Alaska
Fisheries Science
Center

National Marine
Fisheries Service

U.S. DEPARTMENT OF COMMERCE

AFSC PROCESSED REPORT 92-01

Estimation of the Fishing Power Correction Factor

February 1992

ERRATA NOTICE

ESTIMATION OF THE FISHING POWER CORRECTION FACTOR

Russell F. Kappenman
Alaska Fisheries Science Center
Seattle, Washington

1. Introduction

One of the missions of the Alaska Fisheries Science Center is to monitor changes in abundance of many different species of fish in several different areas of the Pacific Ocean. One way of accomplishing this mission is to conduct trawl surveys, where fishing boats trawl for a specified period of time at each of a number of stations in the region of interest.

The fish caught by each trawl are separated according to species, and the weight of each of the species caught is determined. Also recorded is the area swept by each trawl. For any species of interest, the ratio of the weight caught by a trawl to the area swept by the trawl is referred to as the catch per unit effort (cpue) random variable.

Typically, more than one fishing vessel is used in a trawl survey, and for a given species of fish, it may well be that one of the boats is more efficient at catching the species. Consider the hypothetical situation in which two boats could trawl the same area at the same time. Let X and Y represent the cpue's for a given species for the most efficient boat and another, respectively. Then X/Y would be some unknown parameter which is at least one. This unknown parameter is commonly referred to as the fishing power correction (fpc) factor, and it must be estimated by use of cpue observations for the two vessels.

The problem of estimating the fpc may be stated as the following statistics problem. Let X and Y represent two positive random variables whose distributions are unknown but identical, except possibly for the values of the scale parameters of the distributions. That is, the p.d.f.'s of X and Y are of the forms $\frac{1}{b_x}f(\frac{x}{b_x})$ and $\frac{1}{b_y}f(\frac{y}{b_y})$, respectively, where f is some unknown p.d.f. Find an estimate of $r=b_x/b_y$, given random samples of observations of X and of Y.

A good estimate of the fpc is critical, because for estimation of abundance of a species, the cpue's of a vessel are adjusted by multiplying them by an estimate of a fpc. The estimate of abundance is quite sensitive to the estimate of the fpc used.

Procedures for estimating the ratio of scale parameters have been proposed in the statistical literature. Undoubtedly, the most popular one is a jackknife estimator suggested by Miller (1968), who also discussed the others and their drawbacks. Miller's jackknife estimator is discussed in detail in Hollander and Wolfe (1973). At the beginning of the research which led to this paper, the jackknife estimator was applied to many sets of cpue data. Several times it was found that the jackknife estimate was pretty clearly unsatisfactory. At the time, the belief was that poor estimates of the fpc could be produced by the jackknife estimator because of the nature of the cpue distributions. These distributions are routinely heavily skewed to the right.

Thus a search was conducted for a better way of estimating the ratio of scale parameters of distributions which are identical except for the values of their scale parameters. The estimator discussed in Section 2 is the one, of many studied, which performs the best. A bonus of the research is that this estimator appears to perform remarkably better than the jackknife estimator even in situations where the latter estimator should, by its nature, do exceptionally well. These situations will be discussed in Section 3, where some simulation studies, conducted to assess the performances of the estimators, are described.

2. Development of the Ratio Estimator

Suppose that $x_1,...,x_n$ and $y_1,...,y_m$ represent, respectively, random samples of observations of two positive random variables X and Y. Let b_x and b_y denote, respectively, the unknown scale parameters of the distributions of X and Y. The goal is to use the samples of observations of X and Y to estimate $r=b_x/b_y$. We shall assume that the p.d.f.'s of X and Y are identical, except possibly for the values of their scale parameters.

An important implication of the latter assumption is that X and rY have identical distributions. This implication plays a prominent role in the development of an estimator for r. This estimator will now be developed step by step.

