Estimating input sample size for length-frequency data in Stock Synthesis: US longline and US troll fisheries [1]

Hui-Hua Lee

Joint Institute for Marine and Atmospheric Research, University of Hawaii

Pacific Islands Fisheries Science Center

2570 Dole Street, Honolulu, HI 96822, USA

Email: huihua.lee@noaa.gov

[1] Working document submitted to the ISC Albacore Working Group Workshop, 12-19 October 2010, La Jolla, California, USA. The views and opinions expressed or implied in this working paper are those of the author(s) and do not necessarily reflect the position of the author(s) agency(ies) or government(s). Data and results presented in this working paper should be viewed as preliminary and therefore subject to change. This working paper should not be cited without authors consent.

Estimating input sample size for length-frequency data in Stock Synthesis: US longline and US troll fisheries

Hui-Hua Lee

Introduction

In commercial fisheries, the sample of fish of a particular species measured is usually not a random sample of individual fish from the entire population but a sample of *n* clusters (trips or sets). Fish caught together under cluster designs tend to have more similar characteristics, such as length or age, than those in the entire population. Variance within cluster is small but large among clusters. Therefore, a total of *m* fish collected from *n* clusters will contain less information about the population length distribution than *m* fish sampled randomly from population.

One way to measure the information contained in a sample of length measurements is to estimate the number of fish that one would need to sample at random (effective sample size) to obtain the same information on length contained in the cluster samples. In stock synthesis, variability of a length composition is in terms of an effective sample size that can be an input value or iteratively tuned inside the model to achieve internal model consistency. Effective sample size is then the input sample size multiplied by the fishery-specific variance adjustment. However, the determining input sample size is usually somewhat arbitrary. By comparing the variance of the estimator under cluster sampling with the variance of the same estimator under simple random sampling, actual sample size being measured under cluster designs can be adjusted to derive the effective sample size.

Materials and methods

Data used

For the US albacore troll fishery in the North Pacific, two main sources of data were used to assess the precision of length-frequency: 1) actual length sample by trip from a port sampling program and 2) catch information by trip from fishermen logbooks.

For the US longline fishery in the North Pacific, two main sources of data were used to assess the precision of length-frequency: 1) actual length sample by set and trip from an on-board observer program and 2) catch information by set and trip from fishermen logbooks.

Assessing the precision of length-frequency estimates

Mean length was used as an estimator to represent a sample of length distribution (Pennington *et al.* 2002). Suppose population mean length is similar in a year and a random sample of *n* clusters is chosen. In the case of cluster sampling, the population mean length, $\widehat{R}$, is a ratio estimator from sample average length.

$$\widehat{R} = \frac{\sum_{i=1}^{n} M_i \widehat{\mu}_i}{\sum_{i=1}^{n} M_i}, \tag{1}$$

where M_i is the number of fish caught (either actual or estimated) in trip *i* and $\widehat{\mu}_i$ is an estimate of the average length of fish measured in trip *i*.

The variance of population mean length, var($\widehat{R}$), is approximated by

$$\text{var}(\widehat{R}) = \sum_{i=1}^{n} \frac{\left(M_i / \overline{M}\right)^2 \left(\widehat{\mu}_i - \widehat{R}\right)^2}{n(n-1)}, \tag{2}$$

where $\overline{M} = \sum_{i=1}^{n} Mi / n$ is mean number of fish caught for a trip.

Next estimate the variance, $\widehat{\sigma}_x^2$, of the population mean length if m_i fish are randomly measured in each trip (or if all fish are measured). Then

$$\widehat{\sigma}_x^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} \left(M_i / m_i\right)\left(x_{i,j} - \widehat{R}\right)^2}{M - 1}, \tag{3}$$

where $M = \sum_{i=1}^{n} Mi$ is the total number of fish caught in a year and $x_{i,j}$ is the length of the j^{th} fish in trip *i*.

If it were possible to sample *m* fish at random from the population, then the variance of the sample mean would be equal to $\hat{\sigma}_x^2 / m$. The ratio of the variance of the population mean length under cluster design to the variance of the same estimator under simple random sampling is called design effect (*deff*):

$$deff = \frac{\text{var}(\hat{R})}{\hat{\sigma}_x^2/m}, \quad (4)$$

The effective sample, m_{eff}, is then estimated from the actual sample size, m_{act}, adjusted by the design effect as follows.

$$m_{eff} = m_{act}/deff, \quad (5)$$

In other word, if m_{act} fish is sampled randomly, the design effect would be 1. This also implies that the sample mean would have the same precision as an estimate based on a sample of *n* clusters.

