Center for Independent Experts (CIE) independent peer review of "Shore-based and boat-based fishing surveys in Guam, CNMI, and American Samoa: Survey design, expansion algorithm, and a case study"

Dr. John M. Hoenig

November 2021

Executive Summary

Preface. The main report we were given to review did not explain the sampling designs and procedures in enough detail to enable us (the reviewers) to know what is done in the surveys. Consequently, the review team requested and participated in a series of online meetings with NMFS staff scientists to clarify the six designs in use. It is to the NMFS scientists' credit that they were able to provide a great deal of supplemental information in a short period of time. The technical report describing the sampling procedures should be revised to make the design, randomization, field sampling procedures, and data analysis clear. There are enough differences among the surveys that each one should be described separately. However, I believe I have enough information now to proceed with my review of the sampling programs.

Design. The overall sampling design looks reasonable. Effort was made to spread out the data collection over space and time to get representative results, and in most cases the sampling effort was extensive which helps in reducing bias and variance. However, the designs are not as claimed in the technical report, meaning that the designs were not simply stratified random sampling. Rather, the designs in most cases are (closer to) two stage cluster sampling, for example selecting days (clusters) at the first stage and then selecting port or shift (time of day) within a day (within a cluster). The designs approximate even more closely a lattice design. This probably does not cause appreciable bias in the catch estimates but the estimated variances are probably too optimistic (too low). It is not clear in the report how the sample selection procedures (i.e., randomizations) are implemented but in discussions with NMFS scientists we found that the randomizations are often complicated because of logistical and other constraints imposed on the randomization. In some cases, no randomization was done. This makes the properties of the survey design uncertain.

The design of each of the six surveys (3 island groups x 2 types of survey, shore and boat-based) should be reviewed with the aim of relating the design to standard sampling designs. As indicated above, I believe that in most cases the surveys are close to, or can be made to be close to, two-stage cluster sampling or lattice sampling. The appropriate formulae for computing estimates of catch and for estimating variance can then be specified.

Estimates are made for subsets of the fishery called domains. Examples of characteristics used to define domains are type of fishing gear used during the trip, and whether the trip is a charter boat trip or a private trip. This amounts to using post-stratification. There are two reasons to compute estimates for subsets of the fishery: 1) the domains are of intrinsic interest, and 2) for increased efficiency (precision). Sometimes, estimates are needed for subsets of a fishery, e.g., when quota needs to be allocated by port or by mode of fishing. This does not appear to be the case here. By defining domains, one can increase or decrease the survey precision depending on whether the means differ greatly among domains. Defining too many domains can cause problems when sample sizes for the domains approach zero. The solution adopted by the investigators is to pool data over similar domains which defeats the purpose of defining domains. I recommend that the investigators reconsider which strata and domains are important for the assessment and management of the fisheries to see if the sampling designs can be simplified and made to follow standard sampling designs more closely. This can be evaluated by estimating totals and variances with domains collapsed (combined) in various ways.

Expansion. The expansion procedures for estimating catch rate, effort, and catch are appropriate for the most part but not for estimating variance. This is because the estimates are based on stratified random sampling when the actual procedure is closer to cluster sampling or lattice sampling. Assuming stratified random sampling (meaning, assuming that simple random samples are taken within a stratum), when the actual procedure consists of sampling or subsampling clusters of fishing activity, tends to underestimate the variance (because within a cluster the observations tend to be less variable than in the total population as a whole).

Another issue with the estimation of variances is that certain adjustment factors are treated as known constants when they are actually assigned an assumed value or estimated. The uncertainty in these adjustment factors is not taken into consideration.

The variance formulae do not incorporate finite population corrections (fpc). Intuitively, if half of all fishing activity is observed, then the uncertainty in catch pertains only to the half that is not observed since the catch is known exactly for half the fishery. The finite population correction reduces the variance of the estimates according to the fraction of the fishery that is observed. The use of finite population corrections should be investigated when the sampling fractions are high.

