NRC Research Press

ARTICLE

Bootstrapping of sample sizes for length- or age-composition data used in stock assessments

Ian J. Stewart and Owen S. Hamel

Abstract: Integrated stock assessment models derive estimates of management quantities by fitting to indices of abundance and length and age compositions. For composition data, where a multinomial likelihood is often applied, weights are determined by input sample sizes, which can be an important contributor to model results. We used a generic bootstrap method, verified through simulation, to calculate year-specific maximum realized sample sizes from the observation error inherent in fishery biological data. Applying this method to length-composition observations for 47 groundfish species collected during a standardized trawl survey, we found maximum realized sample size to be related to both the number of hauls and individual fish sampled from those hauls. Sampling in excess of 20 fish from each haul produced little increase in most cases, with maximum realized sample size ranging from approximately 2 to 4 per haul sampled. Utilizing these maximum realized sample sizes as input values for stock assessment (analogous to minimum variance estimates) appropriately incorporates interannual variability, and may reduce over-emphasis on composition data. Results from this method can also help determine sampling targets.

Résumé : Les modèles d'évaluation intégrée des stocks permettent d'établir des estimations des quantités gérées en les calant sur des données d'observation, dont les indices d'abondances et les compositions par longueur et par âge. En ce qui concerne les données de composition, auxquelles une probabilité multinomiale est souvent appliquée, les facteurs de pondération sont déterminés par les tailles des échantillons d'entrée, qui peuvent influencer considérablement les résultats du modèle. Nous avons utilisé une méthode d'autoamorçage générique, vérifiée par simulation, pour calculer les plus grandes tailles d'échantillon réalisées selon l'année à partir de l'erreur d'observation inhérente aux données biologiques sur les pêches. En appliquant cette méthode à des observations de composition selon la longueur pour 47 espèces de poissons de fond recueillies dans le cadre d'un relevé au chalut standardisé, nous avons noté que la plus grande taille d'échantillon réalisée était reliée au nombre de traits et aux différents poissons prélevés de ces traits. Le prélèvement de plus de 20 poissons de chaque trait ne produisait qu'une faible augmentation dans la plupart des cas, la plus grande taille d'échantillon réalisée variant de 2 à 4 environ par trait échantillonné. L'utilisation de ces plus grandes tailles d'échantillon réalisées comme valeurs d'entrée pour l'évaluation des stocks (à l'instar d'estimations de la variance minimum) permet une intégration adéquate de la variabilité interannuelle et pourrait réduire l'importance indue conférée aux données de composition. Les résultats produits par cette méthode peuvent également aider à la détermination de cibles d'échantillonnage. [Traduit par la Rédaction]

Introduction

Integrated statistical fisheries stock assessment represents the state-of-the-art for estimating population size and status after a century of fisheries science. From the development of the conceptual basis for these models (Fournier and Archibald 1982), methods have evolved for diverse applications (Megrey 1989). Rapid generalization allowed for statistical fitting to many types of auxiliary data, including indices of abundance and biological information (Quinn 2003). Basic methodology for the general theory of integrated age-structured population models (Hilborn and Walters 1992; Quinn and Deriso 1999) is now widely applied (Maunder and Punt 2013), and the role of uncertainty in stock assessment has become more important (Patterson et al. 2001).

Integrated stock assessment models derive estimates of management quantities, and associated estimates of uncertainty in these quantities, by fitting to observed data, including age- and length-composition data from both fishery-dependent and fishery-independent sources. The relative weighting of each likelihood component is dictated by the variance or sample size assigned to each observation. These sample sizes may vary systematically by source and annually because of changes in sampling programs. Alternative approaches to setting input sample sizes for composition data, and methods for tuning these initial values, often result in different estimates of population parameters and may be an important component of uncertainty (Francis 2011; Hulson et al. 2012); however, there is currently no consensus regarding the best approach to treating input sample sizes.

Realized sample size, as used in this analysis, is not the number of fish measured or aged, but the statistical power of a sample in the context of the assumed error distribution. Multiple sources of variability contribute to the realized sample size from a particular sample. These sources include variability among lengths (or ages; while this analysis will focus exclusively on length, the same considerations apply to age-frequency data) both within a haul (or within a trip for port-sampled fishery data) and variability in the length distributions between hauls, both of which may differ among species and years. Conceptually, if all fish captured in a single haul were always of the same length, and all hauls caught the same number of fish, then one should sample only a single fish from each haul, and the realized sample size would be the

Received 29 May 2013. Accepted 14 January 2014.

Paper handled by Associate Editor Kenneth Rose.

I.J. Stewart* and O.S. Hamel. National Marine Fisheries Service, Northwest Fisheries Science Center, 2725 Montlake Boulevard East, Seattle, WA 98112, USA.
Corresponding author: Ian J. Stewart (e-mail: Ian@iphc.int).
*Present address: International Pacific Halibut Commission, 2320 West Commodore Way, Suite 300, Seattle, WA 98199-1287, USA.

number of hauls. In contrast, if all fish in all hauls are entirely independent of one another, the realized sample size would be a function only of the number of fish measured, regardless of the number of hauls. In the latter case, sampling more individual fish would increase the realized sample size in proportion to the complexity of the underlying distribution. Given that the reality must lie between these two extremes, realized sample size should depend upon the overlap of the geographic distribution of the species and sampling program and the modal structure of the population, as well as the sampling intensity (total number of hauls and individual lengths measured), which will vary annually as these factors vary.

