Western Alaska Salmon Stock Identification Program **Technical Document**:[1] 3

Title: Estimating small proportions **Version :** 1.0

Authors: J. Jasper, C. Habicht, W. Templin

Date: September 10, 2009

Introduction

High statistical power is necessary when attempting to estimate the contribution of stocks which contribute at small proportion to the mixture (e.g. <0.05) in order to detect the presence of these stocks. Along with detecting presence/absence, obtaining unbiased estimates is also important. In other words, we are looking for methods to increase the accuracy and precision of estimates of stocks in mixtures that appear in low proportions. Generally, statistical power is generated through increasing sample sizes within strata; however this is often not an option.

One way to increase power, when faced with several samples of fixed sample size, is make use of a stratified design. However, stratifying means that we must increase the scope of our estimate. For example, consider the contribution made by North Peninsula stocks of sockeye salmon to the harvest in the Ugashik District over a three-year period. The current sampling plan for this district identifies four temporal strata per year. We could provide a separate estimate for each temporal stratum, a separate estimate for each year, or a single estimate over all years and strata. As we broaden the scope of the estimate, we improve precision and accuracy. Our purpose here is to demonstrate this improvement with a simulated example. The North Peninsula/Ugashik scenario was chosen for this example because there is much genetic overlap between stocks of sockeye salmon spawning within the North Peninsula and Ugashik districts.

[1] This document serves as a record of communication between the Alaska Department of Fish and Game Commercial Fisheries Division and the Western Alaska Salmon Stock Identification Program Technical Committee. As such, these documents serve diverse ad hoc information purposes and may contain basic, uninterpreted data. The contents of this document have not been subjected to review and should not be cited or distributed without the permission of the authors or the Commercial Fisheries Division.

Methods

In the Ugashik District in 2008, the estimated composition of the commercial catch of sockeye salmon in all four strata was consistently 85-90% Ugashik fish, 10-15% Egegik fish, and minor contributions from other stocks (Tim Baker, personnel communication). The total harvest in 2008 ranged from 69,000 to 446,000 fish with an average of 250,000 and a standard deviation of 154,000. We assumed 2008 was a typical fishing season in the Ugashik District and composition and harvest numbers from this year were used as a model for this simulation.

For each of three years, mixtures for four temporal strata were generated in proportions similar to those estimated in the Ugashik District in 2008, with the contribution from North Peninsula set at 1.1% for all samples (Table 1). Each mixture was given a sample size of N=380. To generate each mixture, fish were removed from baseline populations and the remaining baseline was used to resolve the mixture. A total of 3 (years) X 4 (strata/year) = 12 (strata) mixtures were generated. Harvest for each stratum in each year was drawn from a normal distribution using the observed mean and standard deviation from 2008 (Table 2).

All mixtures were analyzed with an implementation of the Bayesian mixture model (Pella and Masuda 2001) in WinBUGS (Spiegelhalter et al. 2006) using a flat prior. One chain was run for 25,000 iterations, burning the initial 5,000. The resulting posterior outputs were read into R using the CODA feature (Plummer et al. 2006). All estimates were rounded to the nearest 1/10 of 1%.

To estimate the contribution of North Peninsula fish, three levels of summaries (posterior means and 90% Bayesian confidence intervals, hereafter referred to as confidence intervals) were calculated: 1) a separate estimate for each stratum in each year; 2) a broader estimate combining all strata within each year; and 3) a single grand estimate combining all years and strata.

Summaries for each stratum in each year were calculated by simply taking the mean and quantiles of the posterior outputs. Strata were combined into yearly estimates by weighting them by their respective harvests according to the following equation:

$$p_y = \frac{\sum_{i=1}^{4} H_{y,i} p_{y,i}}{\sum_{i=1}^{4} H_{y,i}} \quad .$$

Where $H_{y,i}$ is the harvest in year y and stratum i; $p_{y,i}$ is the proportion of North Peninsula fish in year y and stratum i; and p_y is the overall proportion of North Peninsula fish in year y. To calculate confidence intervals for p_y, its distribution was estimated via Monte Carlo by re-sampling the posterior output from each of the constituent strata and applying the harvest to the draws according to the above equation.

Similarly, all years were combined by weighting the yearly proportions by the yearly total harvests.

