

Confidence Intervals and Hypothesis Tests for CWT
Contribution Estimates Based on the
Binomial-Hypergeometric Model

Richard Comstock
USFWS

October , 1989

Introduction

In this paper I present confidence intervals and tests of hypothesis for coded wire tag (CWT) contribution estimates. I discuss both nonreplicated and replicated experiments.

The results presented in this paper were developed assuming the binomial-hypergeometric compound distribution model (Newman, 1989). This model supposes that the number of tagged fish caught in a fishery follows the binomial distribution and that the number of tagged fish sampled from a fishery follows the hypergeometric distribution.

Notation

- R_i = Number of tagged fish released from group i
 a_{ij} = Number of group i fish caught in fishery j
 p_{ij} = Contribution of group i to fishery j , $p_{ij} = a_{ij}/R_i$
 $q_{ij} = 1 - p_{ij}$
 N_j = Number of fish caught in fishery j
 n_j = Number of fish sampled in fishery j
 s_{ij} = Number of group i fish sampled from fishery j
 r_{ij} = The proportion of fish in fishery j that are from group i ,
 $r_{ij} = a_{ij}/N_j$

p , a , and r are estimated by:

$$\hat{r}_{ij} = s_{ij}/n_j$$

$$\hat{a}_{ij} = \hat{r}_{ij} * N_j$$

$$\hat{p}_{ij} = \hat{a}_{ij}/R_i$$

The variance of the contribution is expressed as:

$$\widehat{\text{var}}(\hat{p}_{ij}) = \frac{\hat{p}_{ij}\hat{q}_{ij}}{R_i} + \frac{N_j(N_j - n_j)}{R_i^2(n_j - 1)} \hat{r}_{ij}(1 - \hat{r}_{ij}) \quad 1.$$

(Newman, 1989)

Covariances

In developing variance estimates for contribution two types of covariance terms must be considered. First, there is a covariance between the estimates of the contribution of one tag group to two fisheries. An estimate of this covariance term is developed as follows:

$$\begin{aligned}\text{cov}(p_{ij}, p_{ik}) &= E(p_{ij} p_{ik}) - p_{ij} p_{ik} \\ &= (1/R_i^2) (E(a_{ij} a_{ik}) - a_{ij} a_{ik}) \\ E(a_{ij} a_{ik}) &= \sum_{a_{ij}=1}^{R_i} \sum_{a_{ik}=1}^{R_i - a_{ij}} a_{ij} a_{ik} M\left(R_i, \frac{a_{ij}}{R_i}, \frac{a_{ik}}{R_i}, \frac{R_i - a_{ij} - a_{ik}}{R_i}\right) \\ &= R_i(R_i - 1) p_{ij} p_{ik} \sum_{a_{ij}=1}^{R_i} \sum_{a_{ik}=1}^{R_i - a_{ij}} M\left(R_i - 2, \frac{a_{ij} - 1}{R_i - 2}, \frac{a_{ik} - 1}{R_i - 2}, \frac{R_i - a_{ij} - a_{ik}}{R_i - 2}\right) \\ &= R_i(R_i - 1) p_{ij} p_{ik}\end{aligned}$$

where, M is the multinomial density function

$$\therefore \text{cov}(\hat{p}_{ij}, \hat{p}_{ik}) = - p_{ij} p_{ik} / R_i$$

substituting \hat{p}_{ij} and \hat{p}_{ik} for p_{ij} and p_{ik}

$$\widehat{\text{cov}}(\hat{p}_{ij}, \hat{p}_{ik}) = - \hat{p}_{ij} \hat{p}_{ik} / R_i \quad 2.$$