There is an extensive body of analysis suggesting that, because of population clustering, the number of fishery survey hauls sampled is generally a good proxy and primary factor for determining the realized sample size in numbers at length or age (e.g., Pennington and Volstad 1994; Pennington et al. 2002; Aanes and Pennington 2003; Helle and Pennington 2004; Cerviño and Saborido-Rey 2006). Most of the variance in commercial catch-at-age (90%) was found to derive from the trip, rather than the individual fish, in a west coast US fishery sampling program (Crone 1995). The recommended basis for sampling targets has, therefore, been based on the ratio of within- to among-trip variability (Chih 2010). In some cases, the realized sample size may be even less than the number of hauls sampled despite thousands of individual fish observations (Pennington and Volstad 1994).

Determining the realized sample size that results from a survey or fishery sampling program is, therefore, fundamentally more complex than calculating the number of truly random samples needed to describe a particular length- or age-frequency distribution with a certain tolerance for sampling error. Several studies have thoroughly investigated how many random samples are required to summarize a length distribution with a reasonable level of accuracy. This number can range from 75 to 1200, depending upon the size range, modal complexity, and the degree of length bin aggregation (e.g., Worthington et al. 1995; Vokoun et al. 2001; Gerritsen and McGrath 2007).

Regardless of the underlying sampling program, stock assessment scientists are faced with making an assumption regarding the relative statistical weighting for each set of compositional data. Many assessments utilize a multinomial error distribution, which requires an input sample size rather than a variance. Guidance from the available literature suggests a wide range of approaches are taken to setting input sample sizes and for updating these values (or not) during the subsequent analysis. Despite the known presence of interannual variability, when estimating (or iterating) input sample sizes, it has often been convenient to set them equal across all years (Fournier et al. 1990). Methods range from setting fixed values ranging from 200 (Methot 2000) to 400 (Fournier and Archibald 1982), to using the number of fish sampled, but not exceeding a fixed cap (Methot 1989), which may be as large as 1000 (Fournier et al. 1998). More commonly, it is recognized that the sample size should be much smaller than the number of fish sampled, because of clustering by size and (or) age, with a value around 100 for fishery and survey data (McAllister and Ianelli 1997). Bootstrapping, based directly on sampling theory, has been shown to produce reasonably reliable stock assessment behavior (Hulson et al. 2012), although in some well-sampled (unpublished) applications, bootstrapped values may generate sample sizes that appear far in excess of common limits used in stock assessments.

There are two common approaches once initial values have been identified (Hulson et al. 2012): specify initial sample sizes and leave them alone, or tune the initial values, either iteratively or through direct estimation via an additive or multiplicative scaling factor or an alternative error distribution (Deriso et al. 2007; Maunder 2011). Both methods rely to some degree on reasonable initial values that adequately reflect the inherent differences among data sets and among years within data sets. Ultimately, model tuning aims to achieve the goal of internal consistency between assumed data weighting and model fit.

Given the lack of consistency in dealing with input sample sizes (and therefore the weighting of data sources) among many stock assessments, there is a distinct need for general guidance that can be applied to any fisheries data set. The multinomial distribution is a widely used error distribution for length- and age-composition data in integrated stock assessments (e.g., Methot and Wetzel, 2013). Following a bootstrapping approach very similar to Hulson et al. (2012), this study uses length-frequency observations from the Northwest Fisheries Science Center's annual west coast bottom trawl survey, collected from 2003 to 2010, to illustrate an objective method for deriving maximum realized sample sizes suitable for use as starting values in stock assessments. This analysis is based solely on the multinomial error assumption, but the general approach could be applied to any error distribution to be applied. Patterns observed among a large number of species are summarized to generate some general guidelines for applications where specific bootstrapped results may not be available.

Methods

Length-frequency calculations and bootstrapping

To estimate the length-frequency distribution for a particular species in a single year, a simple, design-based approach was used. Let the predicted number of fish ($\hat{N}$) in a haul (h) be denoted

$$\hat{N}_h = \frac{W_h}{W_s} \cdot N_s \quad (1)$$

where W_h is the total haul catch mass for that species and W_s is the total mass of all fish of that species enumerated in the subsample from the haul (N_s; not all fish that are counted are necessarily measured). In many cases the entire catch is enumerated, so $\frac{W_h}{W_s} = 1$, and therefore $\hat{N}_h = N_s$. If all fish in a haul (h) are measured, then the predicted numbers at length ($\hat{L}_{h,b}$) in each size bin ($b = 1$ to n, for n sizes, including separate bins for males and females) are equal to the observed numbers, $N_{h,b}$. Where not all fish have been measured, the proportions in each bin are applied to the predicted number of fish in that haul

$$\hat{L}_{h,b} = \hat{N}_h \cdot \frac{N_{h,b}}{\sum_b N_{h,b}} \quad (2)$$

Within each geographic region (r) defined by depth and latitude (possible in this case because depth is very closely correlated with longitude), the proportions by bin ($\hat{P}_{r,b}$) are then calculated as

$$\hat{P}_{r,b} = \frac{\sum_h \hat{L}_{h,b}}{\sum_h \sum_b \hat{L}_{h,b}} \quad (3)$$

and the total predicted numbers of fish as a function of the spatial area of the region (A_r), the area swept by each haul (a_h; although these are standardized as much as possible) conducted in that region, and the count of all hauls in the region (C_r, including those that captured no fish of the target species)

Table 1. Summary statistics for flatfish and "other" ground fish species.

