created in step (iv) above). See Appendix 6 (page 55) for example *.ctl files for this step and step (iii) below. GPA should be set rather high to guarantee that the realized likelihood value for each simulation is very near the maximum likelihood value (95% - 98%). The maximum number of iterations should also be set high (~1000) to allow sufficient searching for each simulation.

Run SPAM 3.5 using this *.ctl file. It will create a *.rlk file whose first column lists, for each simulated mixture of 792 observations, the support for the observations under the null model.
The first few lines of this *.rlk file were:

```
-4397.2      -4405.5       -200    658967    383758
–4424.8      -4433.9     383758    485374     49810
–4429.6      -4443.6      49810    206221    691902
```

The columns are described in the Output File section (page 15) and in Appendix 7 (page 57).

Check the *.log file to make sure each simulation converged by the GPA criterion. Only adequately converged simulations should be used in approximating the null reference distributions.

(ii) Now create a series of M simulation *.ctl files (one for each of the original mixture samples). Use the settings from step (i) with the following changes. For the first *.ctl file, set the simulation sample size to the size of the first sample, in this case 398, set BEFORE: 0 and AFTER: 394. For the second *.ctl file, set the simulation sample size to the size of the second sample, 394, set BEFORE: 398 and AFTER: 0. See Appendix 6, sections 2 and 3 (page 55) for examples.

In general, put the M mixture samples into some sequence (it doesn't matter how you order them). The *.ctl file for the k^{th} uses the settings from step (i) with the simulation sample size set to the size of the k^{th} mixture sample, BEFORE set to the sum of the sample sizes for the first k-1 mixture samples, and AFTER set to the sum of the remaining M-k sample sizes.

(iii) Run the M simulation *.ctl files in SPAM 3.5, creating M *.rlk files.
The first few lines of the *.rlk file associated with the "Ward's Cove" partition simulations were:

```
-2234.1      -2243.8         -200    658967    383758
-2273.6      -2284.4       383758    485374     49810
-2248.0      -2259.6        49810    206221    691902
```

The first few lines of the *.rlk file associated with the "Salamatof Seafoods" partition simulations were:

```
-2154.8      -2161.7          -200   658967    383758
-2136.4      -2149.5        383758   485374     49810
-2173.8      -2183.9         49810   206221    691902
```

Check the *.log file associated with each *.rlk file to make sure each simulation converged by the GPA criterion. Only adequately converged simulations should be used in approximating the null reference distributions. If a simulation is found not to converge, then it should be removed FROM ALL OF THE SYNCHRONIZED *.rlk files. That is, if the second simulation doesn't converge for one of the *.log files in step (i) or step (iii), then the second row of results should be removed from ALL the *.rlk files before calculating the null reference distribution approximation.

(iv) Sum the first value in the first column in each of the M different *.rlk files created in step (iii); this value is the support for the first set of simulated observations under the general model. Sum the second value in the first column of the M different *.rlk files created in step (iii); this value is the support for the second set of simulated observations under the general model. Continue this process, calculating the support for each of the R sets of simulated observations under the general model:
-2234.1 + -2154.8 = -4388.9 for the first Monte Carlo simulation,
-2273.6 + -2136.4 = -4410.0 for the second Monte Carlo simulation, etc.

(v) Subtract the support for the first set of simulated values under the null model (that is, the first value in the first column of the *.rlk file from step (ii)) from the support for the first set of simulated observations under the general model (calculated in step (v)). This gives one random observation from the null reference distribution. Subtract the support for the second set of simulated values under the null model (that is, the second value in the first column of the *.rlk file from step (ii)) from the support for the second set of simulated observations under the general model. This gives a second random observation from the null reference distribution. Repeat this process to generate all R random observations from the null reference distribution. For example, the log of the likelikelihood ratio is
-4388.9 - -4397.2 = 8.3 for the first Monte Carlo simulation,
-4410.0 - -4424.8 = 14.8 for the second Monte Carlo simulation, etc.

These values should all be >=0. If they aren't, either the simulations are not being properly synchronized between step (i) and step (ii) or the simulations are not all converging properly. Check each of the *.ctl files to make sure the same random number seed, GPA value, and other convergence tolerance values were used. Also check both *.log files to make sure all the

simulations converged using the GPA criterion. If some did not, they should be discarded as they might not have stopped sufficiently close to the maximum likelihood value.

Most runs will stop very near, but not exactly at, the maximum of the likelihood surface (the value of GPA determines how near). If a mixture is just as likely under the full baseline model as under the reduced baseline model, then very small differences in the attained GPA's at the final estimate of the likelihood maximum under the two models may produce very small numerical differences in the likelihood values. This may produce very small negative log likelihood ratios (~-0.01 or so) that in 'fact' are zeros. They should be considered zeros in the null reference distribution.

(vi) Count how many of the R random log likelihood ratios are equal to or larger than the observed log likelihood ratio; call this Q. The approximate Monte Carlo *P value* for the likelihood ratio test is (1 + Q) / (1 + R) (Davison and Hinkley 1997), which in this case is (1+178)/(1+1000) = 0.1788.

The approximate null reference distribution fails to reject the null hypothesis that the two mixture samples come from a common mixture (Figure 3).

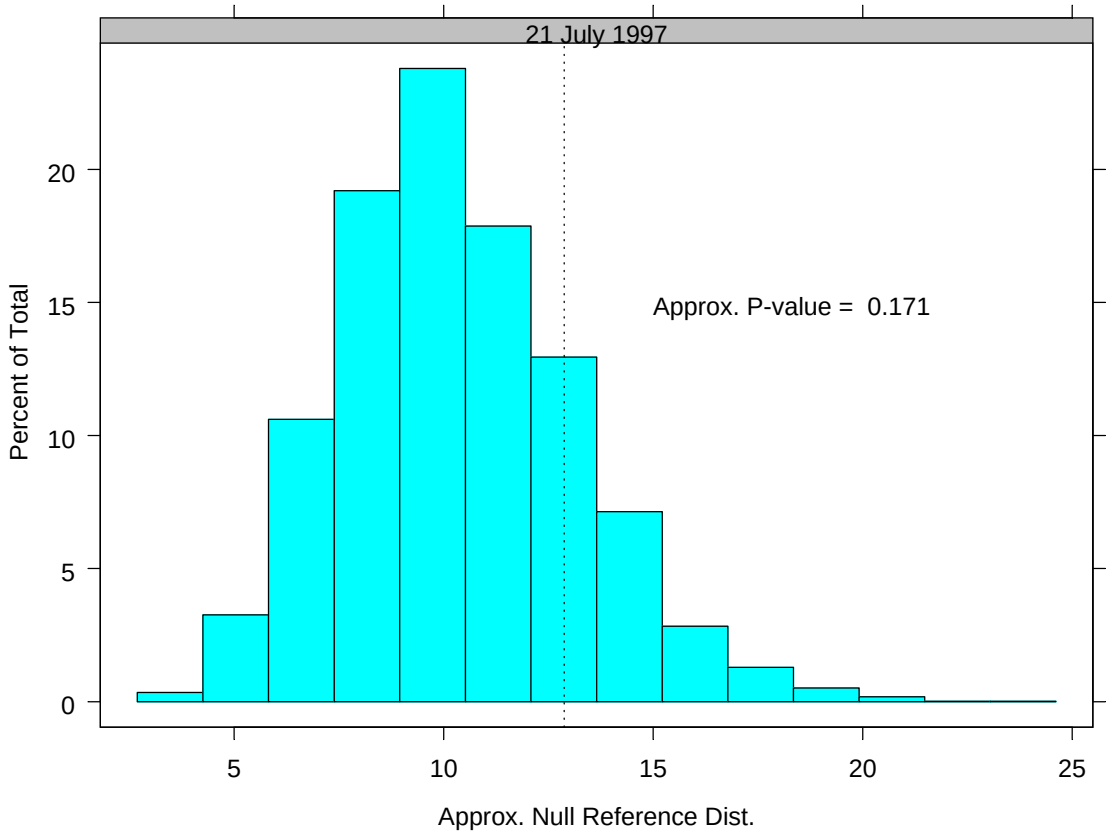


Figure 3. Approximate null reference distribution for the conditional likelihood ratio test of *Ho: Ward's Cove sample and Salamatof Seafoods, Inc. sample are from the same mixture* versus *Ha: the samples come from two different mixtures.* The null reference distribution is approximated by 1000 Monte Carlo simulations conducted in SPAM 3.5. A vertical dashed line marks the observed likelihood ratio. There is little evidence that the samples came from different mixtures. Details regarding data collection, etc., are in Reynolds and Templin (in review). Mixture samples are the 21 July 1997 collections.