Results and discussion

Estimates of the design effect and associated statistics for estimates of albacore length composition are presented in Table 1 and 2 for US troll and US longline fisheries, respectively. The results indicated that the variance of the population mean length under cluster sampling design (var($\hat{R}$)) is larger than the variance of the same estimator under simple random sampling ($\hat{\sigma}_x^2/m$), implying that the estimates of the length distribution from cluster sampling were less precise. In other words, if the actual sample size was used, the variance of the population mean length does not represent the variance from random samples. It is noted that for US troll fisheries, the number of trips used in a year from 1973 to 1976 were small, therefore the estimates of design effect in those years were excluded.

The total number of fish measured by each stratum (fishery/year/quarter) was adjusted by the annual design effect. The results are shown in Table 3 and Table 4 for US troll and US longline fisheries, respectively. The assumption was made that strata with small sample size (less than *deff*) do not represent the entire distribution (e.g.. season 4 in 1982 for US troll fisheries). The input sample size for strata with large sample sizes (actual sample size is over 1,000 for US troll and 250 for US longline fisheries) was defined as 1,000/*deff* or 250/*deff* for US troll and US longline fisheries, respectively. The maximum actual sample size for large sample sizes is chosen arbitrarily so that the input sample size would not exceed the number of trips (Fig. 1). Otherwise, the input sample size was defined as the total number of fish measured divided by *deff*. For US troll fisheries in years 1966-1976 and 1994, the average design effect (1977-2008) was applied to estimate input sample size. The results also showed that smaller input sample sizes leading to less precise length data (ex. 1982 for US troll and 1998 for US longline fisheries). Effective sample sizes (for use in the multinomial error assumption) for each fleet were initially estimated as above and can be iteratively tuned inside the model to achieve internal model consistency (input sample sizes times fishery-specific variance adjustment).

In this paper, the estimation of input sample size is merely a proxy of effective sample size that represents a randomly sampled length distribution. Mean length may not be the best way to represent a length distribution. For example, Gomez-Buckley *et al.* (1999) used cumulative frequency distributions (cdf) from a sample as the representative statistic rather than mean length. An underlying assumption of comparing the annual variances of the mean length under different sampling designs is that seasonal growth is negligible. When seasonal variability is taken into consideration, the input sample size could become larger and some strata did not have enough trips to estimate the sample effect.

Reference

M. Pennington, L.-M. Burmeister, V. Hjellvik. 2002. Assessing the precision of frequency distributions estimated from trawl-survey samples. Fishery Bulletin. 100:74–80.

M. Gomez-Buckley, L. Conquest, S. Zitzer, B. Miller. 1999. Use of statistical bootstrapping for sample size determination to estimate length-frequency distributions for Pacific albacore tuna (*Thunnus alalunga*). FRI-UW-9902. Fisheries Research Institute, School of Fisheries, University of Washington. http://www.fish.washington.edu/research/Publications/pdfs/9902.pdf

Table 1. Summary statistics for assessing the precision of the estimated length distributions of North Pacific albacore caught by the US troll fisheries. The estimated design effect is denoted by *deff* due to the cluster sampling design, *n* is the number of trips at which albacore were caught, *M* is the number of albacore caught, *m* is the number measured, var($\hat{R}$) is the variance estimate of mean length under cluster sampling, and $\hat{\sigma}_x^2/m$ is the variance of sample mean length if fish are randomly measured.