The second type of covariance occurs when two different tag codes are recovered in one fishery. An estimate for this covariance is developed as follows:

$$\begin{aligned}
 \text{cov}(\hat{p}_{ij}, \hat{p}_{lj}) &= E(\hat{p}_{ij} \hat{p}_{lj}) - p_{ij} p_{lj} \\
 &= \frac{N_j^2}{n_j^2 R_i R_l} E(s_{ij} s_{lj}) - p_{ij} p_{lj} \\
 E(s_{ij} s_{lj}) &= \sum_{s_{ij}=1}^n \sum_{s_{lj}=1}^{n-s_{ij}} s_{ij} s_{lj} H\left(N_j, \frac{a_{ij}}{N_j}, \frac{a_{lj}}{N_j}, \frac{N-a_{ij}-a_{lj}}{N_j}\right) \\
 &= \frac{n_j(n_j-1)a_{ij}a_{lj}}{N_j(N_j-1)} \sum_{s_{ij}=1}^n \sum_{s_{lj}=1}^{n-s_{ij}} H\left(N_j-2, \frac{a_{ij}-1}{N_j-2}, \frac{a_{lj}-1}{N_j-2}, \frac{N-a_{ij}-a_{lj}}{N_j-2}\right) \\
 &= \frac{n_j(n_j-1) a_{ij} a_{lj}}{N_j(N_j-1)}
 \end{aligned}$$

where, H is the hypergeometric density function

$$\therefore \text{cov}(\hat{p}_{ij}, \hat{p}_{lj}) = p_{ij} p_{lj} \left(\frac{N_j(n_j-1)}{(N_j-1) n_j} - 1 \right)$$

substituting \hat{p}_{ij} and \hat{p}_{lj} for p_{ij} and p_{lj}

$$\widehat{\text{cov}}(\hat{p}_{ij}, \hat{p}_{lj}) = \hat{p}_{ij} \hat{p}_{lj} \left(\frac{N_j(n_j-1)}{(N_j-1) n_j} - 1 \right) \quad 3.$$

Confidence Interval

Figure 1 contains a frequency distribution of simulated contribution estimates. From this, it seems reasonable to assume that contribution is normally distributed. With this assumption we can develop a confidence interval for the average estimated contribution of k tag groups to m fisheries.

$$\hat{p}_{..}/k = \frac{1}{k} \sum_{i=1}^k \sum_{j=1}^m p_{ij}$$

a $(1-\alpha)\%$ C.I. for $\hat{p}_{..}/k$ is, $\hat{p}_{..}/k \pm Z_{\alpha/2} \sqrt{\text{var}(\hat{p}_{..}/k)}$

$$\text{var}(\hat{p}_{..}/k) = \frac{1}{k^2} \left(\sum_{i=1}^k \text{var}(\hat{p}_{i.}) + 2 \sum_{i < j}^k \text{cov}(\hat{p}_{i.}, \hat{p}_{j.}) \right) \quad 4.$$

$$\text{var}(\hat{p}_{i.}) = \left(\sum_{j=1}^m \text{var}(\hat{p}_{ij}) + 2 \sum_{l < j}^m \text{cov}(\hat{p}_{il}, \hat{p}_{lj}) \right)$$

$\text{var}(\hat{p}_{ij})$ is given by equation 1.

$\text{cov}(\hat{p}_{il}, \hat{p}_{lj})$ is given by equation 2.

$$\begin{aligned} \text{cov}(\hat{p}_{i.}, \hat{p}_{j.}) &= E(\hat{p}_{i.} \hat{p}_{j.}) - p_{i.} p_{j.} \\ &= \sum_{l=1}^m \sum_{k=1}^m \left(E(\hat{p}_{il} \hat{p}_{jk}) - p_{il} p_{jk} \right) \end{aligned}$$

for $l < k$ $E(\hat{p}_{il} \hat{p}_{jk}) = p_{il} p_{jk}$ and $E(\hat{p}_{il} \hat{p}_{jk}) - p_{il} p_{jk} = 0$