Mixture Sample Size Power Analysis

Conditional power analyses can be conducted using SPAM 3.2 or SPAM 3.5. These analyses allow the user to investigate the sample size required to detect a certain sized contribution from a given region in a specific baseline with a given confidence. While such an exercise does not utilize any of the new features in SPAM 3.5, we illustrate the process to help researches better design their mixture studies.

In face of the personnel and economic costs of collecting genetic samples from the commercial troll harvest of chinook salmon (*O. tshawytscha*) during the summer fishery in Southeast Alaska, we conducted an analysis to aid in designing the sampling plan. Chinook from anywhere on the Pacific (California north to Alaska)

could occur in the harvest, so mixed stock analysis of such harvest utilizes the coastwide chinook allozyme baseline (Teel et al. 1999). This baseline contains 254 populations partitioned into 29 regions and gives population allele frequencies for 34 genetic characters. The current analysis dropped the 2 Eastern Russia populations, leaving 252 populations partitioned into 28 regions; 26 characters were employed. Managers were interested in mixture sample sizes ranging from 100 to 800 individuals. We demonstrate the method with a power analysis for the Southeast Management Quadrant of the commercial fishery (for quadrant description, see Pahlke 1995).

SPAM is used to simulate and fit mixture samples of increasing size. By repeatedly simulating samples of a specific size from a specific mixture, one can estimate the expected confidence interval and bias for a specific region contribution conditional on the known true contribution, baseline, and mixture sample size. The results are best viewed graphically (requiring some processing outside of SPAM).

(i) The user selects the relevant baseline and mixture sample sizes of interest. The user also must identify a mixture of interest. The power analysis will be highly influenced by the mixture chosen; if available, estimates from comparable previous studies should be used to identify the contributions.

The baseline for this example has already been described. Sample sizes of interest were 100, 200, …, 800. Based on historical experience and coded wire tag studies, managers defined the a mixture of interest (Scott McPherson, personal communication, May 2001) (see excerpt below). In keeping with GSI protocols for Alaskan commercial fisheries, we report only the region-level contributions. Each region-level contribution was evenly divided among the associated populations.

(ii) For each mixture sample size, N, create a *.ctl file to simulate 1000 samples of this size. Relevant portions of the *.ctl file for the N=100 sample size are shown below with the key options highlighted. Note the simulation mixture contributions given in the *Regions section. One can select to have SPAM calculate a specific bootstrap confidence interval (95% symmetric percentile bootstrap intervals, in this case) from the simulations, or save the estimates to a file for processes outside of SPAM.

```
* regions
   [level] [label]  [region]                                  [estimate]
...
     1       7       Mid and Upper Columbia, Snake Spring  :   0.00
     1       8       Upper Columbia Summer, Fall, Snake F  :   0.05
     1       9       Washington Coastal                    :   0.00
     1       10      Puget Sound                           :   0.00
```

```
        1       11      Lower Fraser                        :   0.00
        1       12      Thompson River                      :   0.00
        1       13      mid and Upper Fraser                :   0.00
        1       14      Strait of Georgia                   :   0.00
        1       15      WCVI                                :   0.05
        1       16      central BC coastal                  :   0.05
        1       17      Skeena                              :   0.025
        1       18      Nass                                :   0.025
        1       19      AK/BC Transboundary                 :   0.00
        1       20      Southern SE AK                      :   0.80
...
```

(iii) Execute each of the `*.ctl` files using SPAM 3.2 or SPAM 3.5.

The rest of the analysis requires processing outside of SPAM. The following calculations and graphics were conducted in S-Plus 2000 (MathSoft 1999).

(iv) For each sample size, gather the following features from the output files:

`*.log` convergence results for each simulated mixture.

`*.rsm` estimated mixture contributions for each population for each simulated mixture.

(v) Exclude the results of any simulations that did not adequately converge. See the discussion of convergence on the SPAM website. Generally, GPA over 95% is the preferred mode of convergence.

(vi) For each region, for each sample size, calculate the symmetric percentile bootstrap confidence interval of the desired confidence level (see UG:3.2 or Lunneborg 2000 for details) and the mean estimated contribution. Note: you can have SPAM do this for you automatically by setting the confidence interval level in the `*Parameters` section of the `*.ctl` file (see UG:3.2). Alternatively, you can process the resampled estimates outside of SPAM to see the impact of a varying the confidence levels.

(vii) Plot the lower bound of the bootstrap confidence interval estimate for each region as a function of sample size. We've separated the regions by their true contribution rate into following ranges: [0%,3%], (3%,5%], (5%,7%], (7%,10%],(10%,15%],(15%,100%] (Figure 4). This clarifies the interaction between sample size, true contribution, and 'detectability'

in terms of nonzero confidence interval lower bound.

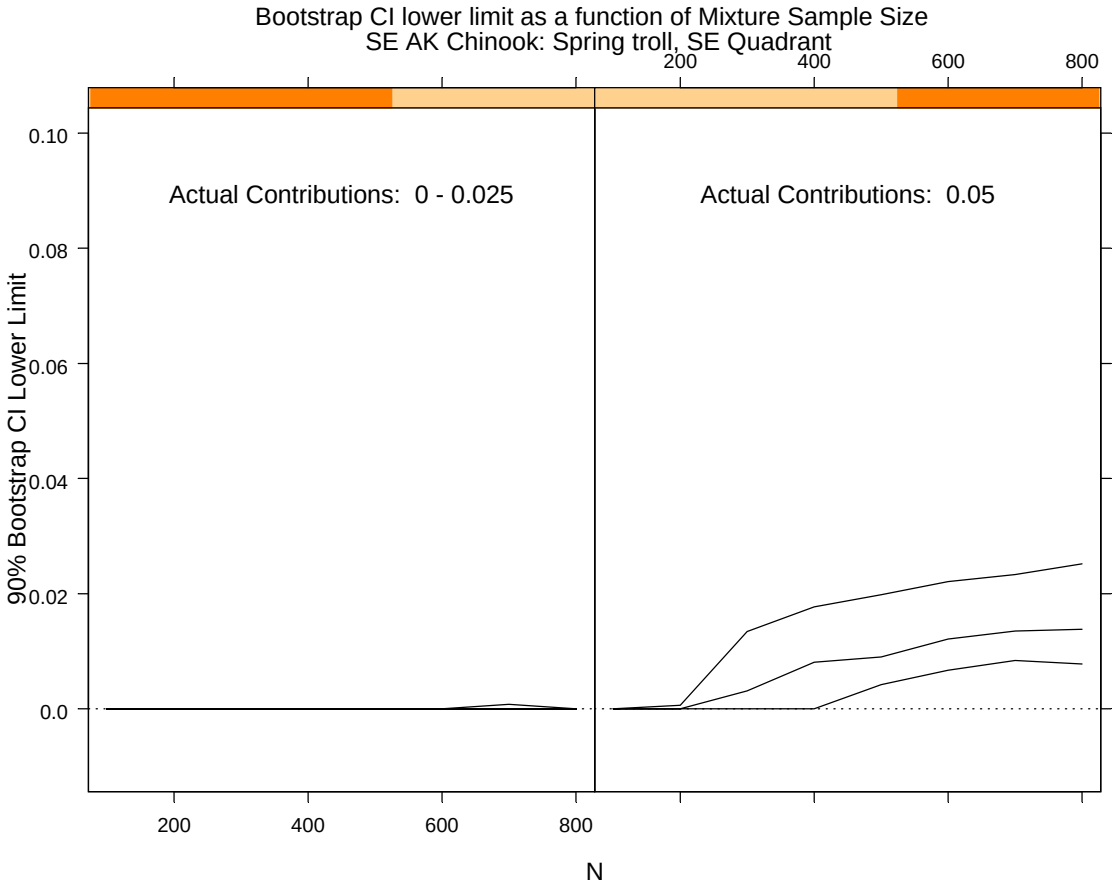


Figure 4 – Lower limit of 90% bootstrap confidence intervals for each of 28 regions in the Coastwide Chinook baseline, as a function of mixture sample size (horizontal axes) and true contribution (individual plots). Note the dashed line marking 0; a confidence interval with zero lower limit is commonly interpreted as signifying a potentially noncontributing region. Regions contributing less than 2.5% are not detected at any of the investigated sample sizes (left graph), except for a bare detection at N=700 ; regions contributing 5% require mixture samples of at least 300 –500 for detection (right graph); the region contributing 80% is always detected (graph not shown).