			Length bins					Effective *N* per year		
Group	Species (area)	Scientific name	No.	Range (cm)	Mean hauls per stratum	Mean hauls per year	Mean fish per year	Min.	Mean	Max.
Flatfish	Arrowtooth flounder	*Atheresthes stomias*	35	13–80	36	213	3534	332	643	1160
	Curlfin sole	*Pleuronichthys decurrens*	24	11–34	12	63	407	109	176	277
	Dover sole	*Microstomus pacificus*	26	15–64	35	523	11 945	519	1572	2369
	English sole	*Parophrys vetulus*	18	12–45	13	125	2803	177	340	718
	Flathead sole	*Hippoglossoides elassodon*	36	13–48	10	41	493	50	153	313
	Pacific sanddab	*Citharichthys sordidus*	31	5–35	30	201	6519	104	620	1707
	Petrale sole	*Eopsetta jordani*	25	15–62	17	261	3724	428	1086	1894
	Rex sole	*Glyptocephalus zachirus*	26	3–52	26	362	9522	412	1011	1487
Others	Big skate	*Raja binoculata*	38	9–190	12	83	262	79	137	241
	California scorpionfish	*Scorpaena guttata*	59	1–59	10	13	183	23	70	143
	California skate	*Raja inornata*	34	13–78	14	66	350	23	144	248
	Lingcod	*Ophiodon elongatus*	42	29–110	32	192	1197	163	314	627
	Longnose skate	*Raja rhina*	27	19–145	25	367	3223	68	615	1350
	Pacific cod	*Gadus macrocephalus*	30	21–78	8	34	281	35	93	217
	Pacific flatnose	*Antimora microlepis*	25	10–49	23	113	1277	257	431	645
	Pacific grenadier	*Coryphaenoides acrolepis*	31	2–32	17	125	3023	378	589	805
	Pacific hake	*Merluccius productus*	51	20–70	21	295	5168	58	250	576
	Sablefish	*Anoplopoma fimbria*	31	31–90	28	413	4622	410	894	1381
	Spiny dogfish	*Squalus acanthias*	39	7–119	21	219	2840	58	222	528
	Spotted ratfish	*Hydrolagus colliei*	31	5–64	31	327	3154	93	206	371

$$\hat{N}_r = A_r \frac{\sum_h \frac{\hat{N}_h}{a_h}}{C_r} \quad (4)$$

This accounts for differing sampling density (number of hauls per area) among regions, and makes the assumption that each haul is an equal estimator of density, regardless of the area swept. Note that the total number of hauls in an area may include a small number in which no lengths were collected because of equipment malfunction, inclement weather precluding full sampling, or other unforeseen circumstances. Finally, the coast-wide estimated proportions by bin are calculated via

$$\hat{P}_b = \frac{\sum_r \hat{N}_r \cdot \hat{P}_{r,b}}{\sum_r \hat{N}_r} \quad (5)$$

Bootstrapped proportions at sex and length bin (B_b) are created by the following procedure:

(*i*) Within each geographic region, draw (with replacement) a random sample of hauls from the set of hauls contributing to the size composition estimate for that region.
(*ii*) Within each sampled haul, draw (with replacement) a random sample of lengths from individual observed lengths, keeping the original number of measurements.
(*iii*) Expand the bootstrapped "data" via eqs. 2, 3, and 5. Note that the scaling factors $\hat{N}_h$ and $\hat{N}_r$ remain unchanged. Therefore, this bootstrap can be interpreted as a simulation of new length observations from the same underlying survey sampling effort, not a resampling of the distribution of hauls among strata possible catches among strata.

For each bootstrap replicate, the realized sample size (R) is calculated via the equation used by McAllister and Ianelli (1997; referred to as effective sample size and consistent only with the use of a multinomial error assumption)

$$R = \frac{\sum_b (P_b \cdot (1 - P_b))}{\sum_b (P_b - B_b)^2} \quad (6)$$

The distribution of realized sample sizes was calculated via 25 000 replicates. This number was found to be adequate for the sample sizes and number of length bins frequently encountered in this data set by evaluating the sampling error among repeated bootstrap exercises for several species. Monte Carlo error for all estimators considered below was found to be <1% at this level of replication.

Simulation testing alternative estimators

To empirically validate the best estimator from a distribution of bootstrapped realized sample sizes, a simple simulation experiment was conducted using data for a typical rockfish species. A "true" length-frequency distribution was created, and random samples of varied sample size were drawn from this distribution. For each replicate, the realized sample size was calculated (via eq. 6), and the distribution of these calculated values was summarized by one of four commonly used estimators: the arithmetic mean, the geometric mean, the median, or the harmonic mean. To test for sensitivity, the exercise was repeated using more or fewer length bins and differing modal structure in the true data based on observed distributions from several representative species.

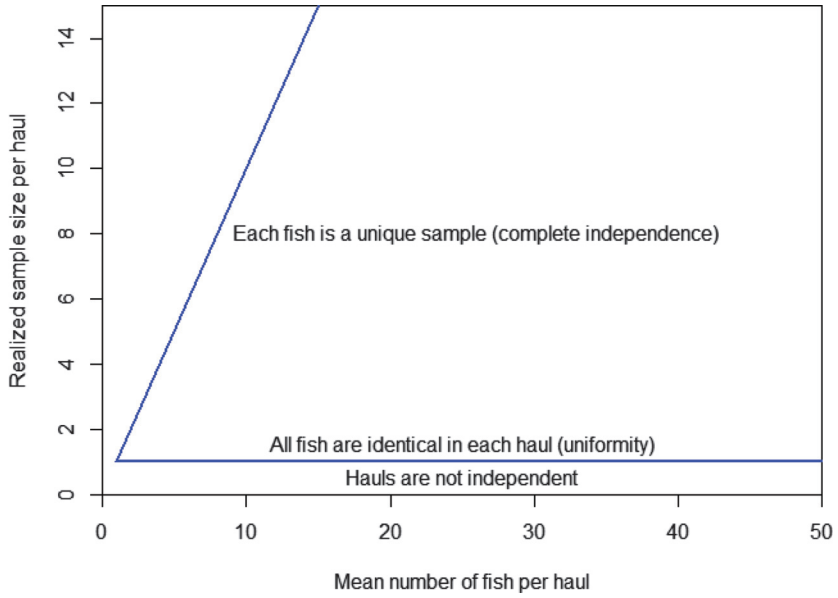
Summary among species

Annual length-frequency distributions and realized sample sizes based on each of the four estimators were calculated for 47 species for which an average of at least 75 length observations per year had been collected by the National Marine Fisheries Service (NMFS) west coast bottom trawl survey (Keller et al. 2007*a*, 2007*b*, 2008) during the period 2003–2010. Species were divided into five ecological-taxonomic groups: flatfish, shelf rockfish, slope rockfish, thornyheads, and "other species" (Tables 1 and 2). Length bins used in the analysis varied in width from 1 to 4 cm, matching the bin structure for that species (if an assessment

Table 2. Summary statistics for rockfish and thornyhead species.