Therefore,

$$\text{cov}(\hat{p}_{i.}, \hat{p}_{j.}) = \sum_{l=1}^m p_{il} p_{jl} \left(\frac{N_l (n_l - 1)}{(N_l - 1) n_l} - 1 \right)$$

and

$$\widehat{\text{cov}}(\hat{p}_{i.}, \hat{p}_{j.}) = \sum_{l=1}^m \hat{p}_{il} \hat{p}_{jl} \left(\frac{N_l (n_l - 1)}{(N_l - 1) n_l} - 1 \right) \quad 5.$$

A test of hypothesis

With the normality assumption we can also develop a test of hypothesis as follows:

$$H_0: \sum_{i \in \theta_1} A_i \hat{p}_i - \sum_{j \in \theta_2} B_j \hat{p}_j = 0$$

$$H_a: \sum_{i \in \theta_1} A_i \hat{p}_i - \sum_{j \in \theta_2} B_j \hat{p}_j < > 0$$

A_i, B_j are constants, θ_1, θ_2 are sets of tag groups

The test statistic is:

$$Z = \frac{\sum_{i \in \theta_1} A_i \hat{p}_i - \sum_{j \in \theta_2} B_j \hat{p}_j}{\sqrt{\text{var}\left(\sum_{i \in \theta_1} A_i \hat{p}_i - \sum_{j \in \theta_2} B_j \hat{p}_j\right)}}$$

Reject if $Z > N_{\alpha/2}$ or $Z < -N_{\alpha/2}$

$$\begin{aligned} \text{var}\left(\sum_{i \in \theta_1} A_i \hat{p}_i - \sum_{j \in \theta_2} B_j \hat{p}_j\right) &= \text{var}\left(\sum_{i \in \theta_1} A_i \hat{p}_i\right) + \text{var}\left(\sum_{j \in \theta_2} B_j \hat{p}_j\right) \\ &\quad - 2 \text{cov}\left(\sum_{i \in \theta_1} A_i \hat{p}_i, \sum_{j \in \theta_2} B_j \hat{p}_j\right) \end{aligned}$$

$\text{var}\left(\sum_{i \in \theta_1} A_i \hat{p}_i\right)$ and $\text{var}\left(\sum_{j \in \theta_2} B_j \hat{p}_j\right)$ are given by eqn. 4, where $1/k$ is replaced by A_i and B_j respectively.

$$\text{cov}\left(\sum_{i \in \theta_1} A_i \hat{p}_i, \sum_{j \in \theta_2} B_j \hat{p}_j\right) = E\left(\sum_{i \in \theta_1} A_i \hat{p}_i \cdot \sum_{j \in \theta_2} B_j \hat{p}_j\right) - \sum_{i \in \theta_1} A_i \hat{p}_i \cdot \sum_{j \in \theta_2} B_j \hat{p}_j$$

$$\sum_{i \in \theta_1} \sum_{l=1}^M \sum_{j \in \theta_2} \sum_{k=1}^M A_i B_j \left(E(\hat{p}_{il} \hat{p}_{jk}) - p_{il} p_{jk}\right)$$

for $k > j$ $E(\hat{p}_{i1} \hat{p}_{jk}) - p_{i1} p_{jk} = 0$, the remaining terms are:

$$\sum_{i \in \theta_1} \sum_{j \in \theta_2} \sum_{l=1}^m A_i B_j p_{i1} p_{jk} \left(\frac{N_1(n_1-1)}{(N_1-1)n_1} - 1 \right)$$

$$\therefore \widehat{COV} \left(\sum_{i \in \theta_1} A_i \hat{p}_{i.}, \sum_{j \in \theta_2} B_j \hat{p}_{j.} \right) = \sum_{i \in \theta_1} \sum_{j \in \theta_2} \sum_{l=1}^m A_i B_j \hat{p}_{i1} \hat{p}_{jk} \left(\frac{N_1(n_1-1)}{(N_1-1)n_1} - 1 \right)$$

As a check on the performance of this test, I performed a simulation. The results are as follows:

Number of iterations = 20000
 Number of tag groups = 4
 Release size (each tag group) = 4000
 Number of other fish released = 16000
 Contribution rate to fishery 1 (all groups) = .02
 Contribution rate to fishery 2 (all groups) = .03
 Fishery sampling rate = .2

$$H_0: 1/2 (p_{1.} + p_{2.}) - 1/2 (p_{3.} + p_{4.}) = 0$$

alpha level of test = .05
 percent times significant = .04945

The percent of times the null hypothesis was rejected agrees well with the alpha value. It appears that the test performs well under the null hypothesis.