(viii) Calculate the bias (mean estimate – true contribution) for each region at each sample size. Plot the bias for each region as a function of the sample size.

Again, we've separated the regions by their true contribution rate (Figure 5).

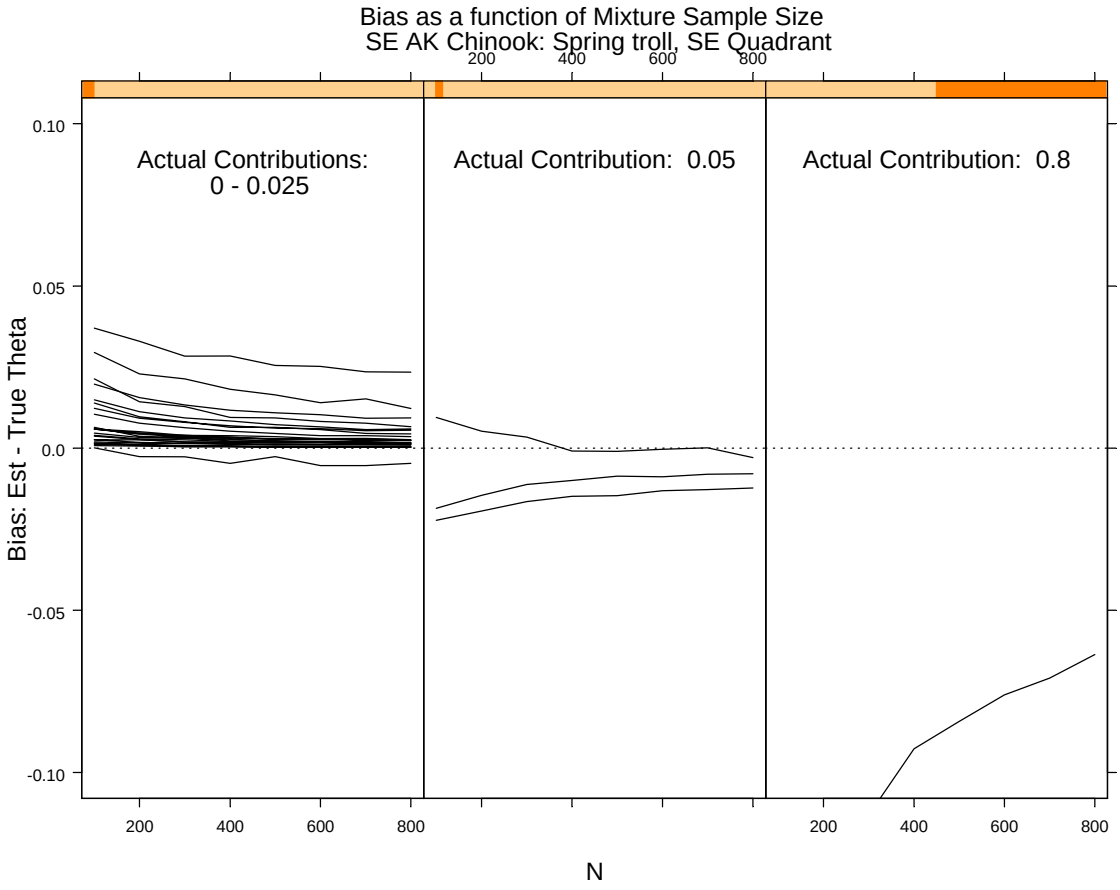
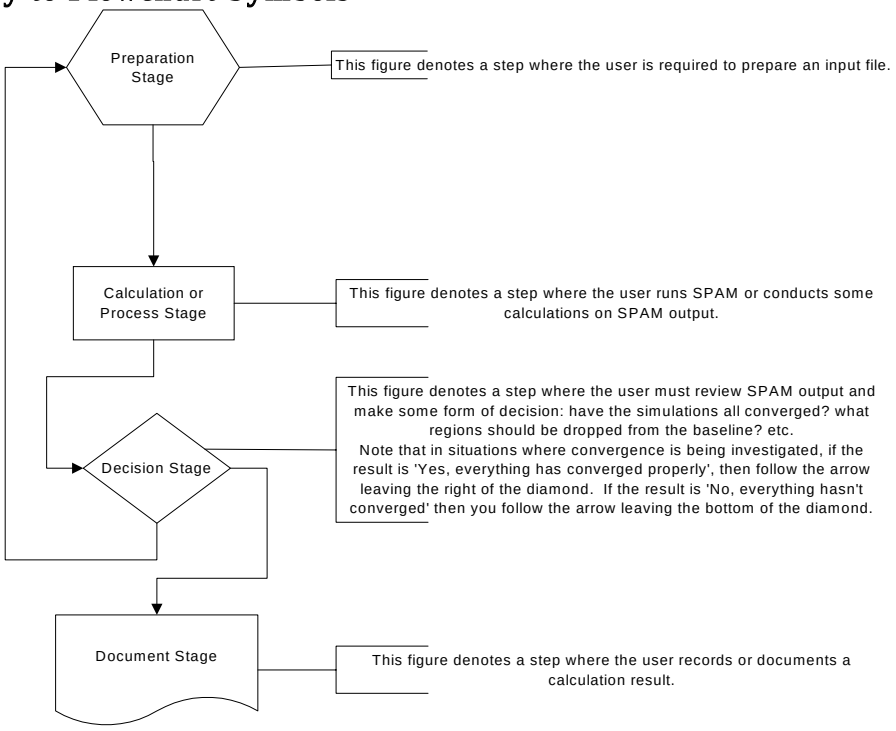


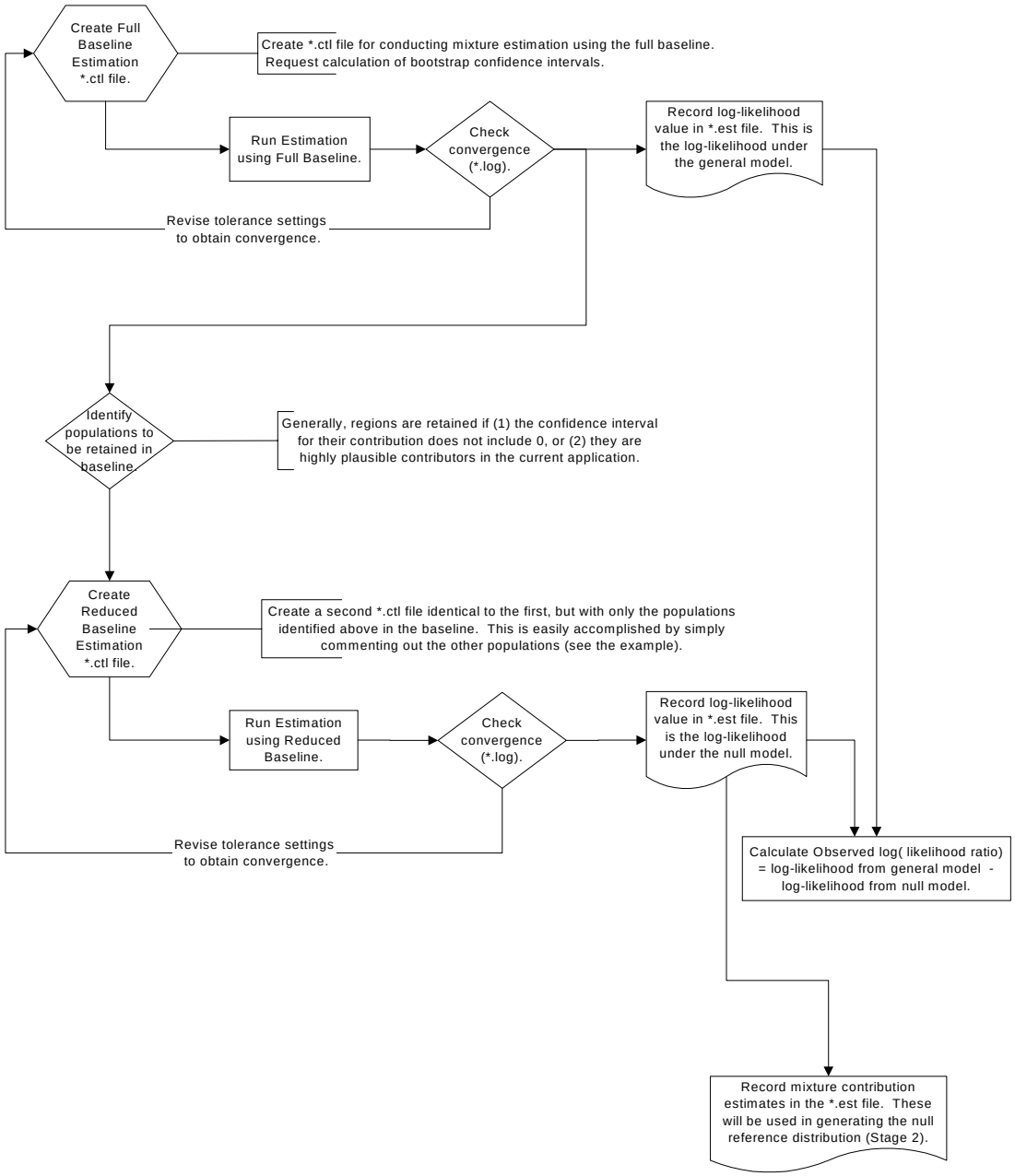
Figure 5 – Bias estimate for each of 28 regions in the Coastwide Chinook baseline, as a function of mixture sample size (horizontal axes) and true contribution (individual plots). Note the dashed line marking 0. As expected, bias magnitude declines with increasing mixture sample size. Regions contributing less than 2.5% are generally slightly overestimated (by ~ 2%) at any of the investigated sample sizes (left plot); regions contributing 5% show slight bias (middle plot); and the region contributing 80% is always underestimated, by as much as 15% for smaller mixture sample sizes (right plot). Note that even with mixture sample sizes of 800, the major contributing region can be underestimated by more than 5% (far right end of curve in far right plot). Underestimation of major contributing stocks increases as the baseline size increases (Reynolds et al. 1, in preparation).