			Length bins					Effective *N* per year		
Group	Species (area)	Scientific name	No.	Range (cm)	Mean hauls per stratum	Mean hauls per year	Mean fish per year	Min.	Mean	Max.
Shelf rockfish	Bank	*Sebastes rufus*	21	12–51	3	11	103	5	20	41
	Bocaccio	*Sebastes paucispinis*	27	25–76	5	35	228	40	59	107
	Calico	*Sebastes dalli*	22	6–27	5	8	74	7	23	59
	Canary	*Sebastes pinniger*	28	13–66	9	43	562	41	79	108
	Chilipepper	*Sebastes goodei*	19	17–52	13	80	2682	46	98	163
	Greenspotted	*Sebastes chlorostictus*	21	9–48	6	37	534	34	148	238
	Greenstriped	*Sebastes elongatus*	37	6–42	13	161	2702	139	344	454
	Halfbanded	*Sebastes semicinctus*	19	6–24	16	56	1506	18	53	74
	Pygmy	*Sebastes wilsoni*	16	9–24	3	11	204	6	40	105
	Redbanded	*Sebastes babcocki*	27	11–62	7	51	203	83	98	121
	Redstripe	*Sebastes proriger*	35	8–42	3	15	436	15	49	143
	Rosethorn	*Sebastes helvomaculatus*	31	7–37	6	52	1204	34	174	297
	Shortbelly	*Sebastes jordani*	25	4–28	9	53	1623	11	58	92
	Squarespot	*Sebastes hopkinsi*	21	9–29	5	10	263	4	37	72
	Stripetail	*Sebastes saxicola*	27	6–32	14	137	3382	155	240	337
	Widow	*Sebastes entomelas*	20	14–51	3	23	158	7	35	63
	Yellowtail	*Sebastes flavidus*	22	15–56	7	39	789	14	116	195
Slope rockfish	Aurora	*Sebastes aurora*	23	9–52	10	86	1580	146	375	546
	Blackgill	*Sebastes melanostomus*	25	13–60	7	33	433	39	83	107
	Darkblotched	*Sebastes crameri*	37	6–37	7	56	971	43	129	227
	Pacific ocean perch	*Sebastes alutus*	24	17–40	8	42	715	31	62	139
	Rougheye	*Sebastes aleutianus*	32	13–74	6	32	128	41	54	79
	Sharpchin	*Sebastes zacentrus*	34	6–39	4	38	1139	46	116	169
	Splitnose	*Sebastes diploproa*	28	5–42	15	137	3878	69	213	372
Thornyheads	Longspine thornyhead	*Sebastolobus altivelis*	31	5–35	23	231	8314	934	1911	3913
	Shortspine thornyhead	*Sebastolobus alascanus*	27	8–74	26	326	5891	699	1475	2812

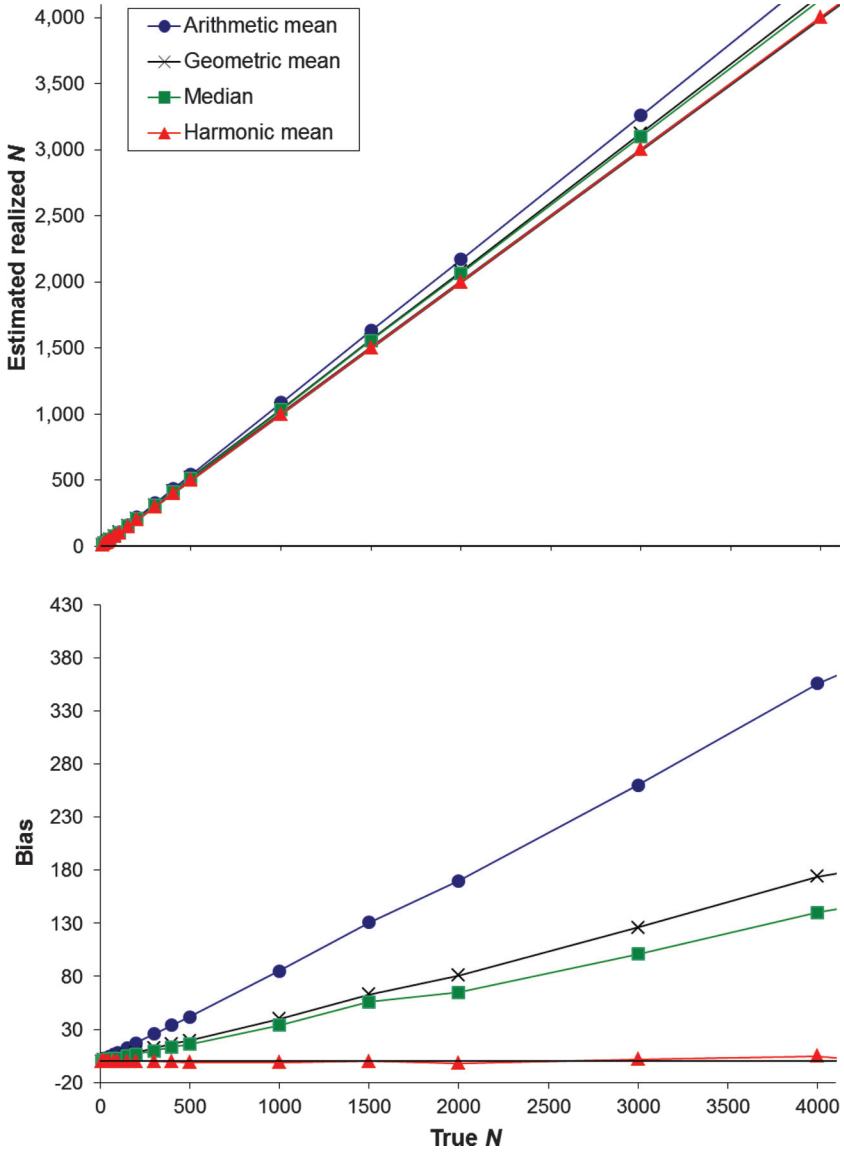
Fig. 1. Conceptual map of realized sample size from a sample as a function of the number of individual fish sampled per haul (or trip).