Replicates

Replicates are typically associated with experimental units (Hurlbert, 1984). They are important for generalizing the results of an experiment to include a greater range of experimental units. In CWT experiments the concept of replicates is somewhat vague because the choice of experimental units is not obvious. In agricultural experiments it seems natural to identify individual plots of land as experimental units. In the case of fish hatcheries possible choices are ponds, years, kin groups, etc., but none of these seems obvious. If study objectives are considered, however, reasonable choices for experimental units, and replicates, can be determined.

In choosing experimental units and replicates we must first consider the basic building block of CWT studies, the tag group. Each fish in the tag group has the same mark. A single contribution estimate will be computed for this group. Events concerning each fish that are of interest to the CWT investigator (i.e. was the fish caught, if so, was it sampled) are Bernoulli trials. Therefore, theoretical equations can be developed for variance estimation using equation 1. If a variance estimate is computed for a single tag group the implicit model is:

$$p = \mu + \varepsilon$$

Where μ is the implicit contribution rate for the tag group. ε is random experimental error that results from the possible outcomes of the Bernoulli trials that comprise the CWT experiment.

Of course the implicit contribution is likely dependent upon the year of release, specific pond conditions, broodstock selection techniques, etc. The error term does not include variation from these sources. Therefore, while the theoretical variance of the tag group could be used to construct a confidence interval for that tag group, it should not be used for a confidence interval for other groups from other ponds or from other years, etc.

If the investigator wanted to generalize the results of the tag group to fish from other ponds or compare the contribution of two tag groups reared in separate ponds, an appropriate model is:

$$P_i = \mu + C_i + \varepsilon$$

Here a pond effect (C) is added to the model. This model assumes that specific pond conditions effect survival. These can be either fixed or random effects. An example of a fixed effect would be if a certain pond has consistently better or worse flows than another. Random effects could result from small variations from planned feeding rates, loading densities, etc. μ is the implicit contribution rate of the hatchery stock during a specific year, before pond conditions are considered. ϵ is defined as above.

If fish from several ponds were marked, each with a separate tag code, the resulting tag groups could be considered as replicates. Contribution could be computed for each replicate (p_i) and the average (\bar{p}) could be used as an estimate of the contribution of the entire group. The variance of \bar{p} could be estimated by computing a sample variance from the set of replicate contribution estimates and applying a covariance adjustment. The correct formula is developed as follows:

k = Number of tag codes

p_i = contribution of group i

$$\bar{p} = \sum_{i=1}^k p_i / k$$

μ = actual contribution rate $p_i \sim N(\mu, \sigma^2)$

$$E\left(\sum_{i=1}^k (p_i - \bar{p})^2 / k(k-1)\right) = \frac{1}{k-1} E((p_1 - \bar{p})^2)$$

Change of variables to make computation less messy

$$X_i = p_i - \mu \quad \text{so} \quad E(X_i) = 0$$

$$p_1 - \bar{p} = (p_1 - \mu) - (\bar{p} - \mu) = X_1 - \bar{X}$$

$$\therefore E((X_1 - \bar{X})^2) = E((p_1 - \bar{p})^2)$$

$$E((X_1 - \bar{X})^2) / k-1 = \frac{1}{k-1} (EX_1^2 + E\bar{X}^2 - 2E(X_1, \bar{X}))$$