Analysis Flowcharts

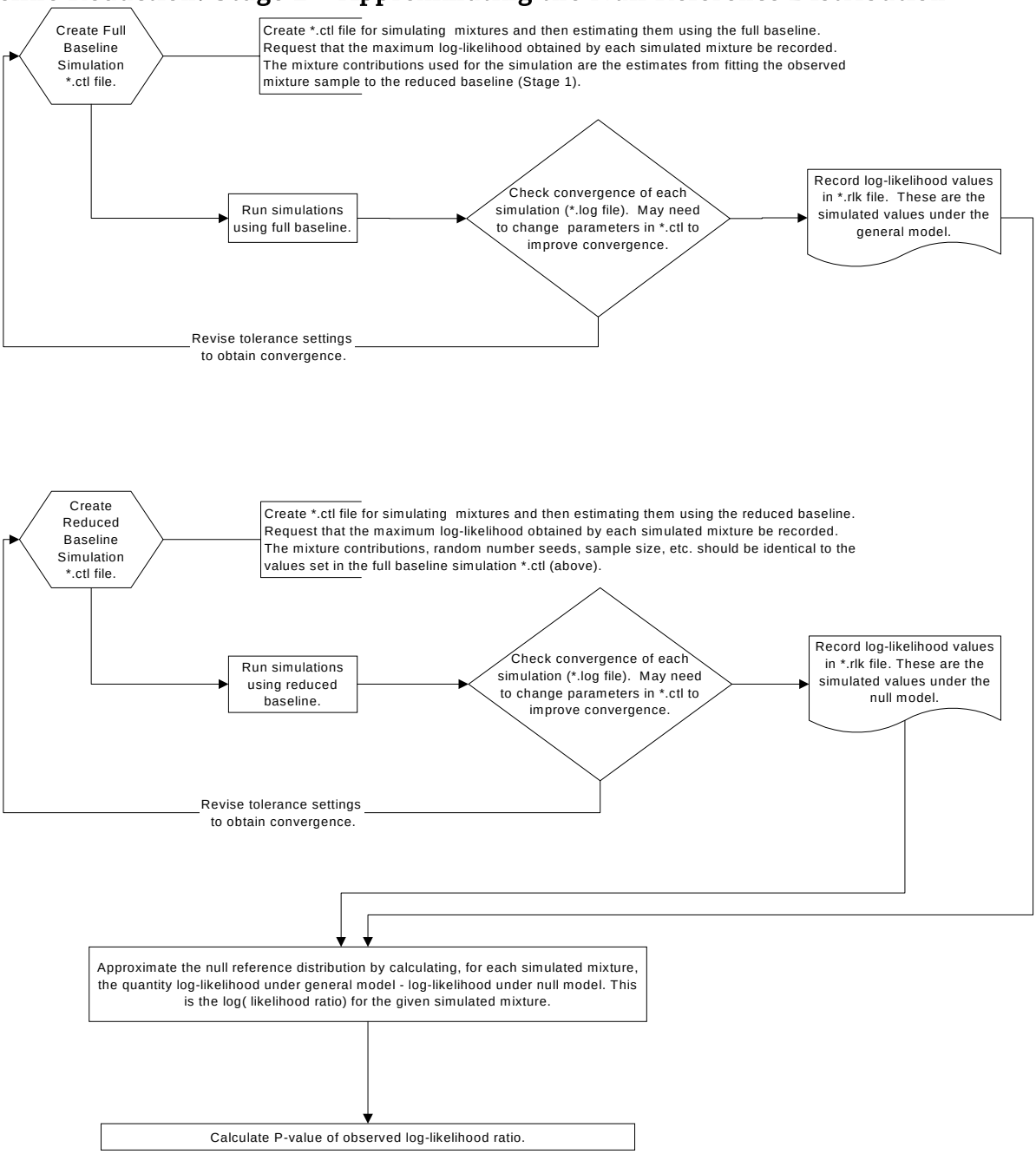
Key to Flowchart Symbols



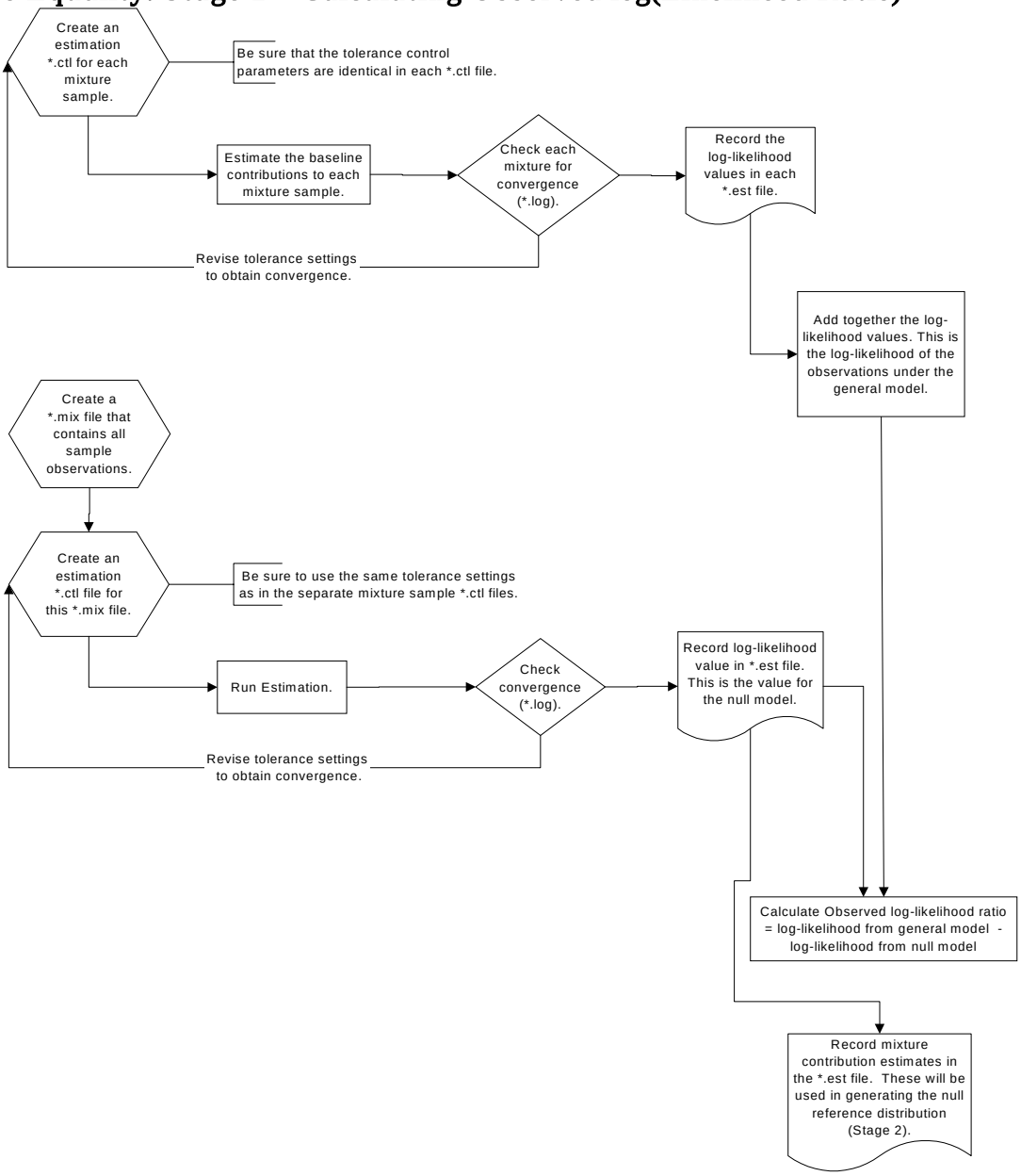
Baseline Reduction: Stage 1 – Calculating Observed log(Likelihood Ratio)