Results

The posterior means and confidence intervals for all three levels are shown in Table 3. For the individual strata (level 1), the estimates tend to be noisy with wide confidence intervals, all of which contain zero when rounded to the nearest one-tenth of one percent. Histograms of the posterior outputs from the first year reveal distributions with large modes at or near zero and long, diffuse tails extending well beyond the mean (Figure 1).

The yearly estimates were better behaved with tighter confidence intervals, one of which excludes zero (Table 3). The posterior distribution of the first yearly proportion is bi-modal, with one mode near zero and the other mode near the true value of 1.1% (Figure 2).

The estimated grand proportion over all years is very near the true value 1.1% and the tight confidence interval excludes zero (Table 3). The posterior distribution is a very well shaped uni-modal distribution whose mode is near 1.1% (Figure 3).

Discussion

Preliminarily, these results appear to give promise to the task of accurately and precisely estimating small proportions, as long a single overall estimate is acceptable. An obvious caveat of this exercise is that there were always 4 North Peninsula fish in every 380-fish mixture, whereas in reality, this proportion would vary across samples if the fishery actually caught 1.1% North Peninsula fish. Also, we failed to fully examine the benchmark scenario of 0.0% North Peninsula fish to see if an overall estimate would exclude zero. Initial explorations show a small, but positive estimate when the true contribution is 0.0%, as is typical of MSA.

Another approach under consideration is to simply pool all the samples; not for the purpose of estimating stock proportions, but rather, for the detection of North Peninsula fish. Detection can be ascertained via confidence intervals, or possibly model selection techniques involving either Bayes factors or deviance information criteria (DIC) that has been adapted specifically towards mixture models. Establishing presence/absence of North Peninsula fish can aid in the assessment of the validity of estimates for small contributions.

A further approach is to analyze several related mixtures simultaneously in a hierarchical setting. In this framework, the prior parameters for the stock proportions would themselves be given a prior distribution that relates the stock proportions from one mixture to the stock proportions of other mixtures and to covariates. Some potential covariates include proximity of the stocks to the fishery, time of the year, magnitude of escapement, results from the Port Moller test fishery, scale patterns or age distributions, etc. These models can improve estimation for any one mixture by borrowing strength from the other mixtures and the covariates. Explorations of these techniques in the current context, as well as others, have been very promising.

Future Analyses

1. Continue the analysis with true contributions of North Peninsula fish that equal {0.00, 0.02, and 0.05}.
2. Repeat the entire analysis with example stocks that are genetically distinct.
3. Investigate Bayesian model selection techniques with respect to the presence of small contributions in large samples through the use of confidence intervals, Bayes factors, and DIC.
4. Develop hierarchical models, with covariates, using known mixtures in realistic proportions. Preceding this exploration would be the identification of covariates that improve explanation of stock proportions.
5. Replicate all analyses multiple times.

Literature Cited

Martyn Plummer and Nicky Best and Kate Cowles and Karen Vines. 2006. CODA: Convergence Diagnosis and Output Analysis for MCMC. R News 6 (1):7-11. http://CRAN.R-project.org/doc/Rnews/.

Pella, J., and M. Masuda. 2001. Bayesian methods for analysis of stock mixtures from genetic characters. Fishery Bulletin 99(1):151-167.

Spiegelhalter, D.J., A. Thomas, and N.G. Best. 1999. WinBUGS Version 1.2 User Manual. MRC Biostatistics Unit.

Technical Committee review and comments

Document 3: Estimating small proportions.

This is a good study of the tradeoffs between detail and uncertainty: the smaller the spatial/temporal scale examined, the less certain the estimate of the interception rate of the stock. It would be useful to clarify two important points. First, there are two general sources of uncertainty in these analyses: A) uncertainty in identifying stock of origin of fish in the sample from the fishery; B) uncertainty in extrapolating from the sample to the entire fishery. The second point is that uncertainty A is the only portion that improved genetic methods can address; uncertainty B is not due to a limitation of GSI but rather to inescapable statistical realities.

The authors give a good discussion of the limitations of their work. The fixed number of N. Peninsula fish in the trials means the uncertainty was underestimated, but the pattern of more accuracy when strata are collapsed still holds. Another item for consideration is the possibility of overdispersion in the data due to a variety of biological processes and difficulties in obtaining a completely random sample.