The choice of estimator for catch rate (cpue) in the roving surveys might be improved. The current procedure is to use the ratio of means (sum of catch / sum of effort). The alternative is to use the mean of ratios (sum of all catch/effort ratios divided by number of interviews). Hoenig et al. (1997) showed the mean of ratios estimator is appropriate provided the fishing efforts are not short; they recommended that, in practice, the mean of the individual catch rates should be calculated after discarding any interview where the fishing effort was less than some small amount, say 30 minutes. The two estimators lead to similar results in expectation so I am not concerned about bias. However, the procedure recommended by Hoenig et al. had lower mean squared error than the ratio of means. The expansion procedure for the access point surveys is appropriate. In these surveys, catch rate is always expressed “per trip” so the mean of ratios and ratio of means estimators are the same and there is no computational issue.

Pooling data across sampling domains is used when sample sizes are very small (less than 3). The procedures used are reasonable. But the need for pooling would be reduced if fewer domains were defined. Some thought should be given to doing this.

Software. As indicated above, I believe that in most cases the sampling design has been mis-specified by the investigators, including for Guam. This should not cause appreciable bias in the estimates of catch rate, effort and catch. But, the estimates of variance are not appropriate in my opinion. It did not seem necessary to review the code in detail given that the code will have to be rewritten. I therefore offer more general suggestions about the computer code.

The code (in three files) was written to handle all of the surveys. But, because surveys vary in their design, some thought should be given to writing separate programs for each survey design. (Some functions will be usable by all programs.)

Programs could include additional features. It may be helpful to compute estimates for various definitions of sampling domains (i.e., to allow for one to explore the effects of collapsing sampling domains into a smaller number of domains). Also, it would be helpful to include sensitivity analysis of the effect of changing the value of various assumed parameters

Example. The example makes the estimation procedure clear and I have no criticisms of this section of the report.

The example illustrates that the defining of many sampling domains causes problems with inadequate sample sizes. This necessitates pooling of samples and weakens the inferences about domains. The example supports the suggestion that the collapsing (combining) of sampling domains should be considered.

Sensitivity analyses. The sensitivity analyses focused on two aspects of the expansion algorithm: choice of pooling algorithm and selection of representative ports. These analyses are good, as far as they go.

With respect to pooling algorithms, I wish the authors had additionally considered collapsing (combining) domains as another option. If one domain has 2 observations and another has one observation, then pooling results in 3 observations for each domain. Collapsing the two domains would result in 3 observations for the collapsed (combined) domains. If one domain has 2 observations and another has 5, then their pooling results in 7 observations for one domain (2 + 5 = 7) while the other domain has 5 observations (since 5 is greater than the minimum requirement of 3). Collapsing the two domains would result in 7 observations for the collapsed stratum. In the calculation of variances, it is not clear what sample size is specified when pooling is used. This needs to be specified and evaluated.

With respect to representative ports, I think the sensitivity analyses are clear, appropriate and appropriately interpreted.

I wish the authors had conducted additional sensitivity analyses for those quantities are assumed or specified based on expert opinion (e.g., p_1, p_2, reference table consisting of method-specific CPUE values).

Best available science. In my opinion, the best available science is being used based on the fact that the overall study design and the implementation are reasonable. However, this does not mean that the sampling can't be improved. I believe each of the survey designs should be reconsidered in light of my comments about cluster sampling versus stratified random sampling. Modifications should be made to design and procedures so that the sampling protocols more closely match a known design; then the formulae for computing point estimates and variances should be revised (especially the latter). Randomization should be introduced in a few places where it is lacking and in a few instances randomization might be made subject to fewer constraints. I consider this fine tuning basically acceptable procedures, that is fine tuning best available science, rather than discarding existing approaches in favor of alternative approaches.

Background

Here, I provide some background on sampling that is at the heart of my evaluation of the sampling designs. I also review the concept of instantaneous counts for estimating fishing effort in support of my conclusions reached below.