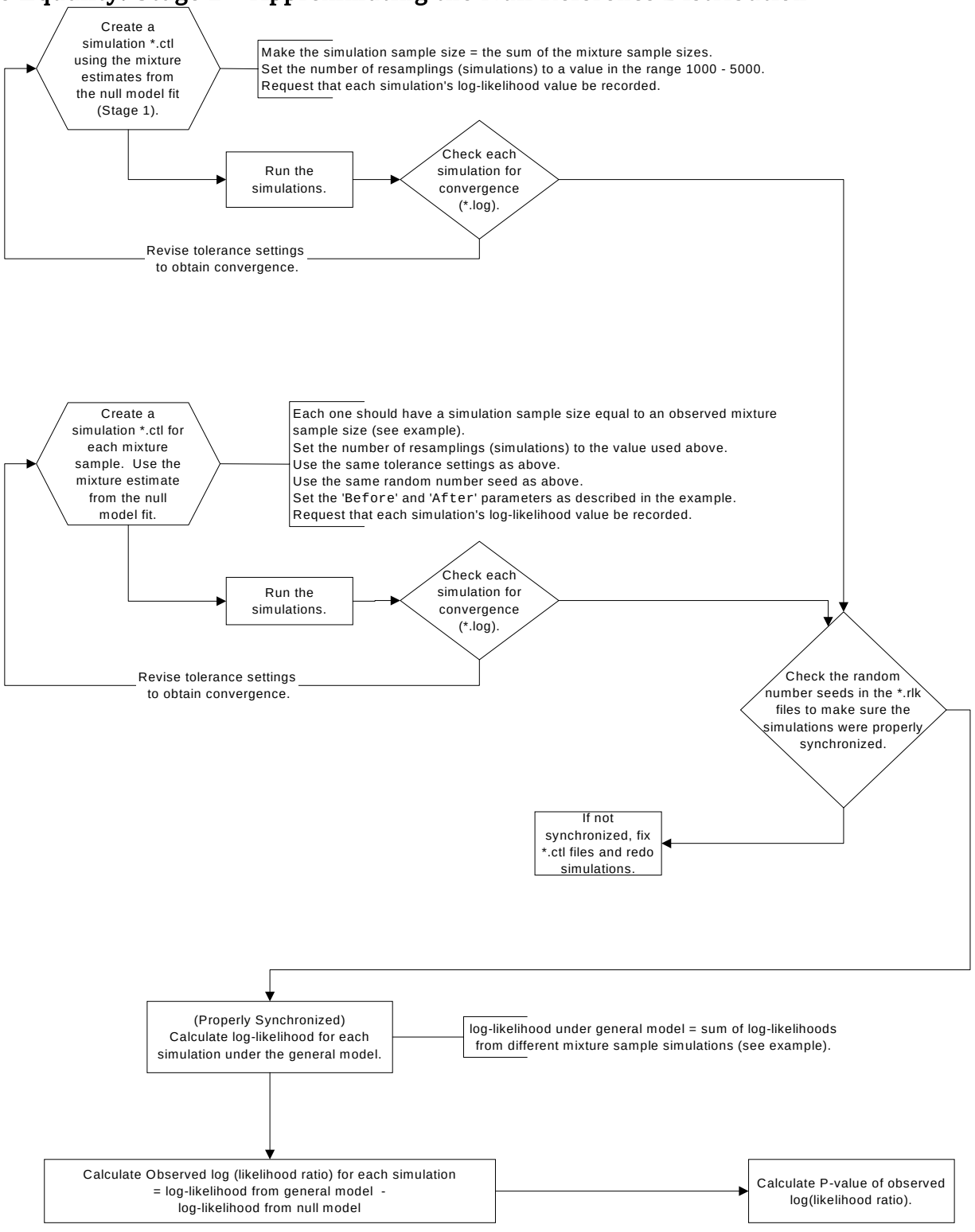
Baseline Reduction: Stage 2 – Approximating the Null Reference Distribution



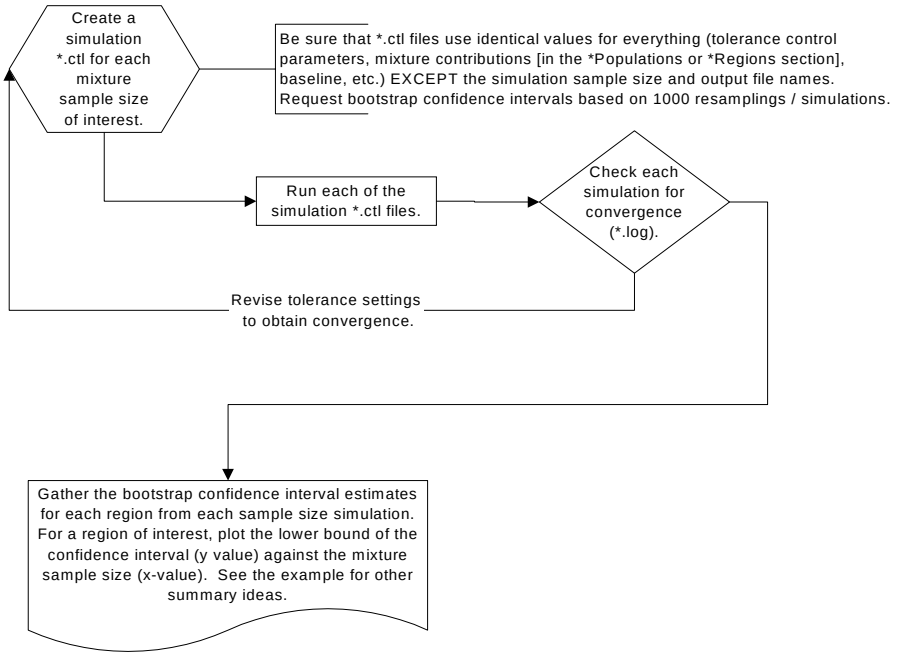
Mixture Equality: Stage 1 – Calculating Observed log(Likelihood Ratio)



Mixture Equality: Stage 2 – Approximating the Null Reference Distribution



Mixture Sample Size Investigations



Correspondence

We welcome correspondence regarding SPAM. If you would like to be included on the mailing list and receive notifications of updates, please contact us at the address below. Please report any bugs as soon as possible so we can assess the problem and make any necessary corrections to the program.

Joel Reynolds – Joel_Reynolds@fishgame.state.ak.us
Bill Templin – Bill_Templin@fishgame.state.ak.us
Lisa Seeb – Lisa_Seeb@fishgame.state.ak.u

Alaska Department of Fish and Game
Division of Commercial Fisheries
Gene Conservation Laboratory
333 Raspberry Road
Anchorage, Alaska 99518
USA

Acknowledgments

SPAM is a continually evolving product. Jerry Pella, Michele Masuda, Richard Gates, and Ed Debevec all developed major portions of the code over the last decade. These efforts arose from the suggestions and requests of researchers and users of the software. Without this feedback, criticism, and, sometimes, praise, few would be motivated to wander into this forest of Fortran. Special thanks goes to Penny Crane and Bill Templin for their patience and perseverance in testing the new features.