existed) or including 20–35 bins to span the observed size range. Geographic stratification (depth × latitude) also reflected the approach taken in the most current assessment, or relied on species presence in each of the five commonly used International Pacific Fisheries Commission areas (≤36, 36–40.5, 40.5–43, 43–47.5, and ≥47.5°N latitude) and three depth zones: shelf (55–183 m), shallow slope (183–549 m), and deep slope (549–1280 m). Geographical stratification for individual species ranged from 2 to 15 areas. Stock assessments are conducted separately for more than one geographic area for a few of these species; for those species, each assessed area was treated separately in this analysis.

Annual realized sample size estimates were compared on a per haul basis for clearer interpretation of the effect of the number of fish sampled. If each individual fish represented a purely random sample from the population, then the realized sample size per haul should increase linearly with the number of fish sampled. If each individual fish within each haul were identical, then the realized sample size should be constant and equal to 1 across all numbers of fish sampled within a haul. If, in addition, hauls within regions show correlation with each other, then the realized sample size could be less than one per haul (Fig. 1).

Fig. 2. Comparison of estimators of realized sample size for simulated data generated from a multinomial distribution based on the sample size and length-frequency distribution of a typical rockfish.



Results

Simulation testing of alternative estimators

The harmonic mean realized sample size was found to recover the true sample size with <1% bias (Fig. 2). Alternative estimators were found to be positively biased in all cases, with the bias increasing as a linear function of the true sample size for each set of simulations. This result merely corroborates previous analytical work by G. Thompson (NMFS, Alaska Fisheries Science Center, Seattle, WA; personal communication). He showed that if the true distribution is multinomial, then the harmonic mean of the distribution of calculated realized sample sizes calculated from repeated sampling will equal the true sample size. For subsequent analysis of trawl survey data, all results are represented by the annual harmonic mean realized sample size over bootstrapped replicates.

Species summaries

The average annual realized sample size was reasonably consistent among taxonomic groups (Fig. 3). All species combined had an average annual effective sample size per survey haul of 2.73, slightly lower for the shelf and slope rockfish species (2.43) and slightly higher for flatfish (3.09). The thornyheads had the highest average realized sample size per haul (6.91), and also the greatest independence among fish within hauls, showing a nearly linear increase in realized sample size per haul as the number of fish sampled per haul increased (Fig. 4).

For all 47 species analyzed (Tables 1, 2), trawl surveys always produced an annual harmonic mean realized sample size of less than 25 per haul sampled, and for most species-years (95%) this number was below 6 (99% below 14). This strongly corroborates the well-published result that individual fish sampled within a particular haul are highly nonindependent samples from the sampled population. Despite frequent annual sampling in excess of 20 fish per haul, neither the flatfish nor the shelf rockfish species showed appreciable improvement in realized sample size as the number of fish was increased. The realized sample sizes calculated for each species varied broadly among years: by a factor of 2–20 depending on the species.

Fig. 3. Distribution of observed realized sample size per survey haul among species groups for all years. Bars and whiskers represent the mean ± 1 SD.