The technical report I reviewed states that stratified random sampling is used in (almost) all of the surveys. As will be discussed below, I believe this is a mischaracterization of the sampling design. For this reason, I describe two common sampling designs here.

In sampling theory, a "design" consists of defining the sampling unit and the variable(s) to be recorded for each sampling unit observed, constructing a "frame" or comprehensive list of all sampling units in the population, and specifying a procedure for selecting sampling units to be observed with known probabilities. The sampling unit might be a location x block of time combination, e.g., Tuesday noon to 6 pm in port B, or Tuesday noon to 6 pm along Route 2. The sampling frame is all locations x days (or shifts) combinations during a larger period of time such as a month or quarter year during which the survey is conducted.

Stratified random sampling consists of dividing the population being surveyed (i.e., the sampling frame) into non-overlapping subsets – called strata - which together encompass the entire population. The strata must be sampled independently which means that what is sampled in one stratum has no bearing on what is sampled in another stratum. Within each stratum, a simple random sample is taken. As an example, suppose we have one survey agent and suppose there are four ports which are to be sampled during a month. The sampling unit is the day x port combination, i.e., there are 30 x 4 = 120 sampling units in the sampling frame. We decide to stratify by day type – weekdays and weekend days such that there are 22 weekdays x 4 ports = 88 sampling units in the weekday stratum and 8 x 4 = 32 sampling units in the weekend stratum. If we decide to sample with high intensity, it may happen that our randomization procedure has the survey agent working 6 or more days in a row (e.g., Tuesday through Sunday). If we impose a constraint that the survey agent can't work more than 5 days in a row then we no longer have stratified random sampling. The reason is that what happens in one stratum affects what can be sampled in the other stratum, thus violating the condition that the strata are sampled independently. As another example, it may happen that the random selection of port x day combinations results in two (or more) ports being sampled in one day. The survey agent cannot be in two places at one time. If we impose the constraint that only one port can be sampled in a day then we no longer have stratified random sampling because we don't have a simple random sample of the sampling units.

In the last example, the imposition of the constraint that only one port can be sampled in a day results in a two stage process. One picks days to sample and, within days, one picks a single port to sample. This is known as two-stage cluster sampling. A day is thought of as the primary sampling unit or cluster, and within the day the four ports constitute secondary sampling units. Then, the variability of the results, i.e., the variance, depends on the variability among days, and the variability within days (among ports). These two components of variance can be estimated and combined to estimate the overall variance of the estimated catch if at least some days have two or more ports sampled. If only one port is sampled at the second stage (i.e., within a day) then it is not possible to estimate the within-cluster variability but it is still possible to obtain a conservative (likely too high) estimate of the variance of the estimated catch (see Cochran 1977, p. 279).

The shore-based surveys evaluated in this report are based on roving survey designs with what Hoenig et al. (1993) call progressive counts. The theory for this kind of survey is unusual in that fishing effort is determined by observing each location along a specified route of interest for just an instant. The concept is simplest for the case of an instantaneous count where the entire route is viewed simultaneously from a suitable vantage point such as an airplane. Then, a count made from, say, a photograph of the entire fishing grounds at a random point in a time interval (say, a shift or a day) is an unbiased estimate of the average number of fishers fishing during the time interval, based on a sample of one instant. The total amount of fishing effort during the time interval is then estimated unbiasedly as the product of the average number of fishers times the length of the time interval. Since only one count was made, it is not possible to estimate the variance for the average number of people fishing (or for the total amount of fishing effort) for that time interval. However, if this procedure is repeated for several time periods (shifts or days) one can estimate a variance from the variability of the counts among time periods. It is not possible to estimate the components of variance – within time period and among time period variability – unless one makes at least two counts (at random times) in some of the time periods.