This work has been supported by: the State of Alaska; the U.S. Fish and Wildlife Service, Office of Subsistence Management, Federal Fisheries Monitoring Program Project FIS 01-070, "Genetic Diversity of Chinook Salmon from the Kuskokwim River"; and the State of Alaska's Southeast Sustainable Salmon Fund Project Number 45044, "Southeast Alaska Chinook Salmon Genetic Stock Identification."

Limited Warranty and Disclaimer

This software and accompanying written materials (including instructions for use) are provided “as is” without warranty of any kind. Further, Alaska Department of Fish and Game (ADF&G) does not warrant, guarantee, or make any representations regarding the use, or the results of use, of the software or written materials in terms of correctness, accuracy, reliability, currentness, or otherwise. The entire risk as to the results and performance of the software is assumed by you. If the software or written materials are defective, you, and not ADF&G or its employees, assume the entire cost of all necessary servicing, repair, or correction.

The above is the only warranty of any kind, either express or implied, including but not limited to the implied warranty of fitness for a particular purpose, that is made by ADF&G. No oral or written information or advice given by ADF&G or its employees shall create a warranty or in any way increase the scope of this warranty and you may not rely on any such information or advice.

Neither ADF&G nor anyone else who has been involved in the creation, production or delivery of this product shall be liable for any direct, indirect, consequential or incidental damages (including damages for loss of business profits, business interruption, loss of business information, and the like) arising out of the use or inability to use such product even if ADF&G has been advised of the possibility of such damages.

Use of this product for any period of time constitutes your acceptance of this agreement and subjects you to it contents.

Literature Cited

Crane, P., Templin, B., Reynolds, J., Eggers, D., and Seeb, L. 2001. Genetic Stock Identification of Southeast Alaska Chinook Salmon Fishery Catches. Final Report For LOA Project NA97FP0272, Alaska Department of Fish and Game, Division of Commercial Fisheries, 333 Raspberry Road, Anchorage, Alaska USA 99518.

Davison, A. C., and Hinkley, D. V. 1997. *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge, U.K.

Edwards, A. E. 1992. *Likelihood*. Cambridge University Press, Cambridge, U. K.

Lunneborg, C. E. 2000. *Data Analysis by Resampling: Concepts and Applications*. Duxbury, Pacific Grove, CA.

MathSoft, Inc. 1999. *S-Plus 2000*. Data Analysis Products Division, MathSoft, Seattle, WA.

McLachlan, G., Peel, D. 2000. *Finite Mixture Models*. Wiley & Sons, New York, New York.

Pahlke, K. A. 1995. Coded wire tagging studies of chinook salmon of the Unuk and Chickamin Rivers, Alaska, 1983 – 1993. Alaska Fishery Research Bulletin, 2: 93 - 113.

Pella, J., Masuda, M, and Nelson, S. 1996. Search algorithms for computing stock composition of a mixture from traits of individuals by maximum likelihood. U.S. Dept. of Commerce, NOAA Technical Memo. NMFS-AFSC-61, 68 p.

Reynolds, J. H. and Templin, W. D. In review. Comparing mixture estimates by parametric bootstrapping likelihood ratios. Journal of Agricultural, Biological and Environmental Statistics.

Reynolds, J. H., Crane, P. A., Templin, W. D., ??? In preparation. Improving mixed stock analysis by testing non-contributing baseline populations.

Reynolds, J. H., Templin, W. D., ??? In preparation. An improved test of the presence of rare populations in mixtures.

Ruzzante, D. E., Taggart, C. T., Lang, S., and Cook, D. 2000. Mixed-stock analysis of Atlantic cod near the Gulf of St. Lawrence based on microsatellite DNA. Ecological Applications, 10: 1090-1109.

Ryan, T. P. 1997. *Modern Regression Methods*. Wiley & Sons, New York, New York.

Seeb, L. W., and Crane, P. A. 1999. Allozymes and mitochondrial DNA discriminate Asian and North American populations of chum salmon in mixed-stock fisheries along the south coast of the Alaska Peninsula. Transactions of the American Fisheries Society, 128: 88 – 103.

Seeb, L. W., Habicht, C., Templin, W. D., Tarbox, K. E., Davis, R. Z., Brannian, L. K., Seeb, J. E. 2000. Genetic diversity of sockeye salmon of Cook Inlet, Alaska, and its application to managment of populations affected by the *Exxon Valdez* Oil Spill. Transactions of the American Fisheries Society, 129: 1223-1249.

Seidel, W., Mosler, K., and Alker, M. 2000. A cautionary note on likelihood ratio tests in mixture models. Annals of the Institute of Statistical Mathematics. 52(3): 481-487.

Stuart, A., Ord, J. K., and Arnold, S. (1999), *Kendall's Advanced Theory of Statistics Vol 2A: Classical Inference and the Linear Model* (6th ed.), New York, Oxford University Press.

Teel, D. J., Crane, P. A., Guthrie, C. M., Marshall, A. R., Van Doornik, D. M., Templin, W. D., Varnavskaya, N. V., and Seeb, L. W. 1999. Comprehensive allozyme database discriminates chinook salmon around the Pacific Rim. (NPAFC document 440) 25 pp. Alaska Department of Fish and Game, Division of Commercial Fisheries, 333 Raspberry Road, Anchorage, Alaska USA 99518.

Wilmot, R. L., Kondzela, C. M., Guthrie, C. M., and Masuda, M. M. 1998. Genetic stock identification of chum salmon harvested incidentally in the 1994 and 1995 Bering Sea trawl fishery. North Pacific Anadromous Fish Commission Bulletin, 1: 285-299.

Wilmot, R. L., C. M Kondzela, C. M. Guthrie III, A. Moles, J. J. Pella, M. Masuda. 2000. Origins of salmon seized from the F/V *Arctic Wind*. (NPAFC Doc.) Auke Bay Fisheries Laboratory, Alaska Fisheries Science Center, NMFS, NOAA, 11305 Glacier Highway, Juneau, AK 99801-8626. XX pp

Appendices

EXAMPLE SPAM FILES

Appendix 1: Example Control Files for Testing sufficiency of a reduced baseline.

1. Example control file (excerpt) for fitting the general model using the full baseline.

```
* Estimate:  SE. AK. Chinook 1999 Opening 1, Full Baseline

...
* control parameters
  number of stocks in analysis     : 252
  number of characters in analysis :  26
  number of bootstrap resamplings  : 100
  maximum number of genotypes      : 520
  maximum number of classes        : 100
  maximum number of iterations     : 2000
  maximum number of missing loci   :   4
  estimate tolerance               : .1E-5
  likelihood tolerance             : .1E-15
  genotype tolerance               : .1E-40
  algorithm tolerance              : .1E-3
  guaranteed percent of maximum    :  90
  random seed                      : -4533880
...

* Populations used in analysis
   [id #]          [population]                          [file]         [lev1]
   1             Mokelumne and Nimbus Hatcheries fall :  kMoaNHa.frq  :       1
   2                          Merced Hatchery fall :  kMerHat.frq  :       1
   3                        Feather Hatchery fall :  kFeaHat.frq  :       1
   4                      Feather Hatchery spring : kFeaHat1.frq  :       1
   5                        Coleman Hatchery fall :  kColeHa.frq  :       1
   6                  Upper Sacramento River winter :  kUSacra.frq  :       1
   7                             Mattole River fall :  kMattoR.frq  :       2
   8                           Van Duzen River fall :  kVDuzRi.frq  :       2
   9                              Salmon Creek fall :  kSalmCr.frq  :       2
...
 252                         Unalakleet River 92 93 :  kUnalaR.frq  :       27


 * regions
   [level] [label]  [region]
     1       1       Central Valley (Sp,F,W)
     1       2       California, S. Oregon coastal
     1       3       Klamath (Sp and F)
     1       4       Mid and North Oregon Coastal
     1       5       Lower Columbia Spring and Fall
     1       6       Willamette
     1       7       Mid and Upper Columbia, Snake Spring
     1       8       Upper Columbia Summer, Fall, Snake F
     1       9       Washington Coastal
     1       10      Puget Sound
     1       11      Lower Fraser
     1       12      Thompson River
     1       13      mid and Upper Fraser
...
.....1       28      Upper Canadian Yukon
```