The first step is to find a number d such that X^d and r^dY^d have, approximately at least, the same normal distribution. The number r is treated as an unknown constant. A value for d may be found by assuming, for some constant r, that $x_1^d,...,x_n^d,r^dy_1^d,...,r^dy_m^d$ are independently and identically normally distributed with an unknown mean and variance. A maximum likelihood estimation argument suggests, as a value for d, the solution, for d, to the equation

$$\frac{n+m}{d}+u-\frac{1}{w}[\Sigma x_i^{2d}\ln x_i-v\Sigma x_i^d\ln x_i+q^2\Sigma y_j^{2d}\ln y_j-vq\Sigma y_j^d\ln y_j]=0, \quad (1)$$

where

$$u=\Sigma\ln x_i+\Sigma\ln y_j,\quad v=\frac{\Sigma x_i^d+q\Sigma y_j^d}{n+m},\quad w=\frac{1}{n+m}[\Sigma(x_i^d-v)^2+\Sigma(qy_j^d-v)^2],$$

$$q=\frac{-s+\sqrt{s^2-4pt}}{2p},\quad p=\frac{n}{n+m}\Sigma y_j^{2d}-\frac{n}{(n+m)^2}(\Sigma y_j^d)^2,\quad s=\frac{m-n}{(n+m)^2}\Sigma x_i^d\,\Sigma y_j^d,$$

and

$$t=\frac{m}{(n+m)^2}(\Sigma x_i^d)^2-\frac{m}{n+m}\Sigma x_i^{2d}.$$

Next, let r_0 represent some arbitrary value for r, and suppose one is forced to decide between two conjectures:

(a) $x_1^d,...,x_n^d,r_0^dy_1^d,...,r_0^dy_m^d$ all have the same distribution and

(b) the distribution of $x_1^d, \ldots, x_n^d$ is different from the distribution of $r_0^d y_1^d, \ldots, r_0^d y_m^d$.

To choose between these two conjectures, one might use the following procedure:

1. Assume, for the moment, that (a) is the correct choice. Calculate a predicted value for x_j^d, $j=1,\ldots,n$, by making use of all observations except for x_j itself, and calculate a predicted value for $r_0^d y_h^d$, for $h=1,\ldots,m$, by making use of all observations except for y_h itself. Because of the way d is chosen, the most reasonable such predicted values for x_j^d and $r_0^d y_h^d$ are, respectively,

$$x_j^{d\,(a)} = \frac{1}{n+m-1}\left[\sum_{i\neq j} x_i^d + \sum_{h=1}^{m} r_0^d y_h^d\right]$$

and

$$r_0^d y_h^{d\,(a)} = \frac{1}{n+m-1}\left[\sum_{j=1}^{n} x_j^d + \sum_{i\neq h} r_0^d y_i^d\right] .$$

2. Calculate the sum of squares of the differences between the x_j^d's and the $r_0^d y_h^d$'s and their predicted values, under the assumption that (a) is the best choice. This sum of squares is

$$S_a = \sum_{j=1}^{n} [x_j^d - x_j^{d\,(a)}]^2 + \sum_{h=1}^{m} [r_0^d y_h^d - r_0^d y_h^{d\,(a)}]^2 .$$

3. Assume that (b) is the correct choice. Calculate a predicted value for x_j^d, $j=1,\ldots,n$, using all observations except for x_j itself, and calculate a predicted value for $r_0^d y_h^d$, $h=1,\ldots,m$, using all observations except for y_h

itself. In this case, the most reasonable such predicted values for x_j^d and $r_0^d y_h^d$ are, respectively,

$$x_j^{d\,(b)} = \frac{1}{n-1} \sum_{i \neq j} x_i^d$$

and

$$r_0^d y_h^{d\,(b)} = \frac{1}{m-1} \sum_{i \neq h} r_0^d y_i^d \; .$$

4. Calculate the sum of squares of the differences between the x_j^d's and the $r_0^d y_h^d$'s and their predicted values, under the assumption that (b) is the best choice. This sum of squares is

$$S_b = \sum_{j=1}^{n} [x_j^d - x_j^{d\,(b)}]^2 + \sum_{h=1}^{m} [r_0^d y_h^d - r_0^d y_h^{d\,(b)}]^2 \; .$$

5. The choice between (a) and (b) may be made by comparing the values of S_a and S_b. Conjecture (a) would be chosen if $S_a < S_b$; otherwise, conjecture (b) would be chosen.

In general, there are two numbers, r_1 and r_2 say, such that conjecture (a) will be chosen by this cross-validation procedure if and only if $r_1 < r_0 < r_2$. The actual values of r_1 and r_2 depend upon the observations $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$. Each is a value of r_0 such that $S_a = S_b$.