Often, one cannot make an instantaneous count of all locations at once. Then one can make a progressive count. Consider first a circular route. One picks a random place along the route to begin the count at the beginning of the time interval and the survey agent travels the route at constant speed to end up at the starting place at the end of the time interval; the agent counts fishers while travelling, viewing each location for just an instant. Each location has equal probability of being counted with respect to time in the time interval. Thus, the count for each location provides an unbiased estimate of the average number of fishers at that location during the time interval. Furthermore, the sum of the counts (over all points on the route) provides an unbiased estimate of the average number of people fishing on the route over the entire time interval. Now consider a linear route between Point A and Point B. One starts the count at the beginning of the time interval at a randomly selected point C between A and B. One then chooses the direction of travel randomly (either towards A or B). Say the direction of travel is towards Point B. Then one travels from C to B counting fishers along the way, then returns to Point A (not counting anglers), and resumes counting as one proceeds from Point A to Point C.

Description of Reviewer's Role in Review Activities

I reviewed the documents provided, submitted queries to the responsible project leaders and received additional documentation (see Appendix 1), and participated in several online meetings to discuss procedures for the review and to discuss aspects of the sampling program.

<u>Online meetings, all with NMFS, CIE, and external reviewers</u>

23 August, Introductory meeting
7 September, 1st meeting to discuss shore-based sampling

9 September, 2nd meeting to discuss shore-based sampling
16 September, 3rd meeting to discuss shore-based sampling
22 September, 4th meeting to discuss shore-based sampling

Summary of Findings (weaknesses and strengths)

1. Evaluate the shore-based and boat-based fishing survey designs and determine if the survey designs are appropriate.

There are enough differences among the six surveys that review of each survey separately is warranted. Because these surveys are complex, I summarize the sampling design for each survey, as I understand it, based on the scheduling procedures. If the Agency disagrees with, or has questions about, my conclusions, it will want to verify that my comments are based on a sound understanding of the procedures in use. My evaluation of what sampling designs are being used differs from the type of sampling stated in the technical report (Table 1).

Following my description and evaluation of the six surveys, I discuss some general design issues including randomization in roving designs, and how many domains should be defined.

Shore-based design, Guam.

The survey design and sample selection procedure is as follows.

A survey lasts one month. Estimates are computed separately for weekends (defined to be Saturdays, Sundays and holidays) and weekdays. There are three regions and two shifts during a day. Both shifts are sampled on every day selected for sampling but estimates are computed by time of day. (The technical document led me to believe that time of day was not an expansion domain but the notes on shore-based sampling distributed by T. Matthews on 20 September clarified that time of day is an expansion domain.)

The fishing day is defined to be from 6:00 am to 2 am. The dividing line between early and late shifts is 6 pm.

For estimating effort, all regions and both shifts are sampled on each day selected to be sampled. Two weekdays are sampled per month and 2 weekend days. Effort is estimated by roving, starting at 6:30 am and finishing 4 to 5 hours later (around 10:30 to 11:30) – whenever the agent finishes travelling the route. This observed effort is assumed to be representative of the period from 6:00 am to 6 pm. Effort for the late shift is determined by roving, starting at 7 pm and ending 4 to 5 hours later. This effort is representative of the effort between 6 pm and 2 am. All 3 regions are covered in each shift.

Catch rate is determined by a separate procedure from effort. Region 1 is sampled at 1 day per month, with the daytype alternating by month. (If in last month's survey, Region 1 was sampled on a weekday, in this month's survey it will be sampled on a weekend.) Region 2 is also sampled at 1 day per month with the daytype being the opposite of that for Region 1. There is an additional constraint that the days selected for Regions 1 and 2 cannot be days already selected for effort sampling. (There can only be one type of survey in any given day because of logistical constraints.) Region 3 is sampled at 2 d/mo. One day is a weekday and one is a weekend. The days are selected from those days not already selected for catch rate or effort sampling.

My evaluation of the sampling design is as follows.