year	*n*	*M*	m	var($\hat{R}$)	$\hat{\sigma}_x^2/m$	*deff*
1973	4	14319	274	0.03	0.139	0.19
1974	6	5213	368	2.71	0.123	22.05
1975	2	1770	99	0.02	0.259	0.08
1976	4	5027	233	3.99	0.172	23.14
1977	272	118835	14517	0.23	0.005	47.19
1978	357	204577	17066	0.30	0.005	59.97
1979	134	81150	6182	0.76	0.013	57.88
1980	109	511639	6348	2.11	0.016	130.48
1981	282	260655	15955	0.13	0.004	30.64
1982	136	119262	8395	8.12	0.015	539.67
1983	296	362528	18969	0.43	0.004	117.03
1984	278	553037	20778	0.63	0.003	218.98
1985	259	311693	18962	0.23	0.003	65.30
1986	114	137770	9957	1.77	0.011	166.67
1987	140	91243	14549	0.11	0.003	41.03
1988	148	365216	12760	0.11	0.002	64.42
1989	60	18471	2767	1.23	0.020	60.15
1990	93	843326	12867	2.86	0.005	614.58
1991	42	83642	5174	0.40	0.013	31.57
1992	133	216449	11774	0.96	0.002	466.42
1995	126	1025444	10414	1.52	0.010	157.16
1996	105	446088	14817	0.05	0.002	23.63
1997	111	262131	12188	1.46	0.005	284.75
1998	36	140274	4038	0.46	0.011	40.67
1999	70	153763	5114	1.43	0.008	172.58
2000	68	349891	5489	1.38	0.009	157.51
2001	98	212582	6770	0.46	0.007	68.15
2002	61	135848	5234	0.47	0.007	69.74
2003	56	157322	4658	0.40	0.012	31.80
2004	241	1150593	19347	0.17	0.001	113.72
2005	261	721465	17527	0.41	0.004	93.35
2006	288	1135928	23657	0.06	0.001	64.57
2007	299	956814	23021	0.18	0.002	86.60
2008	126	306517	11360	0.74	0.004	172.53

Table 2. Summary statistics for assessing the precision of the estimated length distributions of North Pacific albacore caught by the US longline fisheries. The estimated design effect is denoted by *deff* due to the cluster sampling design, *n* is the number of trips at which albacore were caught, *M* is the number of albacore caught, *m* is the number measured, var($\hat{R}$) is the variance estimate of mean length under cluster sampling, and $\hat{\sigma}_x^2/m$ is the variance of sample mean length if fish are randomly measured.

year	*n*	*M*	m	var($\hat{R}$)	$\hat{\sigma}_x^2/m$	*deff*
1994	37	6537	1791	2.94	0.065	45.17
1995	42	5617	1623	2.45	0.089	27.45
1996	51	7365	2125	2.62	0.052	50.83
1997	36	6048	2573	2.60	0.045	58.27
1998	43	4232	2163	6.53	0.101	64.86
1999	40	6951	3327	3.65	0.049	74.84
2000	60	5022	2355	0.83	0.040	20.75
2001	202	21219	10506	0.13	0.007	18.78
2002	223	9802	4823	0.27	0.013	21.40
2003	171	7974	3863	0.10	0.014	6.94
2004	260	9024	4357	0.58	0.016	37.20
2005	308	6742	3501	1.66	0.054	30.77
2006	193	5307	1424	0.94	0.106	8.84
2007	179	6745	1242	3.18	0.199	15.95
2008	12	754	137	23.87	1.511	15.80

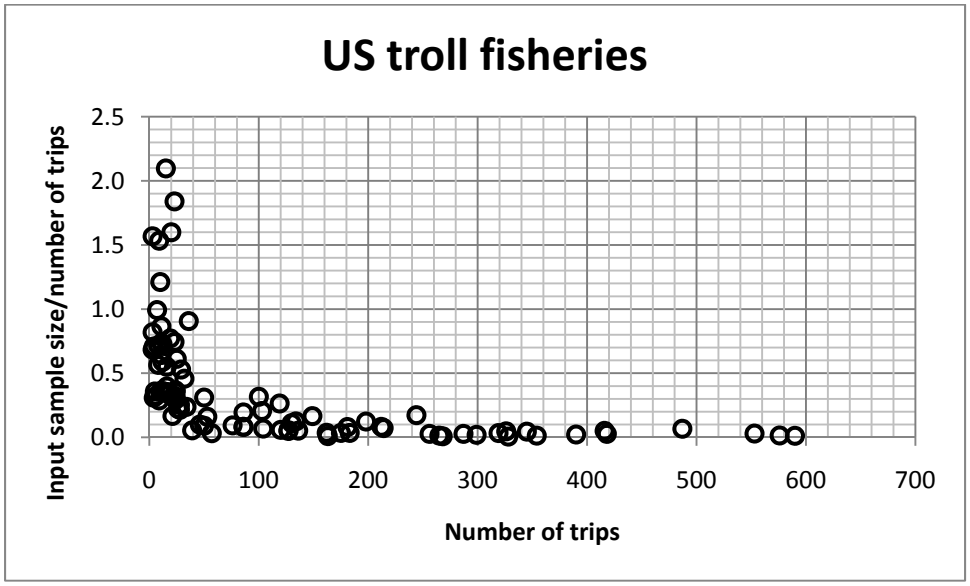
Table 3. The estimated input sample size for each stratum (fishery/year/quarter) for North Pacific albacore caught by the US troll fisheries. The input sample size was estimated as the number of fish measured divided by estimated yearly design effect from Table 1 due to non random samples. For years 1966-1976 and 1994, the average design effect (1977-2008) was applied to estimate input sample size.

