Journal of Statistical Planning and Inference 19 (1988) 253-259
North-Holland

A NOTE ON COMPARISON OF CONDITIONAL MEANS

Gutti Jogesh BABU

The Pennsylvania State University, University Park, PA 16802, U.S.A.

Received 24 September 1986; revised manuscript received 22 June 1987
Recommended by R.N. Bhattacharya

Abstract: Let $(X_1, Y_1), \ldots, (X_p, Y_p)$ be i.i.d. random vectors. This paper is concerned with the estimation of $E(X_1 \mid X_1 > 0)/E(Y_1 \mid Y_1 > 0)$. This problem arises naturally in comparing the fishing ability of different vessels. In this case (X_i, Y_i) denote the catch per tow of the two vessels at the site i or time i. Due to the highly skewed nature of the distributions of the catches the difference of the average means would not give a realistic idea of the relative fishing power of the vessels.

AMS Subject Classification: 62G05.

Key words and phrases: Mean square error; estimation of relative fishing power; bootstrap method.

1. Introduction

The problem considered in this paper arose during the discussions with the scientists at the Northeast Fisheries Center, Woods Hole Laboratory. They periodically conduct surveys to monitor fluctuations in structure and size of fish populations in a way independent of commercial fishery statistics and to obtain data to assess production potential of various species. An important aspect of any long term survey program is standardization of the survey unit. The factors that determine effective unit size are ships, nets, doors, etc. To study the long term effect of pollutants on the stock of the fish, one requires data on the stock for several decades. But, the life of any survey vessel is about 20 years. Further, nets and doors are changed often. Inherent differences in the vessels, nets and doors would introduce potential for bias due to possible differences in the fishing power. This may be the result of differences in size, displacement, horsepower or other factors. So there is a need for conversion factor to standardize the time series data of the stock obtained using different vessels, nets, doors, etc.

To arrive at such estimates experiments were conducted using different ships in essentially the same set of conditions. Paired tows were made using the survey vessels Albatross IV and Delaware II to estimate the possible differences in fishing

power, which may be the result of the differences in size, horsepower or other factors. Byrne and Fogarty (1985) have done some analysis using non-parametric methods by rank transforming the observations.

Due to the highly skewed nature of the distribution of the catches, the difference in the mean catches would not give an efficient estimate of the relative fishing power. In multispecies fish surveys, when large areas are sampled, any particular species usually occupies only a part of the total survey area. In these circumstances, the zero values can be taken to represent areas of unsuitable or unoccupied habitat. The proportion of non-zero values in the sample estimates the proportion of the total survey area that is occupied by the species.

The interpretation of the proportions of non-zeros in a sample as an estimate of habitat area may be vague in some situations, especially for mobile populations. A suitable habitat may change from time to time due to many factors including the timing of the survey, or an area is unoccupied simply because of a low population level. However, keeping the zeroes separate often enables one to fit a relatively simple distribution like the lognormal to the non-zero values. But we just cannot ignore the data pairs with at most one zero value.

In the paper we suggest a nonparametric method to estimate the relative fishing power and develop the asymptotic properties. We also obtain an estimate of the mean square error of the estimate. Using these results the fishing power is estimated in the last section, for 20 species based on the catches by Albatross IV and Delaware II. It is reasonable to assume that no catch by both the vessels at a point do not effect the relative 'fishing power' of the ships. It may, simply, be due to the lack of fish in that area. As a result, it is enough to consider those pairs of the data where at least one component is non-zero. This leads to the consideration of independent vectors $(X_1, Y_1), \ldots, (X_p, Y_p)$ from a common distribution, where for each i, $X_i \geq 0$, $Y_i \geq 0$ and $X_i + Y_i > 0$.