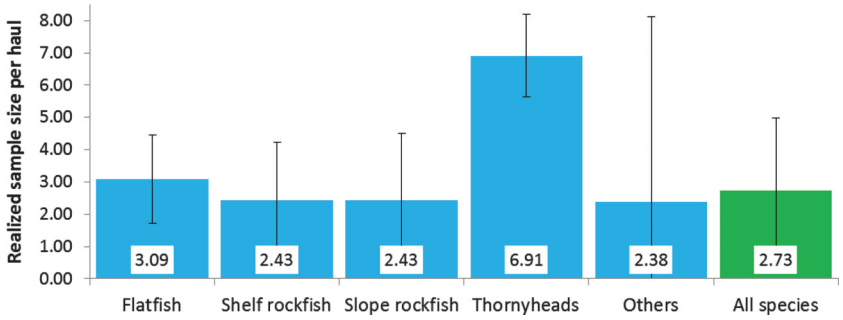
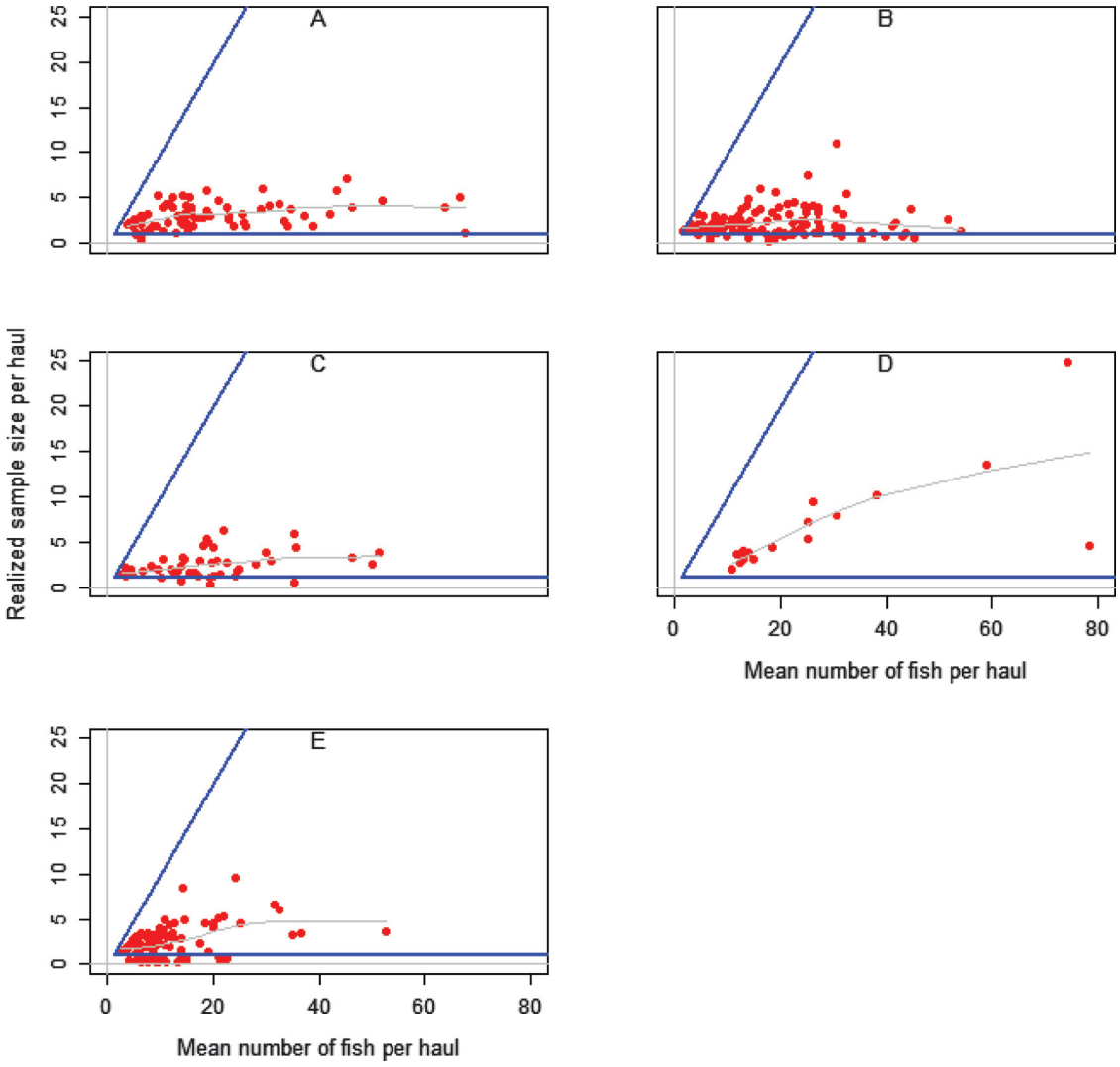


Fig. 4. Relationship between estimated realized sample size and the average number of fish sampled per haul by year for flatfish (panel A), shelf rockfish (panel B), slope rockfish (panel C), thornyheads (shortspine and longspine; panel D), and all others (panel E). Dark lines represent realized sample sizes of 1 per haul (horizontal) and 1 per fish (diagonal); light lines indicate Loess smoothing to aid in visualizing the central tendency of the data.



Discussion

This study generally supports the observation that fisheries data sets containing many sampling units (hauls or trips) are likely to be far more informative than those containing even vast numbers of individual fish sampled from a small number of hauls. Research programs would be well served to remember this, despite the forgone satisfaction of seeing massive numbers of lengths or ages collected. Ideally, bootstrap methods based on sampling theory should be routinely applied to data sets to provide maximum input sample size estimates, derived externally to the stock assessment. These values will inherently incorporate interannual variability in sampling and population-level pro-

cesses. Because of technical requirements, this may not always be possible. Bootstrapping each year for a data set to be used for stock assessment may be technically demanding, but it has been done in some cases (e.g., Hirst et al. 2005; Hulson et al. 2012). For similar cases where a bootstrap is not available, the use of 2–4 times the number of hauls (Fig. 3) may generate approximate starting values that can then be tuned or updated during the assessment process. In some cases, the sample sizes generated via this type of bootstrap procedure may be far too large to use directly in a stock assessment analysis, just as minimum variance estimators for other types of data may be far too small.

This analysis examines a fairly typical data set and provides a comparison among a large number of species to reveal general patterns that may be useful for species currently unanalyzed and for future years' observations. However, strong cohorts, environmental conditions, etc., may change these relationships, even for a particular species over time.

The two thornyhead species are somewhat unique among west coast US groundfish species in that they are nonschooling, their juveniles recruit to the same areas in which the adults live, and both species generally show little geographic structure in biological characteristics. The bootstrapping results for these species suggest a much stronger relationship between the number of individuals measured per haul and increased realized sample sizes than other species evaluated. Flatfish showed little evidence that any additional improvement in realized sample size could be obtained by increasing actual sample sizes above about 20 fish per haul. This is similar to the results obtained by Aanes and Pennington (2003). Slope rockfish, shelf rockfish, and the other species category also showed no clear improvement above a modest number of individual lengths. The breadth of variability within species among years provides a strong confirmation that assumptions of constant sample sizes for all years are not realistic, even for highly standardized surveys. This is likely due to changes in the modal structure of the underlying population, the intended sample size for each species, as well as population clustering. In addition, the best error assumption for composition data may vary among species and may not be well represented by the multinomial distribution.