Year	Season	Number of fish measured	Number of trips	Input sample size
1966	3	7490	127	7.06
1966	4	175	4	1.24
1967	3	5886	86	7.06
1967	4	800	16	5.65
1968	3	6872	104	7.06
1968	4	749	15	5.29
1969	3	4797	86	7.06
1969	4	1150	23	7.06
1970	3	1257	12	7.06
1971	3	1100	11	7.06
1972	3	5452	22	7.06
1973	2	253	5	1.79
1973	3	23747	256	7.06
1973	4	497	10	3.51
1974	3	16911	136	7.06
1974	4	931	19	6.57
1975	3	17134	120	7.06
1975	4	1470	28	7.06
1976	2	1078	24	7.06
1976	3	36042	590	7.06
1976	4	3578	76	7.06
1977	2	328	7	6.95
1977	3	30242	416	21.19
1977	4	4649	103	21.19
1978	2	271	8	4.52
1978	3	31711	553	16.68
1978	4	3433	86	16.68
1979	3	10811	212	17.28
1979	4	989	23	17.09
1980	3	17428	287	7.66
1980	4	1003	24	7.66
1981	2	1766	36	32.63
1981	3	28703	487	32.63
1986	4	1797	28	6.00
1987	3	20152	198	24.37
1987	4	601	32	14.65
1988	3	21287	214	15.52
1988	4	520	34	8.07
1989	3	10581	134	16.63
1990	3	24292	163	1.63
1991	3	12442	100	31.67
1991	4	255	11	8.08
1992	3	23375	268	2.14
1992	4	2545	39	2.14
1994	3	666	3	4.70
1994	4	401	4	2.83
1995	3	15746	183	6.36
1995	4	407	9	2.59
1996	3	32384	244	42.32
1996	4	2531	23	42.32
1997	2	552	6	1.94
1997	3	29068	265	3.51
1997	4	1704	21	3.51
1998	3	15173	149	24.59
1998	4	561	9	13.79
1999	2	1717	17	5.79
1999	3	12097	162	5.79
1999	4	862	14	4.99
2000	3	10432	127	6.35
2000	4	1204	16	6.35
2001	3	13072	181	14.67
2001	4	1760	19	14.67
2002	2	664	11	9.52
2002	3	10924	130	14.34
2002	4	845	10	12.12
2003	3	10546	119	31.45

1981	4	980	20	31.98	2003	4	1266	15	31.45
1982	2	2935	57	1.85	2004	2	1802	24	8.79
1982	3	25208	328	1.85	2004	3	30841	390	8.79
1982	4	156	4	NA	2004	4	1173	16	8.79
1983	2	1748	53	8.54	2005	2	813	12	8.71
1983	3	42379	576	8.54	2005	3	22074	319	10.71
1983	4	1362	24	8.54	2005	4	535	8	5.73
1984	2	2418	50	4.57	2006	3	28853	345	15.49
1984	3	28071	354	4.57	2006	4	3494	50	15.49
1984	4	2524	46	4.57	2007	2	213	3	2.46
1985	2	1418	25	15.31	2007	3	32221	418	11.55
1985	3	26562	326	15.31	2007	4	178	3	2.06
1985	4	2468	29	15.31	2008	3	25417	299	5.80
1986	2	667	11	4.00	2008	4	1930	26	5.80
1986	3	16174	175	6.00					

Table 4. The estimated input sample size for each stratum (fishery/year/quarter) for North Pacific albacore caught by the US longline fisheries. The input sample size was estimated as the number of fish measured divided by estimated yearly design effect from Table 3 due to non random samples.