2. Example control file (excerpt) for fitting the null model. The bootstrap confidence intervals on the region estimates from using the full baseline suggest that only regions 4, 8, 9, 12, 15, 16, and 20 contribute to the mixture. Populations in all other regions are commented out. Note that the number of stocks has been changed in the *Parameters section and that the labels on the remaining regions have been revised. The population listings have been

reordered for visual aid; reordering isn't necessary at this stage.

```
* Estimate:  SE. AK. Chinook 1999 Opening 1, Reduced Baseline

* control parameters
  number of stocks in analysis     : 78
  number of characters in analysis :  26
...

* Populations used in analysis
   [id #]          [population]                              [file]         [lev1]
  32                                 Euchre Creek fall :  kEuchCr.frq  :       1
  33            Elk River and Elk River Hatchery fall :  kERaERH.frq  :       1
...
  80                            Klickitat River summer : kKlickR5.frq  :       2
  81                              Klickitat river fall :  kKlicki.frq  :       2
...
 110                             Naselle Hatchery fall :  kNasHat.frq  :       3
 111                Wynoochee River and Hatchery fall :  kWynoRa.frq  :       3
...
 146                                Spius Creek spring :  kSpiCrs.frq  :       4
 147                               Nicola River summer :  kNicRis.frq  :       4
...
 /  1             Mokelumne and Nimbus Hatcheries fall :  kMoaNHa.frq  :       1
 /  2                               Merced Hatchery fall :  kMerHat.frq  :       1
...
/ 252                            Unalakleet River 92 93 :  kUnalaR.frq  :       27

 * regions
   [level] [label]  [region]
 \    1       1       Central Valley (Sp,F,W)
 \    1       2       California, S. Oregon coastal
 \    1       3       Klamath (Sp and F)
     1       1       Mid and North Oregon Coastal
 \    1       5       Lower Columbia Spring and Fall
 \    1       6       Willamette
 \    1       7       Mid and Upper Columbia, Snake Spring
     1       2       Upper Columbia Summer, Fall, Snake F
     1       3       Washington Coastal
 \    1       10      Puget Sound
 \    1       11      Lower Fraser
     1       4      Thompson River
 \    1       13      mid and Upper Fraser
...
```

Appendix 2: Estimation files for calculating the observed likelihood ratio in testing baseline reduction.

1. Example estimation file (excerpt) from fitting the general model using the full baseline. The log-likelihood of the observed mixture under the general model is highlighted in bold.

```
 * Estimation: SE AK Chinook Opening 1 estimate, Full Baseline

Iterations               149
Log likelihood          -2806.76
Percent of maximum      90.36
Maximum estimate update .715E-04
…
```

2. Example estimation file (excerpt) from fitting the null model using the reduced baseline. The log-likelihood of the observed mixture under the null model is highlighted in bold. The population contribution estimates from this file will be used in generating the null reference distribution (Appendix 3).

```
 * Estimation: Opening#1 estimate, subsample mixture Sub 12 Jan 01 MC test

Iterations                49
Log likelihood          -2815.86
Percent of maximum      90.63
Maximum estimate update .110E-03

…
     Population                     Estimate Score
  1  Euchre Creek fall               .0000    -49.1
  2  Elk River and Elk River Hatche  .0000    -78.4
  3  Sixes River fall                .0000    -80.8
  4  South Fork Coquille River fall  .0080       .0
  5  Coquille River and Bandon Hatc  .0000    -17.1
  6  Millicoma River fall            .0464       .0
  …
```

Appendix 3: Control files for simulating the null reference distribution for testing a baseline reduction.

1. Control file for simulating mixtures under the null model (reduced baseline), then fitting them using the general model (full baseline). Key settings are in bold. Note that the random number seed and number of simulations must be identical to those in the control file below (fitting to the reduced model) in order to guarantee that the same mixture samples are simulated in both situations.

   The population estimates are from fitting the observed mixture to the null model (the second *.est file in Appendix 2), with a value of zero being assigned to any population in a region not included in the null model's baseline.

```
* Simulate:  SE AK Chinook 1999 Opening 1. Simulate mixtures from null
\           model fit but fit to general model (full baseline).

* options selected for optimization
  use IRLS algorithm                     :  off
  fixed baseline frequencies             :  off
  …
  compute likelihood ratio               :  on
  resample using mixture frequencies     :  on
  resample using baseline frequencies    :  off

* control parameters
  number of stocks in analysis     : 252
  number of characters in analysis :  26
  number of bootstrap resamplings  : 1000
  simulation sample size           :  328
  maximum number of genotypes      : 520
  maximum number of classes        : 100
  maximum number of iterations     : 2000
  maximum number of missing loci   :   4
  estimate tolerance               : .1E-9
  likelihood tolerance             : .1E-15
  genotype tolerance               : .1E-40
  algorithm tolerance              : .1E-3
  guaranteed percent of maximum    :  99
  random seed                      : -4533880
…
* Populations used in analysis
   [id #]          [population]                              [file]         [lev1] [estimate]
   1             Mokelumne and Nimbus Hatcheries fall :  kMoaNHa.frq  :       1   0
   2                          Merced Hatchery fall :  kMerHat.frq  :       1   0
   3                        Feather Hatchery fall :  kFeaHat.frq  :       1   0
   …
  32                             Euchre Creek fall :  kEuchCr.frq  :       4   0
  33           Elk River and Elk River Hatchery fall :  kERaERH.frq  :       4   0
  34                             Sixes River fall :  kSixRiv.frq  :       4   0
  35                 South Fork Coquille River fall :  kSFCoqu.frq  :       4   .008
  36          Coquille River and Bandon Hatchery fall :  kCoRaBH.frq  :       4   0
  37                          Millicoma River fall :  kMillic.frq  :       4   .0464
…
```

2. Control file for simulating mixtures under the null model (reduced baseline), then fitting them using the null model. Key settings are in bold. Note that the random number seed and number of simulations must be identical to those in the control file above (fitting to the general model) in order to guarantee that the same mixture samples are simulated in both situations.

The population estimates are from fitting the observed mixture to the null model (the second *.est file in Appendix 2).

```
* Simulate:  SE AK Chinook 1999 Opening 1. Simulate mixtures from null
\            model fit and fit to null model (reduced baseline).

* options selected for optimization
  use IRLS algorithm                    :  off
  fixed baseline frequencies            :  off
  …
  compute likelihood ratio              :  on
  resample using mixture frequencies    :  on
  resample using baseline frequencies   :  off

* control parameters
  number of stocks in analysis     :  78
  number of characters in analysis :  26
  number of bootstrap resamplings  : 1000
  simulation sample size           :  328
  maximum number of genotypes      : 520
  maximum number of classes        : 100
  maximum number of iterations     : 2000
  maximum number of missing loci   :   4
  estimate tolerance               : .1E-9
  likelihood tolerance             : .1E-15
  genotype tolerance               : .1E-40
  algorithm tolerance              : .1E-3
  guaranteed percent of maximum    :  99
  random seed                      : -4533880
…
* Populations used in analysis
   [id #]          [population]                              [file]         [lev1] [estimate]
  32                                Euchre Creek fall :  kEuchCr.frq  :       1   0
  33            Elk River and Elk River Hatchery fall :  kERaERH.frq  :       1   0
  34                                  Sixes River fall :  kSixRiv.frq  :       1   0
  35                   South Fork Coquille River fall :  kSFCoqu.frq  :       1   .008
  36          Coquille River and Bandon Hatchery fall :  kCoRaBH.frq  :       1   0
  37                              Millicoma River fall :  kMillic.frq  :       1   .0464
   …
```

Appendix 4: Control files for calculating observed likelihood ratio in testing Mixture Equality.