A hierarchical framework for analyses is suggested. This could be a great idea – samples from a stratum in one year could have information that could improve estimates from the same stratum in other years. However, the variable assumed to have a hierarchical structure needs careful consideration. On biological grounds, it's reasonable to expect similar fractions of a specific population will be in a fishing district each year. However, the fraction this represents of the fish in the district will vary proportionally to the abundance of the source stock and inversely with the abundance of the other stocks that also frequent the district. It may not be optimal to assume, for example, that the proportion of the catch in the Ugashik district of N. Peninsula origin fits a hierarchical model.

We'd like to see these analyses focused more closely on questions of concern to managers and resource users. The current focus of the simulations, on the ability to detect and estimate the contribution of stocks that constitute a small fraction of the catch, is useful but could be made more so. For most management concerns, I think the number of fish intercepted will be more relevant than the fraction of the catch they constitute.

For instance, those whose stocks are potentially intercepted are interested in whether the fishery is intercepting a 'large' portion of their stock. 'Large' needs to be defined in terms of its effect on the intercepted stock. Relevant simulations should focus on whether a 'large' interception can be detected and its magnitude reliably estimated. These users are also interested in reducing this interception. Thus, identifying the spatial and temporal distribution of this interception is also important.

Conversely, the concern of those participating in the interception fishery is having their fishery unnecessarily restricted. Simulations focused on the probability of estimating a 'large' interception when in fact the interception is 'small' would be most relevant.

[*Unedited comments from "Panel comments October 2009.doc" related to Technical Document 3*.]

Table 1. Compositions of generated mixtures by stratum in each of three years. Compositions resemble those estimated in the 2008 Ugashik District fishery.

	Percentage			
Region	Stratum 1	Stratum 2	Stratum 3	Stratum 4
North Peninsula	1.1	1.1	1.1	1.1
Ugashik	90.0	86.8	86.8	84.2
Egegik	8.9	12.1	12.1	14.7
Naknek	0.0	0.0	0.0	0.0
Alagnak	0.0	0.0	0.0	0.0
Kvichak	0.0	0.0	0.0	0.0
Nushagak	0.0	0.0	0.0	0.0
Wood	0.0	0.0	0.0	0.0
Igushik	0.0	0.0	0.0	0.0
Togiak	0.0	0.0	0.0	0.0
Other	0.0	0.0	0.0	0.0

Table 2. Simulated harvest (X 10,000) by year and stratum. Harvests were drawn from a normal distribution using the mean and standard deviation observed in the 2008 Ugashik District fishery.

Stratification		Harvest
Year 1	Stratum 1	7.5
	Stratum 2	33.8
	Stratum 3	28.1
	Stratum 4	19.9
	Yearly	89.3
Year 2	Stratum 1	25.9
	Stratum 2	24.6
	Stratum 3	37.4
	Stratum 4	43.5
	Yearly	131.4
Year 3	Stratum 1	14.9
	Stratum 2	39.8
	Stratum 3	43.0
	Stratum 4	16.2
	Yearly	113.9
	Overall	334.6

Table 3. Posterior means and Bayesian confidence intervals (90% CI) for the percentage of North Peninsula fish caught in the simulated harvest of sockeye salmon in the Ugashik District fishery over three years. Three levels of estimates were estimated: 1) individual estimates for each stratum in each year; 2) yearly estimates combining all strata in each year; and 3) overall grand estimate combining all years. As the level of the estimate increases, the confidence intervals get narrower.

Level	Stratification		Mean	90% CI	
				5%	95%
Individual strata					
	Year 1	Stratum 1	0.6	0.0	3.3
		Stratum 2	1.3	0.0	4.9
		Stratum 3	1.6	0.0	5.3
		Stratum 4	1.1	0.0	5.0
	Year 2	Stratum 1	0.4	0.0	2.2
		Stratum 2	2.7	0.0	7.0
		Stratum 3	0.6	0.0	2.5
		Stratum 4	1.2	0.0	4.4
	Year 3	Stratum 1	0.3	0.0	1.9
		Stratum 2	1.3	0.0	5.3
		Stratum 3	0.6	0.0	3.1
		Stratum 4	0.4	0.0	2.0
Yearly					
		Year 1-all strata	1.3	0.0	3.2
		Year 2-all strata	1.2	0.2	2.5
		Year 3-all strata	0.8	0.0	2.4
Across years					
		Over all years	1.1	0.4	2.0