Thus any number between r_1 and r_2 might be regarded as a reasonable estimate of r, if one uses the procedure described above to decide whether or not an estimate is reasonable. When one is forced to select just one estimate of r, the intuitively best choice would be the value of r_0 which makes S_a as small as it can possibly be relative to S_b. This is the value of r_0 which minimizes the difference $S_a - S_b$, and it is the estimate proposed here as an alternative to the jackknife estimate of the ratio of scale parameters.

To summarize, the proposed estimate of the ratio of scale parameters r is the value of r which minimizes the function g(r), where g(r) is given by

$$g(r)=\left(\frac{n+m}{n+m-1}\right)^2\left[\Sigma\left(x_i^d-\frac{\Sigma x_i^d+r^d\Sigma y_j^d}{n+m}\right)^2+r^{2d}\Sigma\left(y_j^d-\frac{\Sigma y_j^d+\{\Sigma x_i^d\}/r^d}{n+m}\right)^2\right]$$

$$-\left(\frac{n}{n-1}\right)^2\left[\Sigma\left(x_i^d-\frac{\Sigma x_i^d}{n}\right)^2-\left(\frac{m}{m-1}\right)^2\left[\Sigma\left(y_j^d-\frac{\Sigma y_j^d}{m}\right)^2\right]r^{2d}\right. \qquad (2)$$

and d satisfies (1).

3. Simulation Studies

To assess the performance of the ratio of scale parameters estimator proposed in Section 2, some simulation studies were conducted.

A number of well-known distributions were used as models for the distributions of X and Y. These included members of the lognormal, gamma, and Weibull families of distributions. The members of the lognormal family used were the ones corresponding to the shape parameter (σ) values 2.2, 1.3, 0.8, and 0.5. The members of the gamma family used were the ones corresponding to the shape parameter values 0.5, 1, 2, and 4, and the members of the Weibull family used were the ones corresponding to the shape parameter values 0.65, 1, 1.5, and 2.2. The lognormal, gamma, and Weibull distributions, whose shape parameter is c, are denoted here by LN(c), G(c), and W(c), respectively.

These distributions, for the shape parameter values we are considering here, are all fairly highly skewed to the right. The skewness increases as the shape parameter value increases in the lognormal distribution case, but it increases as the shape parameter value decreases in the gamma distribution and Weibull distribution cases.

Also used as models for the distributions of X and Y were some distributions derived from members of the Tukey lambda family. If

$$w=\frac{u^{\lambda}-(1-u)^{\lambda}}{\lambda},$$

where u is a uniform (0,1) random variable, then w has a lambda distribution. This distribution, denoted here by L(λ), is a symmetric distribution, and its range is $(-\frac{1}{\lambda},\frac{1}{\lambda})$, if $\lambda>0$. If $\lambda\le 0$, the range is $(-\infty,\infty)$. The tail weights of the lambda distribution decrease as λ increases. L(-1) is much like a Cauchy distribution, and L(0.1349) is very similar to the normal distribution. For $\lambda>0.1349$, the tails of the lambda distribution are lighter than normal distribution tails, and L(1) is, in fact, a uniform distribution.

Two of the distributional models used for X and Y were the distributions of $w+\frac{1}{\lambda}$, for $\lambda=0.1349$ and $\lambda=0.5$. These are distributions of positive random variables with symmetric distributions. One of these has normal-like tail weights, and the other has lighter than normal tails.

Another distributional model used for X and Y was a L(-1) distribution truncated at its 0.01 and 0.99 percentiles and shifted so that the left endpoint of the range was zero. Thus this model is a symmetric, heavy tailed distribution of a positive random variable.

For the simulation studies, two sample sizes were used, a moderate one n=m=30 and a large one n=m=100. For each of these sample sizes and each of the distributions described above, 500 times random samples of size n and m were generated. Each time a pair of random samples was generated, the samples were used to obtain the Miller (1968) jackknife estimate of the ratio of scale parameters, denoted here by f_1, and the ratio of scale parameters estimate developed in Section 2 and denoted here by $\hat{r}_2$.

Of interest is the performance of $\hat{r}_2$ relative to that of $\hat{r}_1$. One measure of this relative performance is the ratio of the root mean squared error (RMSE) of $\hat{r}_2$ to the RMSE of $\hat{r}_1$. It is easily shown that this ratio of RMSE's does not depend upon the actual value of r. Thus r was taken to be one.