1. Control file for calculating log-likelihood of Ward's Cove sample under general model (samples come from different mixtures) (excerpt).

```
* estimation: Central District Drift Fishery 1997 Opening 2: harvest mixture sample from
\             Ward's Cove.  Fitting to general model to get log-likelihood of observed
\             mixture sample.

* options selected for optimization
  use IRLS algorithm                     :  off
  fixed baseline frequencies             :  off
  print mixture file                     :  off
  print baseline relative frequencies    :  off
  print conditional genotype prob.       :  off
  print conditional population prob.     :  off
  compute likelihood confidence intervals :  off
  compute infinitesimal jacknife std. dev.:  off
  resample using mixture frequencies     :  off
  resample using baseline frequencies    :  off

* control parameters
  number of stocks in analysis     :  44
  number of characters in analysis :  29
  number of bootstrap resamplings  : 0
  maximum number of genotypes      : 400
  maximum number of classes        :  40
  maximum number of iterations     : 1000
  maximum number of missing loci   :  2
  estimate tolerance               : .1E-8
  likelihood tolerance             : .1E-9
  genotype tolerance               : .1E-45
  algorithm tolerance              : .1E-9
  guaranteed percent of maximum    :  98
…
* Populations used in analysis
    [id #]  [population]             [file]        [lev1]  [lev2]
    1   Byers Lake                 : sByersL.frq  :  1    :  1
…
   61   Tustumena Lake             : sTustum.frq  :  6    :  4
…
```

2. The control file for calculating the log-likelihood of the Salamatof Seafoods, Inc. sample under the general model (samples come from different mixtures) differs only in the mixture file reference, as does the control file for estimating the log-likelihood under null model (pooled mixture samples). In this latter case, the referenced mixture file contains the combined samples.

Appendix 5: Estimation files for calculating the observed likelihood ratio for testing mixture equality (excerpts).

1 Observed log likelihood under general model of samples originating from different mixtures.

Ward's Cove sample log-likelihood:

```
  * Estimation: Central District Drift Fishery 975

  Iterations               86
  Log likelihood         -2251.04
```

Salamatof Seafoods, Inc. sample log-likelihood:

```
* Estimation: Central District Drift Fishery 97 rep 2

  Iterations              265
  Log likelihood         -2279.54
...
```

Log likelihood under general model: -2251.04 + -2279.54 = -4530.58

2 Observed log likelihood under null model.

```
* Estimation: Central District Drift Fishery 1997 Opening 2, Pooled Processors Samples

  Iterations               50
  Log likelihood         -4543.46
...

       Population                    Estimate  Std.Err.   CV     Score
   1  Byers Lake                     .0387     .0291    .75         .0
   2  Stephan Lake                   .0000     .0000    .00      -137.9
   3  Larson Lake                    .0001     .0019    22.        -3.3
   4  Birch Creek                    .0000     .0000    .00       -89.0
   5  Red Shirt Lake                 .0000     .0000    .00      -107.9
...
  43  Glacier Flat/Nikolai creeks    .0073     .0211    2.9         .0
  44  Tustumena Lake                 .0148     .0291    2.0         .0
       Unknown                       .0038
```

Resulting observed log (likelihood ratio) =
(support | general model) – (support | null model) = -4530.58 – (-4543.46) = 12.88.

Note that the null model mixture estimates are used in the Monte Carlo simulation *.ctl files (Appendix 6).

Appendix 6: Control files for simulating the null reference distribution when testing mixture equality (excerpts).

1. Control file for simulating mixtures under the null model (all samples come from a common mixture), then fitting them using the null model (step (i) on page 27). Key settings are in bold. Note that the random number seed must be identical to those in the control file below (fitting to the general model) in order to guarantee that the same mixture samples are simulated in both situations.

   The population estimates are from fitting the observed mixture to the null model (the second *.est file in Appendix 5). Though baselines resampling is used, in general one should consider the discussion of this issue in the example.

```
* Simulation: Central District Drift Fishery 1997 Opening 2 – simulating
\             samples and fitting under the null model (common source mixture)
\             in order to approximate the null reference distribution.

* options selected for optimization
  print likelihoods                     :  on
  resample using mixture frequencies    :  on
  resample using baseline frequencies   :  on
  print bootstrap estimates             :  on

* control parameters
  number of stocks in analysis     :  44
  number of characters in analysis :  29
  number of bootstrap resamplings  : 5000
  maximum number of genotypes      : 400
  maximum number of classes        :  40
  maximum number of iterations     : 1000
  maximum number of missing loci   :  2
  estimate tolerance               : .1E-8
  likelihood tolerance             : .1E-11
  genotype tolerance               : .1E-45
  algorithm tolerance              : .1E-9
  guaranteed percent of maximum    :  98
  simulation sample size           : 792
  confidence intervals             :  90
  random seed                      : 100000

…

* Populations used in analysis
    [id #]  [population]            [file]       [lev1] [ESTIMATE]
    1   Byers Lake                 : sByersL.frq  :  1  : .0390
…
   61   Tustumena Lake             : sTustum.frq  :  6  : .0149

…

* run
```

2. Control file for simulating the "Ward's Cove" samples and calculating their support under the general model where samples come from possibly different mixtures (step (iii) on page 27). Key settings are boldfaced. Note the use of the BEFORE and AFTER options in this and the next *.ctl file. Though baselines resampling is used, in general one should consider the discussion of this issue in the example.

```
* Simulation: Central District Drift Fishery 1997 Opening 2 – simulating
\             samples and fitting under the general model (different source mixtures)
\             in order to approximate the null reference distribution.
\             Ward's Cove Samples.

* options selected for optimization
  print likelihoods                      : on
  resample using mixture frequencies     :  on
  resample using baseline frequencies    :  on
  print bootstrap estimates              :  on

* control parameters
  number of stocks in analysis     :  44
  number of characters in analysis :  29
  number of bootstrap resamplings  : 5000
...
  simulation sample size           : 398
  number of null observations BEFORE: 0
  number of null observations AFTER: 394
  confidence intervals             :  90
  random seed                      : 100000

* Populations used in analysis
    [id #]  [population]             [file]       [lev1] [ESTIMATE]
    1   Byers Lake                 : sByersL.frq  :  1  : .0390
…
   61   Tustumena Lake             : sTustum.frq  :  6  : .0149
…
```

3. Control file for simulating the "Salamatof Seafoods, Inc." samples and calculating their support under the general model where samples come from possibly different mixtures (step (iii) on page 27). Key settings are boldfaced. Note the use of the BEFORE and AFTER options in this and the previous *.ctl file. Though baselines resampling is used, in general one should consider the discussion of this issue in the example.

```
* Simulation: Central District Drift Fishery 1997 Opening 2 – simulating
\             samples and fitting under the general model (different source mixtures)
\             in order to approximate the null reference distribution.
\             Salamatof Seafoods, Inc. Samples

* options selected for optimization
  print likelihoods                      : on
  resample using mixture frequencies     :  on
  resample using baseline frequencies    :  on
  print bootstrap estimates              :  on

* control parameters
  number of stocks in analysis     :  44
  number of characters in analysis :  29
  number of bootstrap resamplings  : 5000
...
  simulation sample size           : 394
  number of null observations BEFORE: 398
  number of null observations AFTER: 0
  confidence intervals             :  90
  random seed                      : 100000

* Populations used in analysis
    [id #]  [population]             [file]       [lev1] [ESTIMATE]
    1   Byers Lake                 : sByersL.frq  :  1  : .0390
…
   61   Tustumena Lake             : sTustum.frq  :  6  : .0149
…
```

Appendix 7: Monte Carlo Simulation Output for testing mixture equality: `*.rlk` files (excerpts).

1. Support of Monte Carlo simulations under null model that all observations come from a common mixture (column 1). The first simulation has a support of –4397.2, the second a support of –4424.8, etc. The values in column 2 are currently not utilized. As described earlier (page 15), the values in columns three through five are the current seed for the random number generator at different points in the Monte Carlo simulation process. This information can be used as a check on the synchronization of simulations across SPAM 3.5 calls.

```
-4397.2      -4405.5       -200    658967    383758
–4424.8      -4433.9     383758    485374     49810
–4429.6      -4443.6      49810    206221    691902...
```

2. Support for partitioned Monte Carlo simulations – components of support under general model (samples come from possibly different mixtures): Ward's Cove component.

```
-2234.1      -2243.8        -200  658967    383758
-2273.6      -2284.4      383758  485374     49810
-2248.0      -2259.6       49810  206221    691902...
```

3. Support for partitioned Monte Carlo simulations – components of support under general model: Salamatof Seafoods component.

```
-2154.8      -2161.7        -200  658967    383758
-2136.4      -2149.5      383758  485374     49810
-2173.8      -2183.9       49810  206221    691902...
```