Let us ignore for the moment the fact that the sampling for effort does not cover the entire day sampled but, rather, has a fixed starting time and a variable ending time. Then the design for effort on weekdays resembles stratification by day type and simple random sampling of days within strata. That is, because both of the two shifts are sampled on any day selected for sampling, the sampling approximates a simple random sample of days. The effort and catch rate surveys are done on different days so, strictly speaking, they are not independent and this complicates the theoretical treatment of the sampling design. However, given the low amount of sampling per month for weekdays (2 days for catch rate, 2 for effort) it does not appear that the requirement, that there can only be one type of survey per day, would have much effect on scheduling the sampling.

Because the survey agent cannot be in two places at once, the entire fishery is not viewed all day long and effort is not known perfectly for any effort-sampling day, the use of a finite population correction for effort estimation would not be appropriate here.

The design for effort on weekends is more complicated. A large proportion of the weekend days are sampled (there are approximately 10 weekend/holiday days per month, of which 2 are sampled for effort and 2 for catch rate). This causes the effort survey for weekends to depart from a simple random sample of days but this may not be an important consideration.

If late shifts have different efforts and catch rates than early shifts then it may be efficient to compute estimates separately for each shift and then add the early and late shift estimates to get the total. However, the two estimates are not independent. Therefore, the covariance between the early and late shift estimates needs to be considered. The variance of the total is

Var[early + late shift estimates] = Var[early shift] + Var[late shift] + 2 Cov[early, late] .

The covariance is easily estimated.

A problem with this design that should be corrected is that the effort sampling is not conducted with uniform intensity throughout the shift. This is due to two aspects. First, the starting time is always fixed at 6 am and 7 pm. There is no chance of sampling from 6 am to 6:30 am, from approximately 11:30 am until 7:00 pm, and from approximately midnight to 2 am. Second, the time is takes to cover the sampling route is approximately four to five hours. I recommend that the time spent travelling the sampling route be fixed at five hours and the survey agents be required to stick to a schedule that causes them to finish five hours after they begin. This is most easily accomplished by setting a series of mileposts or checkpoints. The survey agents can then modify their rate of travel to ensure they arrive at each milepost at the required time. The starting time of the route can then be randomized in order to have uniform probability of being in the field at all times of the shift (this can be a little tricky, - see Hoenig et al. (1993) for a procedure for scheduling progressive counts). It appeared from the technical report that the starting location along the route is not randomized. However, in the notes on shore based surveys distributed by Matthews on 20 September it is stated that the starting location is in fact randomized.

The catch rate survey is complicated and I am not sure to what standard design the sampling might resemble. According to the notes on shore based sampling distributed by T. Matthews on 20 September, “Catch rate surveys are stratified by type of day and region” and “An expansion domain is specified by the: Type of day (weekday or weekend), Time of day …, Region, but only for the hook and line fishing method". But, with only one observation of region 1 per month it is not possible to obtain separate estimates for this region for weekdays and weekend days. Similarly, for region 2: one can't estimate two catch rates from one observation. For region 3, there is one observation on a weekday and one observation on a weekend day so one can obtain an estimate of weekday catchrate and weekend catchrate but one cannot estimate the variances without further information, e.g., by pooling data over months. (For example, the method of collapsed strata could be used to pool data from adjacent months.)

I question why one would want to treat region as a domain since it does not appear that region-specific estimates of catch are needed.

I think the design for shore-based catch rate sampling in Guam should be given more thought and simplified if possible.

Shore-based design, American Samoa.

The survey design and sample selection procedure is as follows.

A survey takes place over a period of one month, so there are 12 surveys conducted per year. Estimates are computed separately for weekends (defined to be Saturdays and holidays) and weekdays. Sample selection (i.e., sampling assignments for the survey agents) is done first for the weekends and then, contingent on the weekend selections, for the weekdays. Thus, weekends and weekdays are not sampled independently since the choices for weekends affects what is selected for weekdays.