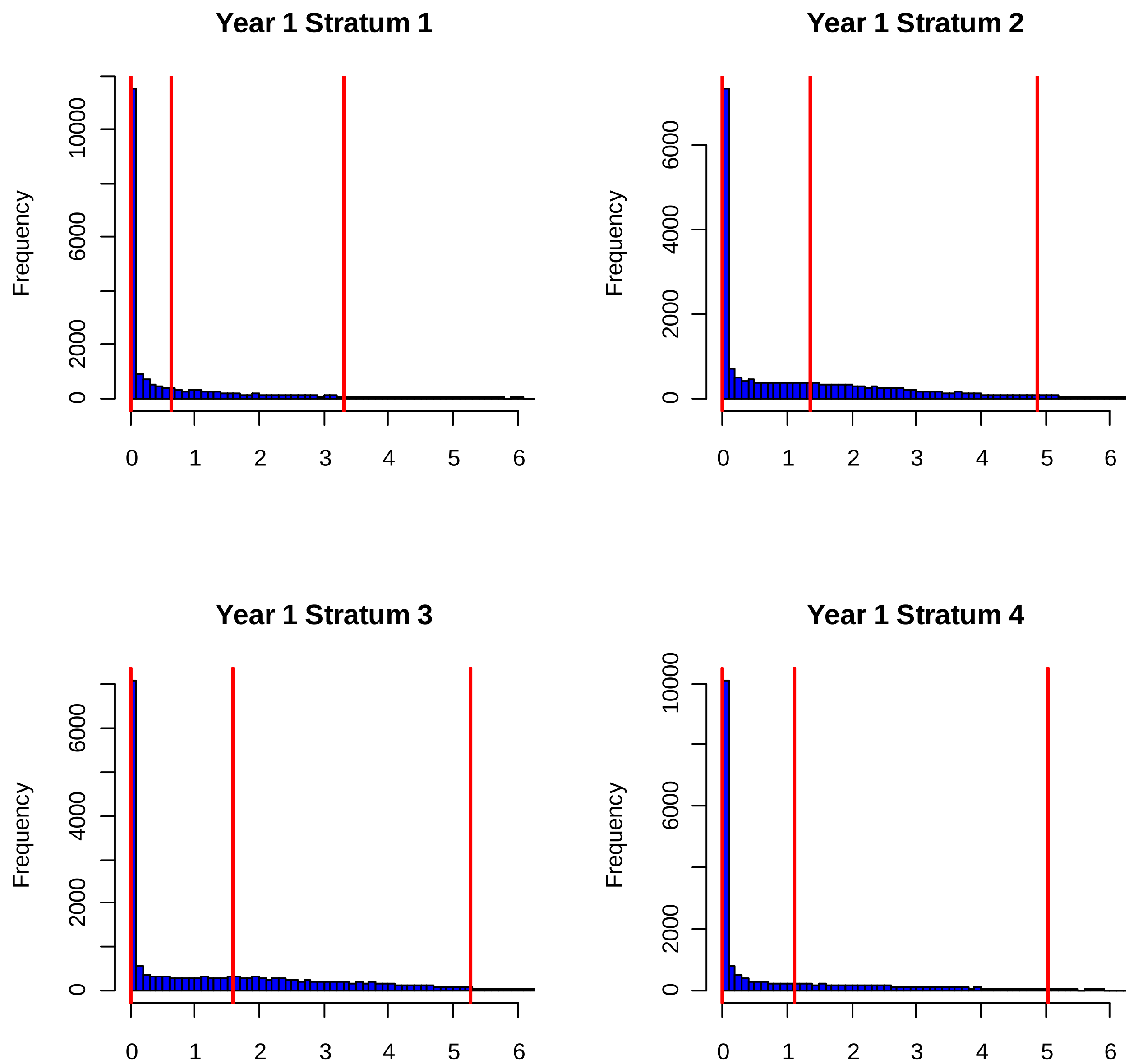

Figure 1. Posterior distributions of North Peninsula's percent contribution to a simulated fishery in the Ugashik District. Plots shown are for the four strata in Year 1 and are typical of those observed in other years. Red vertical lines represent the mean and upper and lower bounds of a 90% confidence interval.