Let $x = E(X_1) > 0$ and $y = E(Y_1) > 0$. The relative fishing power θ is defined as

$$\theta = \frac{E(X_1 \mid X_1 > 0)}{E(Y_1 \mid Y_1 > 0)}.$$

It is natural to take $\hat{\theta} = \tilde{X}/\tilde{Y}$ as an estimate of θ, where

$$\tilde{X} = \left(\sum_{i=1}^{p} X_i \right) / (\text{no. of non-zero } X_j)$$

and

$$\tilde{Y} = \left(\sum_{i=1}^{p} Y_i \right) / (\text{no. of non-zero } Y_j).$$

($\hat{\theta}$ is defined to be zero if either all $X_j = 0$ or if all $Y_j = 0$.)

The main object of the paper is to find out how good this estimator is. In the next section we study the asymptotic properties of the estimator.

Main results

We start with some notation. Let $N_i=1$ if $X_i>0$ and zero if $X_i=0$ and let $M_i=1$ if $Y_i>0$ and zero if $Y_i=0$. Let $\bar{X}$, $\bar{Y}$, $\bar{M}$ and $\bar{N}$ denote the sample means of X_i, Y_i, M_i and N_i respectively. Note that $M_i+N_i-1=M_iN_i$ for all i, since we assumed that $X_i\geq 0$, $Y_i\geq 0$ and $X_i+Y_i>0$. If there is a catch of fish by a vessel then there is a minimum amount of fish caught (either in number or in weight). So it is natural to assume that for some $\delta>0$, $P(Y_1\geq\delta)+P(Y_1=0)=1$. Let $X_0=(\bar{X}-x)/x$, $Y_0=(\bar{Y}-y)/y$, $M_0=(\bar{M}-m)/m$ and $N_0=(\bar{N}-n)/n$, where $n=P(X_1>0)$, $m=P(Y_1>0)$. For any $s\geq 2$, let

$$\varrho_s=E|X_1/x|^s+E|Y_1/y|^s.$$

Theorem 1. *Suppose $\varrho_6<\infty$ and for some $\delta>0$, $m\geq\delta$, $n\geq\delta$ and $P(Y_1\geq\delta)+P(Y_1=0)=1$. Then*

$$E(\tilde{X}/\tilde{Y})=\theta+(a/p)+O(p^{-2}), \tag{1}$$

where

$$a=\theta(\sigma_{xm}+\sigma_{ny}-\sigma_{xy}-\sigma_{my}+\sigma_y^2-\sigma_{mn}-\sigma_{xn}+\sigma_n^2),$$

$$\sigma_{xm}=(1/xm)\,\mathrm{Cov}(X_1,M_1), \qquad \sigma_y^2=y^{-2}\,\mathrm{Var}(Y_1)$$

and the other σ's are defined similarly. The $O(\cdot)$ term depends only on $\delta>0$ and ϱ_6.

Proof. We first note that if $Z_1,\ldots,Z_p$ are i.i.d. random variables with mean zero and finite 6-the moment, then

$$E\left(\frac{1}{p}\sum_{i=1}^{p}Z_i\right)^6=O(p^{-3}E(Z_1^6)). \tag{2}$$

We also note that for $z\neq -1$,

$$(1+z)^{-1}=1-z+z^2-z^3+z^4-z^5(1+z)^{-1}. \tag{3}$$

Let

$$W_0=(X_0+M_0-Y_0-N_0)(1-Y_0-N_0+N_0Y_0+Y_0^2+N_0^2)$$
$$+(X_0M_0-N_0Y_0)(1-Y_0-N_0).$$

Let $Z_0=Y_0+N_0+Y_0N_0$. Using (3) and $m\geq\delta$, $n\geq\delta$ we obtain on $\bar{M}\bar{N}\neq 0$,

$$((\hat{\theta}-\theta)/\theta)=[(\bar{X}\bar{M}ny/\bar{Y}\bar{N}xm)-1]$$
$$=(X_0+M_0-Y_0-N_0+X_0M_0-N_0Y_0)(1+Y_0+N_0+N_0Y_0)^{-1}$$
$$=(X_0+M_0-Y_0-N_0+X_0M_0-N_0Y_0)(1-Z_0+Z_0^2-Z_0^3+Z_0^4-Z_0^5(1+Z_0)^{-1})$$
$$=[(X_0+M_0-Y_0-N_0)+(X_0M_0-N_0Y_0)][1-Z_0+Z_0^2+O[|Z_0|^3+|Z_0|^4$$
$$+|Z_0|^5(1+Z_0)^{-1}])$$