For both weekends and weekdays, the days are divided into three shifts and three regions. Thus, the sampling unit is the day x shift x region combination, e.g., "2nd Saturday of the month, shift 3, region 1" is a sampling unit. There are effectively two survey agents available to be assigned to sampling units. (Technically, there are two survey teams with each team consisting of two people. However, the two members of the team function as one sampler, i.e., they can't split up and go to separate regions or sample at different times.) Each survey agent works two weekends in the month. Four combinations of day x shift x region are selected for a month without replacement. A further restriction is that you can't have two teams in the same region on the same day at the same time; this causes an additional restriction because shift two overlaps shifts one and three in time. So, on a given day in a given region, one can have the two teams sampling shifts one and three but not shifts one and two or shifts two and three. Any day might have 0, 1 or 2 teams sampling (this is true of weekdays as well as weekend days). After scheduling the sampling for weekends in a month, the schedule for week days is determined separately for each week of the month. Each survey agent works six days a week. This means the agent works 5 weekdays (if the agent does not work on Saturday in that week) or works 4 weekdays (if the agent also works on Saturday that week). The sample selection is done for one survey agent in a week by selecting (4 or 5) days to sample and then selecting the shift x region for each say selected. Then the sample selection is done for the second agent subject to the restriction that the two agents can't be in the same region at the same time (i.e., same day, same region, same shift, or same day, same region, adjacent shifts).

The sampling route is travelled six times in a six-hour shift, with effort and catch rate being sampled in alternating circuits of the route. Each circuit of the route begins at one of two (fixed) starting points which are alternated.

There are 4 or 5 Saturdays in a month and 9 shift x region combinations for a total of 36 or 45 sampling units. Four of these are sampled per month for a sampling fraction of 11 or 9%. For weekdays, scheduling is done by week. Each agent works 4 or 5 weekdays per week so 8 to 10 sampling units are selected per week out of a total of 5 x 9 =45 sampling units. The sampling fraction can thus vary from 18 to 22%.

My evaluation of the sampling design is as follows.

Let us ignore for the moment the fact that the three shifts overlap in time. In this case, the design approximates a two stage cluster sampling design within stratified random sampling. One stratifies approximately by day type. It is not exactly stratification because the choices made in scheduling sampling on weekends affects the sample selection for weekdays, so the two “strata” are not sampled independently although the co-dependence in the two sampling schedules is minimal. Then, within each day-type stratum one uses two-stage cluster sampling where the day is sampled at the first stage (the day is the cluster), and the shift x region combination is selected at the second stage. One could modify the current sampling design so that both teams sample the same days. Then, each cluster (day) sampled would have 2 of the 9 sampling units (shift x region) sampled. The variance of the total would have two components, the between day and within day variability. The variance of the total would thus be

Var(estimated total) = $(1 – f_1)$ x first stage variance + $(1 – f_2)$ x second stage variance

where $f_1 = n/N$ = number of days sampled divided by number of days available to be sampled = fraction of days sampled, and

$f_2 = m/M$ = number of sampling units in a day that are sampled (= 2) divided by number of sampling units in a day (= 9).

For weekends, f_1 = 2 days out of 4 or 5; for week days, f_1 = 4 or 5 out of 5. For both weekends and weekdays, $f_2 = 2/9$. Thus, the finite population corrections f_1 and f_2 lead to substantial reductions in variance. They essentially say, the closer you come to sampling everything, the less uncertainty in the total.

Under the present design, the number of shift x region combinations sampled on a given day is random rather than being fixed at 2. One can still estimate the within (second stage) component of variability from those days where two sampling units were selected. It would be a rather straightforward modification to make the number of units selected at the second stage be a constant, 2.

In the current design, the three shifts overlap in time. Conceptually, one can divide the data into subsets corresponding to nonoverlapping periods of time and treat the data as coming from three time-of-day strata. However, the observations among the strata are not independent. Although this doesn’t make it difficult to compute point estimates of total catch, it makes it difficult to assess variability in the estimates, i.e., to compute valid estimates of variance.

I would recommend that the overlapping time shifts be eliminated. There are various ways this could be accomplished and the various options should be evaluated in light of sampling experience over the years.