$$EX_1^2 = \text{var}(X_1)$$

$$E\bar{X}^2 = \text{var}(\bar{X})$$

$$-2E(X_1, \bar{X}) = -2 E\left(X_1 \sum_{i=1}^k X_i / k\right)$$

$$= -\frac{2}{k} (EX_1^2 + \sum_{i=2}^k EX_1 X_i)$$

$$= -\frac{2}{k} \text{var}(X_1) - \frac{2(k-1)}{k} \text{cov}(X_1, X_2)$$

therefore,

$$E((X_1 - \bar{X})^2) / k - 1 = \frac{1}{k-1} (\text{var}(X_1) + \text{var}(\bar{X}) - \frac{2}{k} \text{var}(X_1) - \frac{2(k-1)}{k} \text{cov}(X_1, X_2))$$

$$\text{since } \text{var}(X_1) = k \text{ var}(\bar{X}) - (k-1) \text{cov}(X_1, X_2)$$

$$E((X_1 - \bar{X})^2) / k - 1 = \text{var}(\bar{X}) - \text{cov}(X_1, X_2)$$

$$\therefore \hat{\text{var}}(\bar{X}) = E\left(\sum_{i=1}^k (X_i - \bar{X})^2 / k(k-1)\right) + \hat{\text{cov}}(X_1, X_2)$$

$$\text{and } \hat{\text{var}}(\bar{p}) = E\left(\sum_{i=1}^k (P_i - \bar{p})^2 / k(k-1)\right) + \hat{\text{cov}}(P_1, P_2)$$

$\hat{\text{cov}}(P_1, P_2)$ is given by equation 5. P_1 and P_2 should be replaced by \bar{p} for computing this value.

See the Appendix for simulation tests of this correction.

This variance estimate includes the experimental error described above, plus between pond variation. This would be the appropriate variance to use in establishing confidence limits for the contribution estimate of marked and associated unmarked ponds. Of course, more complicated models may be appropriate. Stock and year effects could easily be added.

If replicates are included in an experimental design, tests of hypothesis can be performed using analysis of variance methods (Snedecor and Cochran, 1967). For example, suppose that an investigator wants to test the hypothesis that two different stocks have the same contribution rate. Suppose that each stock is reared in several ponds. The investigator might choose a design as follows:

	Stock 1	Stock 2
Pond 1	P_{11}	P_{21}
Pond 2	P_{12}	P_{22}

Here two ponds are selected for marking from each of the stocks, for a total of 4 tag groups. As shown above the design does not allow enough degrees of freedom for the residual mean square. However, since the s_{ij} represent the summation of Bernoulli trails we can use other methods to compute residual mean square.

If a tag group were randomly divided into equally sized subgroups then each subgroup would have the same expected contribution. That is:

$$P(\text{fish is caught} \mid \text{fish is from subgroup } i) = P(\text{fish is caught})$$

Therefore, any fish from the group that was caught in a fishery would have equal probability of being from any of the subgroups. That is:

$$P(\text{caught fish is from subgroup } i) = P(\text{caught fish is from subgroup } j)$$

Given this we can obtain a residual mean square by randomly dividing the recoveries from each tag group into subgroups and treating each subgroup as a replicate. The design then becomes:

	Stock 1	Stock 2
Pond 1	P_{111}	P_{211}
	P_{112}	P_{212}
Pond 2	P_{121}	P_{221}
	P_{122}	P_{222}

Because each group is reared in a separate pond the design is nested.

Since the p_{ijk} are not independent the assumptions required for analysis of variance will not be fully met. However, the simulations below show that the test does perform well. Given the covariance terms described above it should be possible to adjust the mean square computation to eliminate the bias. I have not, as yet, developed this adjustment.