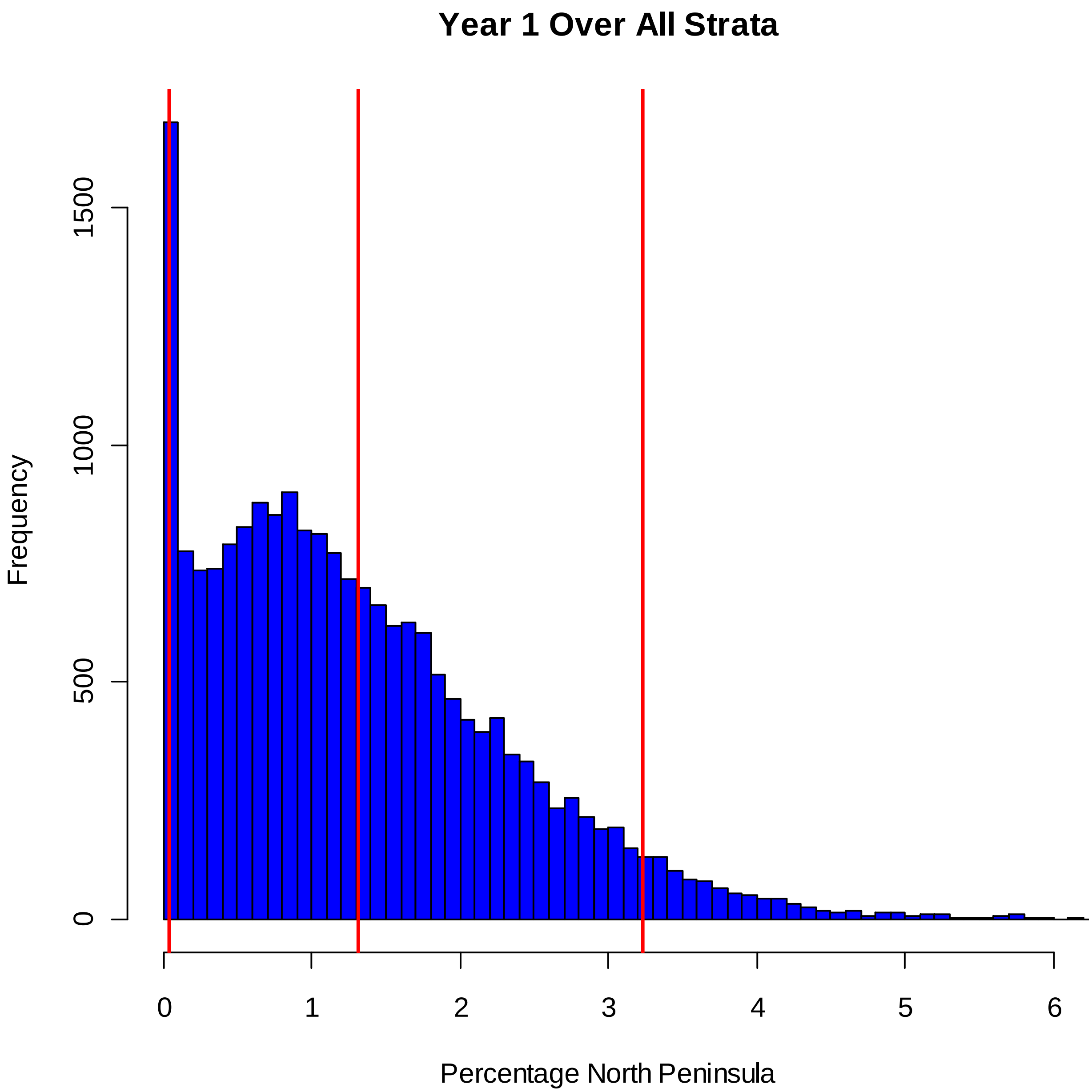

Figure 2. Posterior distribution of North Peninsula’s annual percent contribution to a simulated fishery in the Ugashik District. Plot shown is for Year 1 and is typical of those observed in other years. Red vertical lines represent the mean and upper and lower bounds of a 90% confidence interval.