The starting point (location) for driving the survey route is not randomized. To be assured of unbiased estimates it is necessary to randomize the starting point so that any location can be sampled at any time with uniform probability. This is admittedly logistically inconvenient because it requires time to travel to the randomized starting point prior to doing any sampling and also to return from the randomized end point at the end of the sampling. Nonetheless, the randomization is necessary to ensure unbiasedness.

Shore-based design, (Commonwealth of the Northern) Mariana Islands.

The survey design and sample selection procedure is as follows.

A survey lasts 3 months. Within a survey, estimates are computed separately for weekends and weekdays. In each day, there are four 6-hour shifts. Within a shift, the survey agent does one of the two following:

PIP		IPI	
Participation (effort)	2 hours	Interview	2 hours
Interview (catch rate)	2 hours	Participation	2 hours
Participation (effort)	2 hours	Interview	2 hours

For weekdays, 16 shifts are selected for the 3-month survey. The same is true for weekends: 16 shifts per 3-month survey. The randomization procedure is complicated. Each shift is sampled four times in three months, twice using the PIP configuration and twice using the IPI configuration. (Note: 4 shifts x 4 days devoted to each shift =16 shifts.) Selection of sampling units (day x shift combination) are subject to the additional constraints that 1) a shore based survey day can't be the day after a boat survey, and 2) two consecutive shifts can't be sampled (1 & 2, 2 & 3, 3 & 4, and 4 & 1 the next day). There is also a monthly structure imposed on the randomization so that there is an alternation of some sort in what is sampled.

My evaluation of the sampling design is as follows.

Because of the complicated system of randomization, I am not sure what are the properties of this kind of sampling. For weekdays, the sampling intensity is fairly low (16 shifts spread out over about 65 weekdays = approximately 0.25 shifts per day). Therefore, the randomization constraints may not have much impact. The same cannot be said for weekend sampling where the sampling intensity is much higher (16 shifts spread over about 26 weekend days = approximately 0.6 shifts per day).

One might think of this design as being stratified by time of day. But, this implies shifts are sampled independently of each other. This is clearly not the case because the sampling constraints do not allow for the theoretical possibilities of selecting 3 or 4 shifts in a single day to be sampled, nor do the constraints allow for adjacent shifts to be sampled. One might think of the design as being two stage cluster sampling in which days are selected at the first stage and one or two shifts are selected at the second stage (i.e., within a day). Again, the constraints do not allow for the possibility of selecting adjacent shifts within a day which could occur in 2-stage cluster sampling.

Perhaps the sampling design could be related to a lattice design where balance can be forced with respect to shift (and order).

I think some careful analysis would be necessary to determine the properties of the sampling design currently in use. I understand that the randomization procedure adopted is designed to force balance in the sampling so that the results are forced to be representative. But, the way this has been done makes the design very complicated so the properties of the sampling are not immediately apparent.

I suggest that one should take a step back and look at some simpler designs with the goal of relating the designs to known sampling techniques. One can then try to modify a design to deal with logistic constraints or to achieve greater efficiency through forced balance. For example, one could randomly sample days, and then randomly choose to sample in a day either shifts 1 and 3 or 2 and 4. This would be two-stage cluster sampling with days selected at the first stage and one pair of shifts selected at the second stage. Because only one sampling unit (one pair of shifts) is sampled at the second stage one would not be able to estimate the within and between day components of variability but one could still get a conservative (probably too large) estimate of the overall variance (see Cochran 1977, p. 279). If this is logistically inconvenient, one might consider having just two 8-hr shifts during a day. The two-stage cluster sampling design with one shift chosen at the second stage (within days) would still apply. One would presumably want to do an auxiliary study to determine that the remaining 8 hours of the day do not have much fishing effort or catch.

The starting point for roving is randomly selected from just two possibilities (the north or south end points of the route). To have unbiased results, one should randomly select a starting point along the route.

Boat-based design, Guam.

The survey design and sample selection procedure is as follows.