Test 1

Iterations = 20000
 Number of fisheries = 2
 release (each mark group) = 4000
 number unmarked released = 16000
 fishery sampling rate = .2
 alpha value = .05
 contribution rates (all groups):
 fishery 1 = .03
 fishery 2 = .02

Percent of trials where stock effect (ms stock/ ms pond) was significant = 4.83%

Percent of trails where pond effect (ms pond/ ms residual) was significant = 4.95%

Test 2

Iterations = 5000
 Number of fisheries = 2
 release (each mark group) = 4000
 number unmarked released = 16000
 fishery sampling rate = .2
 alpha value = .05
 contribution rates:

stock	pond	fishery	contribution
1	1	1	.02
1	1	2	.01
1	2	1	.04
1	2	2	.03
2	3	1	.02
2	3	2	.01
2	4	1	.04
2	4	2	.03

Percent of trials where pond effect (ms pond/ ms residual) was significant = 84.78

Test 3

Iterations = 5000

Number of fisheries = 2

release (each mark group) = 4000

number unmarked released = 16000

fishery sampling rate = .2

alpha value = .05

contribution rates:

stock	pond	fishery	contribution
1	1	1	.02
1	1	2	.01
1	2	1	.02
1	2	2	.01
2	3	1	.04
2	3	2	.02
2	4	1	.04
2	4	2	.02

Percent of trials where stock effect (ms stock/ ms pond) was significant = 57.8%

Percent of trials where pond effect (ms pond/ ms residual) was significant = 5%

References

- Cochran, W.G. 1977. Sampling Techniques. John Wiley and Sons, New York.
- Snedecor, G.W. and Cochran, W.G. 1967. Statistical Methods The Iowa State University Press, Iowa.
- Newman, K. 1989. Variance estimation for contribution estimates based on sample recoveries of coded wire tagged fish. Proceedings of the Symposium on Marking and Tagging, American Fisheries Society, Bethesda, Maryland (In press).
- Hurlburt, S.H. 1984. Pseudoreplication and the Design of Ecological Field Experiments. Ecological Monographs 54(2), 1984, p.p. 187-211