$$= (X_0+M_0-Y_0-N_0)(1-Y_0-N_0+N_0Y_0+Y_0^2+N_0^2+O[|Y_0^2N_0|+|N_0Y_0^2|])$$
$$+(X_0M_0-N_0Y_0)(1-Y_0-N_0+O[Y_0^2+N_0^2])$$
$$+O\Big\{[|X_0|+|\bar{M}_0|+|\bar{Y}_0|+|\bar{N}_0|]\Big[\sum_{j=3}^{4}(|Y_0|^j+|N_0|^j)$$
$$+ny(|Y_0|^5+|N_0|^5)/\bar{Y}\bar{N}\Big]\Big\}$$
$$= W_0+O\Big\{[|X_0|+|M_0+|Y_0|+|N_0|]$$
$$\times\Big[\sum_{j=3}^{4}(|Y_0|^j+|N_0|^j+ny((|Y_0|^5+|N_0|^5)/\bar{Y}\bar{N})\Big]\Big\}. \quad (4)$$

Observe that

$$|xmyE(X_0M_0Y_0)| = |p^{-2}E(X_1-x)(Y_1-y)(M_1-m)|$$
$$\leq p^{-2}E|(X_1-x)(Y_1-y)|$$
$$\leq p^{-2}(\mathrm{Var}(X_1)\mathrm{Var}(Y_1))^{1/2} \leq p^{-2}\sigma_x\sigma_y << p^{-2}. \quad (5)$$

We also have by (2), and Hölders inequality that, for $j \leq 5$,

$$E(|X_0||Y_0|^j) \leq (E|X_0|^6)^{1/6}(E(|Y_0|^{6j/5}))^{5/6} << p^{-(j+1)/2}. \quad (6)$$

Also note that

$$E|X_0M_0Y_0^2| \leq \{E(X_0^2M_0^2)E(Y_0^4)\}^{1/2}$$
$$\leq [E(X_0^6)E(M_0^6)]^{1/6}(E(Y_0^6))^{1/3} = O(p^{-3}). \quad (7)$$

By (4), (5), (6) and (7) we have

$$E(|((\hat{\theta}-\theta)/\theta)-W_0|I(\bar{M}\bar{N}\neq 0)) << E(|W_0|I(\bar{M}\bar{N}=0))+p^{-2}. \quad (8)$$

Since $m \geq \delta$ and $n \geq \delta$,

$$P(\bar{M}=0)+P(\bar{N}=0) \leq (1-m)^p+(1-n)^p \leq e^{-pm}+e^{-pn} \leq 2e^{-p\delta}. \quad (9)$$

So we have, with probability $\geq 1-2e^{-p\delta}$, that $\bar{Y}\bar{N} \geq (\delta/2p)$. By (9),

$$E(|W_0|I(\bar{M}\bar{N}=0)) << \sqrt{E(W_0^2)P(\bar{M}\bar{N}=0)} << e^{-\delta p/2} << p^{-2}. \quad (10)$$

The theorem now follows from (4)–(10) and the inequalities similar to (5), (6) and (7) for the other terms appearing in (4).

Remark. The estimate (1) still holds with an error term $O(p^{-3/2})$ instead if we assume $\varrho_5 < \infty$.

Theorem 2. *Suppose $\varrho_8 < \infty$ and the other conditions of Theorem 1 hold. Then the mean square error of the estimator $\hat{\theta}$ of θ satisfies*

$$pE(\hat{\theta}-\theta)^2 = b^2+O(p^{-1}), \quad (11)$$

where

$$b^2 = (xm/ny)^2 \operatorname{Var}\left(\frac{X_1}{x} + \frac{M_1}{m} - \frac{Y_1}{y} - \frac{N_1}{n}\right).$$