The goal of the simulation studies was to get, for a rather broad range of sampling distributions and sample sizes, estimates of the ratio of the RMSE's for the two estimators of the ratio of scale parameters. Table 1 gives the estimates of RMSE($\hat{r}_2$)/RMSE($\hat{r}_1$) obtained by the studies.

TABLE 1

Ratios of the root mean squared error of $\hat{r}_2$ to the root mean squared error of $\hat{r}_1$.

Sampling Distribution	Sample Size n=m=30	n=m=100
LN(2.2)	0.13	0.06
LN(1.3)	0.30	0.21
LN(0.8)	0.39	0.32
LN(0.5)	0.39	0.35
W(0.65)	0.69	0.58
W(1)=G(1)	0.78	0.76
W(1.5)	0.78	0.81
W(2.2)	0.68	0.71
G(0.5)	1.04	1.00
G(2)	0.65	0.63
G(4)	0.51	0.53
L(-1)	0.04	0.05
L(0.1349)	0.27	0.27
L(0.5)	0.90	0.91

Note that almost all of the ratios in Table 1 are less than one, and most are really quite small. The implications of this are that $\hat{r}_2$ is generally more efficient than $\hat{r}_1$ over a broad range of types of sampling distributions, and, most of the time, is substantially more efficient. Further, a comparison of the two columns in Table 1 indicates that the efficiency of $\hat{f}_2$ over $\hat{f}_1$ generally increases with sample size.

One other aspect of the results of the simulation studies is particularly noteworthy and somewhat surprising. As was pointed out above, the L(0.1349) distribution is very similar to a normal distribution. Since $\hat{f}_1$ is essentially the ratio of one sample standard deviation to another sample standard deviation, one would expect that $\hat{f}_1$ will perform extremely well when the samples are drawn from L(0.1349) distributions. The simulation studies estimates of the ratios of RMSE($\hat{f}_2$)/RMSE($\hat{f}_1$), for n=m=30 and n=m=100, are both 0.27, indicating that $\hat{f}_2$ performs much better.

4. Estimation of the Fishing Power Correction Factor

The results of the simulation studies, discussed in Section 3, demonstrate the potential of the ratio of scale parameters estimation procedure, developed in Section 2, for providing good, robust estimates. This estimator of the ratio of scale parameters is the estimator which would be recommended for use in estimating the fpc, for the case where a specified vessel's cpue's are to be adjusted by multiplying them by an estimate of the fpc. The specified vessel need not be the least efficient one.

Often though, the cpue's for the vessels are used to select a vessel as being the more efficient of two vessels for a given species. The cpue's of the vessel which appears to be the least efficient are then multiplied by an estimate of the fpc which is also based on the vessels' cpue's. This practice is normally followed when the index used to monitor change in abundance, of a given species, is an estimate of biomass. For this case, the fpc must be at least one. This procedure, in effect, defines an fpc estimator which is slightly different from the ratio of scale parameters estimator developed in Section 2. The latter estimator is essentially based on the assumption that the vessel, whose cpue's are to be adjusted, is known before the cpue's are observed.

The practice of using cpue's to select a most efficient vessel and to estimate an fpc causes the fpc estimator to be positively biased, if, in fact, the fpc is one.

The method used, in Section 2, to develop an estimator of the ratio of scale parameters has a distinct advantage which may be used to produce an fpc estimator with reduced bias. One may make use of r_1, defined in Section 2 to be the smallest solution of $g(r)=0$, where $g(r)$ is given by (2). Any estimator of the form

$$ar_1+b\hat{r}_2\ , \qquad (3)$$

where $a>0$, $b>0$, and $a+b=1$, will, by the argument of Section 2, be a reasonable estimator of r. Further, it will have less positive bias than $\hat{r}_2$ has, when $r=1$. At the Alaska Fisheries Science Center, (3), with $a=b=\frac{1}{2}$, is now being used to estimate the fpc, because it appears as though the amount of positive bias in the resulting estimator, when the fpc is equal to one, is about the same as the amount of negative bias, when the fpc is 1/0.75.

REFERENCES

Hollander, M. and Wolfe, D.A.(1973). Nonparametric Statistical Methods. John Wiley and Sons.

Miller, R.G.(1968). Jackknifing variances. Annals of Mathematical Statistics, 39, 567-582.