Year	Season	Number of fish measured	Number of trips	Input sample size
1994	1	104	11	2.30
1994	3	176	7	3.90
1994	4	1471	12	5.54
1995	1	641	14	9.11
1995	2	233	13	8.49
1995	3	267	10	9.11
1995	4	482	9	9.11
1996	1	494	16	4.92
1996	2	1075	20	4.92
1996	3	136	9	2.68
1996	4	420	12	4.92
1997	1	779	12	4.29
1997	2	1348	11	4.29
1997	4	399	9	4.29
1998	1	222	10	3.42
1998	2	158	3	2.44
1998	3	425	12	3.85
1998	4	1197	18	3.85
1999	1	1529	14	3.34
1999	2	340	11	3.34
1999	3	130	3	1.74
1999	4	1419	12	3.34
2000	1	292	18	12.05
2000	2	122	9	5.88
2000	3	643	11	12.05
2000	4	3024	64	12.05
2001	1	4059	75	13.31
2001	2	3866	31	13.31
2001	3	1355	43	13.31
2001	4	1375	74	13.31
2002	1	876	77	11.68
2002	2	2868	59	11.68
2002	3	669	43	11.68
2002	4	450	54	11.68
2003	1	793	64	36.01
2003	2	2799	50	36.01
2003	3	97	26	13.97
2003	4	158	36	22.76
2004	1	702	55	6.72
2004	2	394	54	6.72
2004	3	1952	74	6.72
2004	4	1121	86	6.72
2005	1	1488	97	8.13
2005	2	996	73	8.13
2005	3	302	77	8.13
2005	4	882	88	8.13
2006	1	868	86	28.27
2006	2	340	63	28.27
2006	3	235	44	26.57
2006	4	163	26	18.43
2007	1	545	58	15.67
2007	2	92	30	5.77
2007	3	62	21	3.89
2007	4	499	75	15.67
2008	1	951	92	15.83
2008	2	343	63	15.83
2008	3	228	35	14.43
2008	4	440	29	15.83



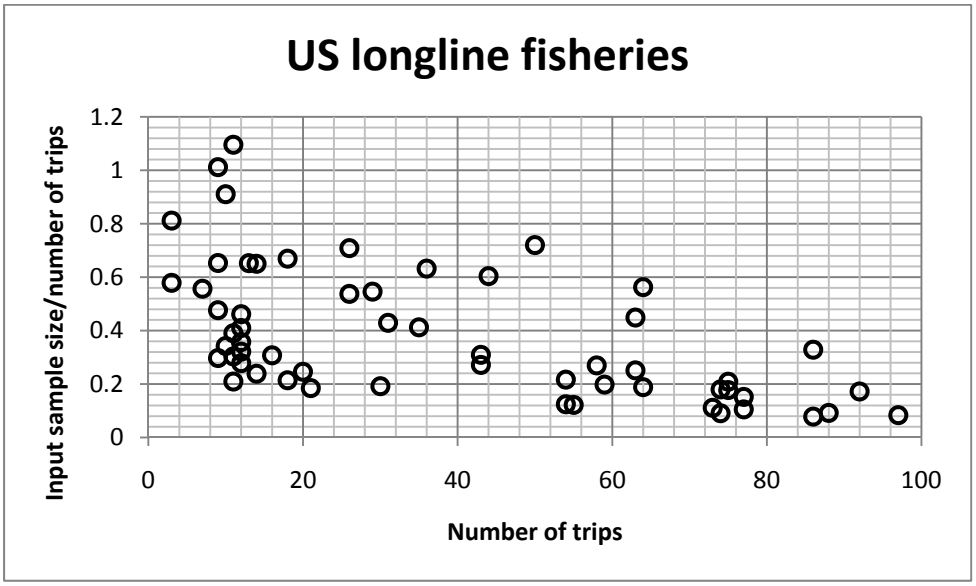


Figure 1. The relationship between number of trips and estimated input sample size per trip. Most of the estimated input sample size per trip are below 1 and the more trips involved in the sampling, smaller input sample size per trip estimated.