Proof. We have for $z \neq -1$,

$$\begin{aligned}(1+z)^{-2} &= 1-2z+3z^2-4z^3+5z^4-(6z^5+5z^6)(1+z)^{-2} \\ &= 1-2z+O(z^2+z^4+|z|^5(1+z)^{-1}+z^6(1+z)^{-2}).\end{aligned}$$

Since $z^2+|z|^3+z^4 = z^2(1+|z|+z^2) \leq 2(z^2(1+z^2))$, the term z^3 can be suppressed. As in the proof of Theorem 1, we have on $\bar{M}\bar{N} \neq 0$,

$$\begin{aligned}((\hat{\theta}-\theta)/\theta)^2 = (X_0+M_0-Y_0-N_0+X_0M_0-N_0Y_0)^2[1-2Y_0-2N_0+O(|N_0Y_0|) \\ +O(Y_0^2+N_0^2+Y_0^4+N_0{}^4+((|Y_0|^5+|N_0|^5)/\bar{Y}\bar{N}) \\ +O((Y_0^6+N_0^6)/(\bar{Y}\bar{N})^2)].\end{aligned}$$

As in the proof of Theorem 1, we obtain, using (2), (5), (6), (7), (9) and similar inequalities that

$$\begin{aligned}E((\hat{\theta}-\theta)/\theta)^2 &= E(X_0+M_0-Y_0-N_0)^2+O(p^{-2}) \\ &\quad +O(E((X_0+M_0-Y_0-N_0)^2 I(\bar{M}\bar{N}=0))) \\ &= p^{-1}\left(\operatorname{Var}\left(\frac{X_1}{x}+\frac{M_1}{m}-\frac{Y_1}{y}-\frac{N_1}{n}\right)\right)+O(p^{-2}).\end{aligned}$$

Remark. The estimate (11) still holds with an error term $O(p^{-1/2})$ if instead we assume $\varrho_7 < \infty$.

The main term b^2 in (11) is not known in general. But this can be estimated by several methods like bootstrap, jackknife or simply by

$$s_p^2 = p^{-1}(\bar{X}\bar{M}/\bar{N}\bar{Y})^2\left(\frac{1}{p}\sum_{i=1}^{p}\left(\frac{\bar{X}_i}{\bar{X}}+\frac{M_i}{\bar{M}}-\frac{Y_i}{\bar{Y}}-\frac{N_i}{\bar{N}}\right)^2\right).$$

It is not difficult to see that the bootstrap as well as jackknife estimate of the mean square error is $s_p^2 + O(p^{-2})$ almost surely under $\varrho_8 < \infty$.

From the proofs given above it is clear that $\sqrt{p}(\hat{\theta}-\theta)$ is asymptotically distributed as normal with mean zero and variance b^2. In fact we can obtain the order of the eror term in this approximation.

Theorem 3. *If $\varrho_3 < \infty$ and if $b > 0$, then*

$$\sup_x |P(\sqrt{p}(\hat{\theta}-\theta) \leq xb) - \Phi(x)| << p^{-1/2}\log p, \tag{12}$$

where Φ denotes the standard normal distribution function.

Remark. The extra $\log p$ term in (12) appears because of the approximation of $\hat{\theta}-\theta$ by a sample mean of i.i.d. random variables.

To prove this we require the following moderate deviations result.

Lemma. *Let $\{Z_i\}$ be i.i.d. mean zero random variables with variance 1 and $E|Z_1|^r < \infty$. Then*

$$P\left(\left|\sum_{i=1}^{r} Z_i\right| > \sqrt{r \log r}\right) << r^{-1/2}. \tag{13}$$

This is a trivial consequence of Theorem 4 of Michel (1974).