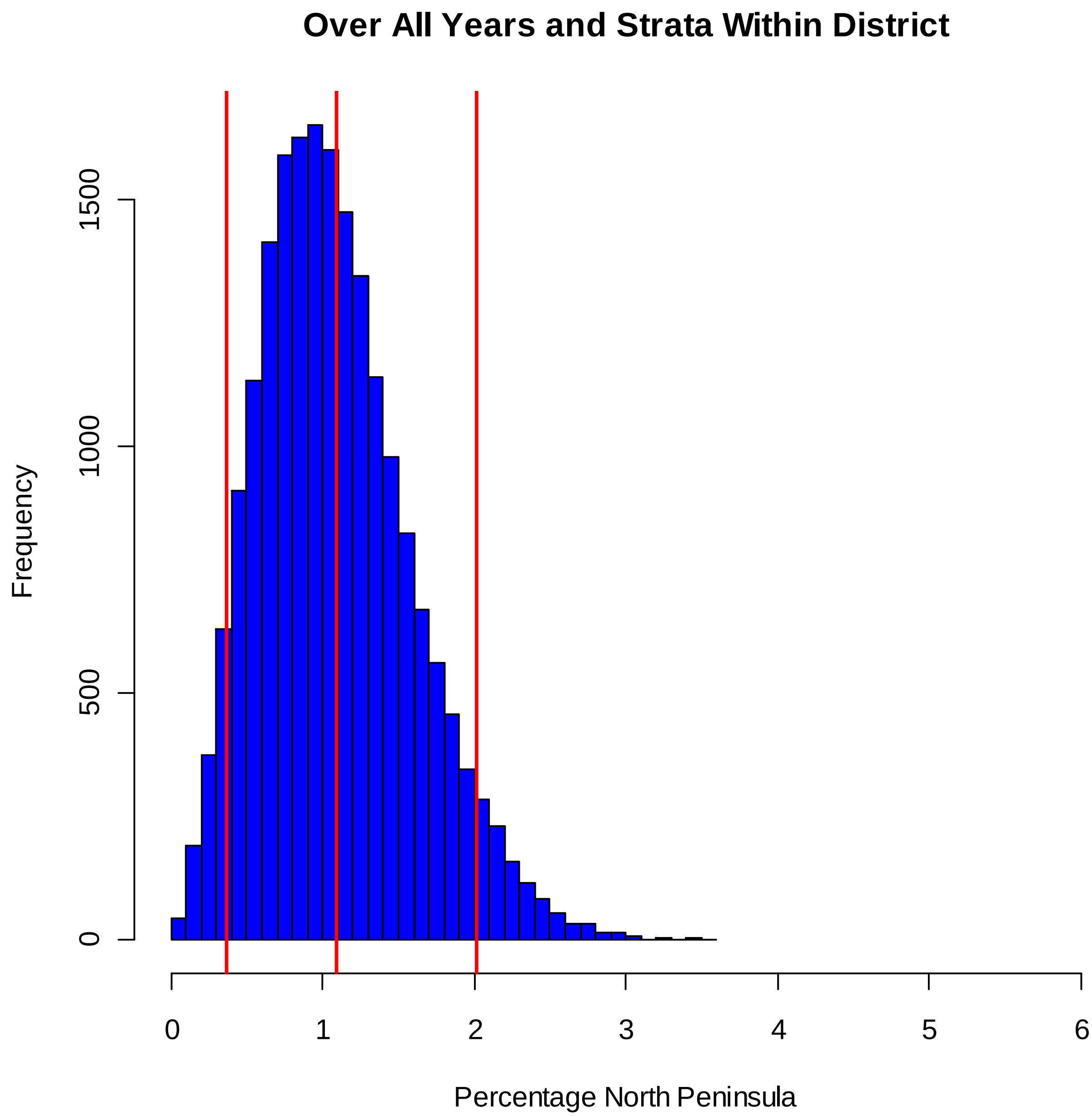

Figure 3. Posterior distribution of North Peninsula's overall percent contribution to a simulated fishery in the Ugashik District. Red vertical lines represent the mean and upper and lower bounds of a 90% confidence interval.