Although the multinomial error distribution is frequently employed for compositional data in stock assessments, there are better alternatives (e.g., Hrafnkelsson and Stefánsson 2004), and a substantial amount of research illustrating that observed variance structures are frequently not well approximated by the multinomial. Dispersion has been found to be greater than even the overdispersed multinomial of McAllister and Ianelli (1997). The variance-covariance structure is likely poorly approximated by the multinomial (Miller and Skalski 2006), and bootstrapping can be used to estimate length-specific variances directly (e.g., Hirst et al. 2005) if these are more applicable for stock assessment. The general approach of bootstrapping the raw data, and summarizing the annual maximum realized sample size based on sampling variability, should be applied to each data set, and adjusted to be consistent with the error distribution to be used in the subsequent stock assessment. For this reason, the specific results presented here are relevant only to analyses utilizing the multinomial. Future simulations should investigate whether these results would differ for applications using alternative error distributions.

One approach to using this type of approach for an actual stock assessment would be to start with initial values from a bootstrap, and then adjust these values down as needed for an entire data source, while retaining the relative interannual variability. This iterative tuning is frequently performed using a single multiplicative scalar for each data source included in the assessment. Where this approach is used, one useful guide may be the harmonic mean of effective sample sizes among years for that data source observed in the assessment (calculated based on model fit). This approach is consistent with the simulation results presented here, showing that the use of the mean of a collection of realized sample sizes is always biased high, but the harmonic mean is unbiased. There is an important (and currently untested) logical assumption made here: that summary of observations among years (strictly different multinomial distributions) can be likened to summary observations from a single multinomial. This should be explored in future studies. Regardless of the method ultimately employed in stock assessments, bootstrapped realized sample sizes represent maximum values that should not be exceeded in stock assessment assumptions. Reductions may be necessary because of a multitude of factors including representativeness of survey sampling, methodological changes, and other sources of error (specifically, a lack of explicit processes in the assessment model which generate systematic differences between the underlying population and the observed samples) in the assessment model, incorrect error distribution assumptions, and many others. The potential costs of overstating the sample sizes for compositional data in stock assessments are an overstatement of certainty and biased results (Francis 2011).

This analysis does not suggest what sample sizes to use for stock assessment per se, nor does it compare alternate error distributions from which to derive them. The approach taken here does provide a general sample size calculation with two important properties: (*i*) interannual heterogeneity in sampling, length modality, and population clustering are accounted for; and (*ii*) based purely on the sampling properties, input sample sizes used in stock assessments should not exceed the bootstrapped values. External definition of maximum annual sample sizes (analogous to a minimum variance estimate) can provide an important limit for subsequent iterative- and estimation-based approaches to tuning of these values in stock assessments. Without such a priori identification of sampling theory based maximum values, a stock assessment analyst may easily be lulled by the thousands of fish measured or aged into placing too much emphasis on compositional data. This may be a serious source of bias if it precludes satisfactory fitting to primary indices of stock abundance (Francis 2011). It can also provide a basis for refining sampling targets and designing experiments to increase understanding of the processes generating compositional samples and their influence on stock assessment results.

Acknowledgements

We thank the many scientists working aboard the NWFSC bottom trawl survey who have collected over a million individual fish lengths used in this analysis. Grant Thompson provided very helpful discussion of analytic methods and conceptual issues regarding treatment of effective sample size in stock assessments, as well as comments on an early version of the manuscript. Stacey Miller assisted with several summaries of stock assessment results and methods being employed that were integral in refining the goals of this work. John Wallace provided editorial comments on the manuscript. This research was largely completed while the lead author was employed by the NMFS Northwest Fisheries Science Center and benefitted greatly from support by and numerous suggestions from many of the stock assessment staff. Comments from two anonymous reviewers substantially improved this paper.

References

Aanes, S., and Pennington, M. 2003. On estimating the age composition of the commercial catch of Northeast Arctic cod from a sample of clusters. ICES J. Mar. Sci. **60**: 297–303. doi:10.1016/S1054-3139(03)00008-0.

Cerviño, S., and Saborido-Rey, F. 2006. Using the bootstrap to investigate the effects of varying tow lengths and catch sampling schemes in fish survey. Fish. Res. **79**: 294–302. doi:10.1016/j.fishres.2006.03.021.

Chih, C.-P. 2010. Incorporating effective sample sizes into sampling designs for reef fish. Fish. Res. **105**: 102–110. doi:10.1016/j.fishres.2010.03.008.

Crone, P.R. 1995. Sampling design and statistical considerations for the commer-

cial groundfish fishery of Oregon. Can. J. Fish. Aquat. Sci. **52**(4): 716–732. doi:10.1139/f95-072.

Deriso, R.B., Maunder, M.N., and Skalski, J.R. 2007. Variance estimation in integrated assessment models and its importance for hypothesis testing. Can. J. Fish. Aquat. Sci. **64**(2): 187–197. doi:10.1139/f06-178.

Fournier, D., and Archibald, C.P. 1982. A general theory for analyzing catch at age data. Can. J. Fish. Aquat. Sci. **39**(8): 1195–1207. doi:10.1139/f82-157.

Fournier, D.A., Sibert, J.R., Majkowski, J., and Hampton, J. 1990. MULTIFAN a likelihood-based method for estimating growth parameters and age composition from multiple length frequency data sets illustrated using data for southern bluefin tuna (*Thunnus maccoyii*). Can. J. Fish. Aquat. Sci. **47**(2): 301–317. doi:10.1139/f90-032.

Fournier, D.A., Hampton, J., and Sibert, J.R. 1998. MULTIFAN-CL: a length-based, age-structured model for fisheries stock assessment, with application to South Pacific albacore, *Thunnus alalunga*. Can. J. Fish. Aquat. Sci. **55**(9): 2105–2116. doi:10.1139/f98-100.

Francis, R.I.C.C. 2011. Data weighting in statistical fisheries stock assessment models. Can. J. Fish. Aquat. Sci. **68**(6): 1124–1138. doi:10.1139/f2011-025.