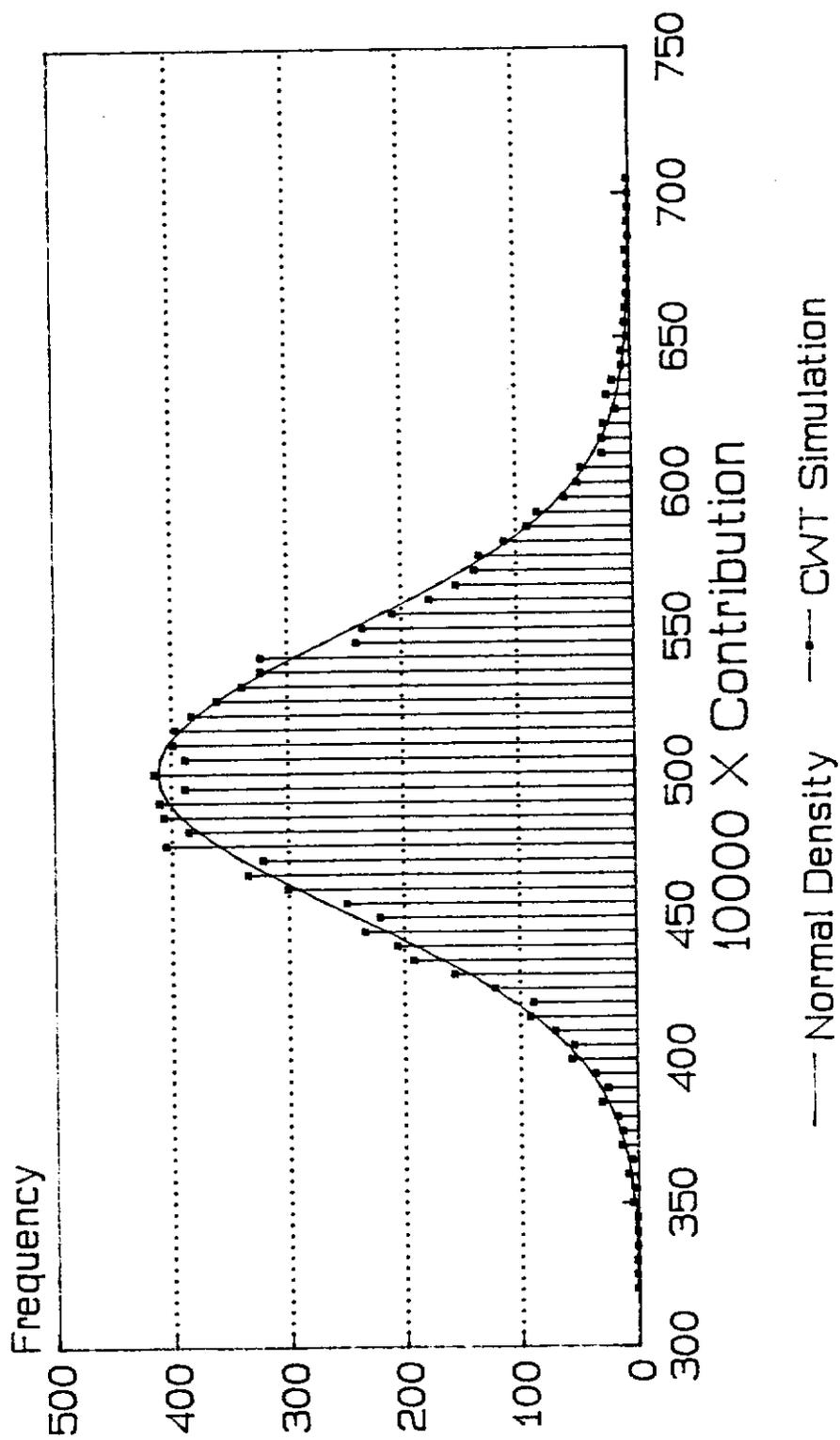


Figure 1, Frequency Distribution of Simulated Contribution. Contribution = 0.05, Iterations = 10000

Appendix

The following tables are modified versions of those presented at the July 12 meeting. They now contain a column for the adjusted replicate standard deviation estimate.

Table 3. Result of CWT simulation, Series 1, where Number of "other" fish varies.

<hr/>						
		Number released tag group 1 (R1)	=	5000		
		Number released tag group 2 (R2)	=	5000		
		Contribution, all groups	=	.05		
		Fishery sampling rate	=	.2		
		Number of iterations	=	1000		
	<hr/>					
R3	R3 / (R1+R2+R3)	Emp. p_4	Emp. sd(p_4)	Replicate sd(p_4)	Adjusted Replicate	Newman's Long sd(p_4)
0	0%	.04994	.002117	.004974	.00218	.002178
5000	33%	.05005	.003419	.005162	.00365	.003375
10000	50%	.05008	.003992	.005095	.00399	.003835
20000	67%	.04996	.004184	.004927	.00420	.004246
30000	75%	.05005	.004446	.004993	.00446	.004441
40000	80%	.05021	.004403	.005090	.00468	.004560
80000	89%	.04997	.004711	.004992	.00476	.004743
120000	92%	.05002	.004851	.004934	.00478	.004817
160000	94%	.05013	.004874	.004870	.00475	.004860
<hr/>						

Table 4. Result of CWT simulation, Series 2, where fishery sampling rate varies.

<hr/>					
Number released tag group 1 (R1) = 5000 Number released tag group 2 (R2) = 5000 Number released "other" fish (R3) = 10000 Contribution, all groups = .05 Number of iterations = 1000					
<hr/>					
Fishery Sampling Rate	Emp. p_4	Emp. $sd(p_4)$	Replicate $sd(p_4)$	Adjusted Replicate	Newman's Long $sd(p_4)$
.2	.05004	.003828	.004901	.00374	.003834
.4	.05012	.002846	.003463	.00287	.002915
.6	.04980	.002598	.002861	.00255	.002529
.8	.05000	.002503	.002472	.00234	.002398
1	.04991	.002127	.002128	.00212	.002179
<hr/>					