Proof of Theorem 3. Observe that on $\bar{M}\bar{N} \neq 0$,

$$\begin{aligned}((\hat{\theta}-\theta)/\theta) &= (X_0+M_0-Y_0-N_0)\left(1+\frac{ny-\bar{N}\bar{Y}}{\bar{N}\bar{Y}}\right)+(M_0X_0-N_0Y_0)\frac{ny}{\bar{N}\bar{Y}} \\ &= (X_0+M_0-Y_0-N_0)-W_p(ny/\bar{N}\bar{Y}),\end{aligned}$$

where

$$W_p = [(X_0+M_0-Y_0-N_0)(Y_0+N_0+Y_0N_0)+N_0Y_0-M_0X_0].$$

By Chebyshev's inequality, we have

$$P(\bar{M} < \tfrac{1}{2}m)+P(\bar{N} < \tfrac{1}{2}m) << \frac{1}{p}. \tag{14}$$

Note that on $\bar{M} \geq \frac{1}{2}m$ and $\bar{N} \geq \frac{1}{2}n$,

$$\bar{Y}\bar{N} \geq (\tfrac{1}{4}nm\delta). \tag{15}$$

Also we have

$$|W_p| \leq 2(|X_0|+|Y_0|+|M_0|+|N_0|)(|Y_0|+|N_0|)+|N_0Y_0|+|M_0X_0|.$$

If $\varrho_3 < \infty$ then by the lemma, we have for some $A > 0$,

$$\begin{aligned}P(p|W_p| > 18A^2 \log p) << {} & P(\sqrt{p}|X_0| > A\sqrt{\log p}) \\ & + P(\sqrt{p}|Y_0| > A\sqrt{\log p}) \\ & + P(\sqrt{p}|N_0| > A\sqrt{\log p}) \\ & + P(\sqrt{p}|M_0| > A\sqrt{\log p}) << p^{-1/2}.\end{aligned} \tag{16}$$

The result now follows from the Berry–Esseen Theorem (see Theorem 12.4 of Bhattacharya and Rao (1976)), (9), (14), (15) and (16).

Data analysis

A total of 20 species were identified for the analysis. Table 1 gives the results of the analysis.

Table 1
Catch in numbers

Species	Fishing power	Mean sq. Error	Bias	No. of nonzero catches		
				Albat.	Delaware	At least one ship
Smooth Dogfish	1.1563	0.0554	0.0227	16	18	24
Little Skate	0.6675	0.0097	0.0256	76	72	93
Silver Hake	1.8075	0.1508	0.0135	111	108	124
White Hake	0.6831	0.0759	0.0760	20	20	32
Spotted Hake	0.7109	0.0295	0.0237	24	24	32
Fourspot Flounder	1.1219	0.0920	0.0129	64	69	80
Yellowtail Flounder	0.7116	0.0099	0.0003	48	50	55
Winter Flounder	1.2644	0.0212	0.0097	48	60	64
Windowpane	1.2169	0.0264	0.0023	48	59	71
Butterfish	1.2990	0.0678	0.0301	95	98	112
Scup	1.1400	0.0892	0.0554	32	32	43
Longhorn Sculpin	0.7320	0.0112	0.0153	45	39	48
Sea Raven	1.3792	0.2007	0.0506	27	34	43
Northern Sea Robin	1.0506	0.0416	0.0168	25	20	32
Ocean Pout	0.8111	0.0193	0.0195	17	19	22
Goose Fish	0.6000	0.0419	0.0450	24	32	43
Amer. Lobster	1.0575	0.0201	0.0046	58	61	76
Sea Scallop	0.9128	0.0297	0.0620	28	25	34
Shortfin Squid	0.4889	0.0146	0.0218	74	81	91
Longfin Squid	1.1650	0.0483	0.0188	96	97	113

Acknowledgements

The author would like to express his thanks to Mike Pennington of Northeast Fisheries Center, Woods Hole Laboratory, for the discussions he had with him, and Mike Fogarty and Byrne of the same laboratory for the data set and for the preprint mentioned in the references. Critical comments and suggestions of the referee helped a great deal in improving the paper.

References

Bhattacharya, R.N. and R. Ranga Rao (1976). *Normal Approximation and Asymptotic Expansions.* John Wiley & Sons, New York.

Byrne, C.J. and M.J. Fogarty (1985). Comparison of the fishing power of two fisheries research vessels. NAFO SCR DOC. 85/90 Serial No. N1065. North Atlantic Fisheries Organization.

Michel, R. (1974). Results on probabilities of moderate deviations. *Ann. Probab.* 2, 349–353.