Gerritsen, H.D., and McGrath, D. 2007. Precision estimates and suggested sample sizes for length-frequency data. Fish. Bull. **106**: 116–120.

Helle, K., and Pennington, M. 2004. Survey design considerations for estimating the length composition of the commercial catch of some deep-water species in the northeast Atlantic. Fish. Res. **70**: 55–60. doi:10.1016/j.fishres.2004.06.011.

Hilborn, R., and Walters, C.J. 1992. Quantitative fisheries stock assessment: choice, dynamics and uncertainty. Chapman and Hall, London. 570 p.

Hirst, D., Storvik, G., Aldrin, M., Aanes, S., and Huseby, R.B. 2005. Estimating catch-at-age by combining data from different sources. Can. J. Fish. Aquat. Sci. **62**(6): 1377–1385. doi:10.1139/f05-026.

Hrafnkelsson, B., and Stefánsson, G. 2004. A model for categorical length data from groundfish surveys. Can. J. Fish. Aquat. Sci. **61**(7): 1135–1142. doi:10.1139/f04-049.

Hulson, P.-J.F., Hanselman, D.H., and Quinn, T.J. 2012. Determining effective sample size in integrated age-structured assessment models. ICES J. Mar. Sci. **69**: 281–292. doi:10.1093/icesjms/fsr189.

Keller, A.A., Horness, B.H., Simon, V.H., Tuttle, V.J., Wallace, J.R., Fruh, E.L., Bosley, K.L., Kamikawa, D.J., and Buchanan, J.C. 2007*a*. The 2004 U.S. West Coast bottom trawl survey of groundfish resources off Washington, Oregon, and California: estimates of distribution, abundance, and length composition. U.S. Dept. Commer., NOAA Tech. Memo. NMFS-NWFSC-87. 134 p.

Keller, A.A., Simon, V.H., Horness, B.H., Wallace, J.R., Tuttle, V.J., Fruh, E.L., Bosley, K.L., Kamikawa, D.J., and Buchanan, J.C. 2007*b*. The 2003 U.S. West Coast bottom trawl survey of groundfish resources off Washington, Oregon, and California: estimates of distribution, abundance, and length composition. U.S. Dept. Commer., NOAA Tech. Memo. NMFS-NWFSC-86. 130 p.

Keller, A.A., Horness, B.H., Fruh, E.L., Simon, V.H., Tuttle, V.J., Bosley, K.L., Buchanan, J.C., Kamikawa, D.J., and Wallace, J.R. 2008. The 2005 U.S. West Coast bottom trawl survey of groundfish resources off Washington, Oregon, and California: estimates of distribution, abundance, and length composition. U.S. Dept. Commer., NOAA Tech. Memo. NMFS-NWFSC-93. 136 p.

Maunder, M.N. 2011. Review and evaluation of likelihood functions for composition data in stock-assessment models: estimating the effective sample size. Fish. Res. **109**: 311–319. doi:10.1016/j.fishres.2011.02.018.

Maunder, M.N., and Punt, A.E. 2013. A review of integrated analysis in fisheries stock assessment. Fish. Res. **142**: 61–74. doi:10.1016/j.fishres.2012.07.025.

McAllister, M.K., and Ianelli, J.N. 1997. Bayesian stock assessment using catch-age data and the sampling - importance resampling algorithm. Can. J. Fish. Aquat. Sci. **54**(2): 284–300. doi:10.1139/f96-285.

Megrey, B.M. 1989. Review and comparison of age-structured stock assessment models from theoretical and applied points of view. Am. Fish. Soc. Symp. **6**: 8–48.

Methot, R.D. 1989. Synthetic estimates of historical abundance and mortality for northern anchovy. Am. Fish. Soc. Symp. **6**: 66–82.

Methot, R.D. 2000. Technical description of the Stock Synthesis assessment program. U.S. Dept. Commer., NOAA Tech. Memo. NWFD-NWFSC-43. 46 p.

Methot, R.D., and Wetzel, C.R. 2013. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fish. Res. **142**: 86–99. doi:10.1016/j.fishres.2012.10.012.

Miller, T.J., and Skalski, J.R. 2006. Integrating design- and model-based inference to estimate length and age composition in North Pacific longline catches. Can. J. Fish. Aquat. Sci. **63**(5): 1092–1114. doi:10.1139/f06-022.

Patterson, K., Cook, R., Darby, C., Gavaris, S., Kell, L., Lewy, P., Mesnil, B., Punt, A.E., Restrepo, V., Skagen, D.W., and Stefansson, G. 2001. Estimating uncertainty in fish stock assessment and forecasting. Fish Fish. **2**: 125–157. doi:10.1046/j.1467-2960.2001.00042.x.

Pennington, M., and Volstad, J.H. 1994. Assessing the effect of intra-haul correlation and variable density on estimates of population characteristics from marine surveys. Biometrics, **50**: 725–732. doi:10.2307/2532786.

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

Quinn, T.J.I. 2003. Ruminations on the development and future of population dynamics models in fisheries. Nat. Res. Modell. **16**: 341–392. doi:10.1111/j.1939-7445.2003.tb00119.x.

Quinn, T.J.I., and Deriso, R.B. 1999. Quantitative fish dynamics. Oxford University Press, New York. 542 p.

Vokoun, J.C., Rabeni, C.F., and Stanovick, J.S. 2001. Sample-size requirements for evaluating population size structure. N. Am. J. Fish. Manage. **21**: 660–665. doi:10.1577/1548-8675(2001)021<0660:SSRFEP>2.0.CO;2.

Worthington, D.G., Fowler, A.J., and Doherty, P.J. 1995. Determining the most efficient method of age determination for estimating the age structure of a fish population. Can. J. Fish. Aquat. Sci. **52**(11): 2320–2326. doi:10.1139/f95-224.