A survey lasts one month. Estimates are computed separately for weekends and weekdays. Within a daytype (weekend or weekday), four estimates are computed: one for each of the three main ports and one for all of the lesser (unsampled) ports combined. It appears one day per month is devoted to each port on weekdays and one day per month is devoted to each port on weekend days and, similarly, one weekday and one weekend day per month is devoted to the lesser (unsampled) ports. No attempt is made to distinguish between (i.e., estimate) day and night fishing. The survey procedure for the three main ports is different from the procedure for the lesser ports. For the three main ports, there are two shifts in a day and both are sampled in a given day. The am shift starts from 5 – 6 am and extends to 11 - 12 noon. The pm shift starts at 4 pm and extends to midnight. It is believed that virtually all trips are tallied in a day in which sampling occurs. (Tallying is done on the basis of identifying boat trailers and noting slips at docks that are unoccupied. Thus, individual boats or trailers are identified). Interviewing takes place while boats are being tallied. On a sampling day, one port is sampled. For the lesser sampled ports, roving allows trailers to be counted in the lesser sampled ports. However, no interviews are obtained so the catch/trip in the lesser sampled ports is assumed to be the same as in the smallest 2 of the 3 heavily sampled ports.

My evaluation of the sampling design is as follows.

It would appear that the primary sampling unit is the day x port combination. The responses recorded from each sampling unit are the day’s effort and the day’s catch rates. The document states that stratification occurs by day type and port. Stratification implies that the different strata are sampled independently. This is clearly not the case here because if a team is in one port it can't be in another port at the same time, thus demonstrating what happens in one "stratum" affects what happens in the others.

The design would be two-stage cluster sampling (pick days, and within days pick a port) except that the investigators have imposed further constraints to achieve balance, i.e., each port is sampled the same number of times (i.e., once in a month) as the other ports. I believe this renders the design a lattice design. The design could be made into two-stage cluster sampling if the ports were selected randomly for each sampling (such that, for example, Port A might be sampled twice, Port B not sampled at all, and Port C sampled once during a particular month). The advantage is that a standard design with known properties would be used. The disadvantages are that sampling is not balanced over the ports (which intuitively seems less precise) and it may not be possible to make estimates for a particular port in a particular month. The importance of having port-specific estimates is not clear.

I am not sure how many times each port is sampled. According to Table 3, 8 days are sampled in total in a month which, when apportioned to daytype and port, comes to 1 day sampled for each port in a daytype stratum. This would not give the replicability of sampling needed to estimate variances. Therefore, I presume data from multiple months are used to compute variances (and probably point estimates). This is hinted at in the statement (p. 9 of the technical report) "The survey days are selected at pre-determined frequencies *over broader scheduling periods* [emphasis added]." This adds another dimension over the lattice design. It is possible to compute variances with this kind of design if one is willing to assume some interactions are zero. I am sympathetic to the desire to impose balance to the sampling structure to force representativeness of the results. But, the properties of the design need to be established and documented.

As it stands now, I think the boat-based survey gives reasonable estimates because of forced balance in sampling but the variances are not credible. There are some caveats: the results are contingent on the assumptions that 1) lack of sampling between noon and 4 pm does not cause bias, 2) catch rate in the unsampled ports is the same as in the two smaller sampled ports, 3) trailers for pleasure boaters are not included in the count of trailers of fishers.

I think the sampling design for Guam boat-based fishing should be re-evaluated and a clearly specified design with known properties should be adopted.

<u>Boat-based design, American Samoa</u>.

The survey design and sample selection procedure is as follows.

Surveys last one month. There are four ports that are sampled; no estimates are made for landings outside the four ports. All four ports are sampled during a sampling day. The investigators believe they can tally every boat coming in during a day. The fishing day is defined to be from 5 am to 9:30 pm (but some sampling is done from 9:30 pm to 5 am to quantify landings at this time; this is called the graveyard shift). Separate estimates are made for weekday and for weekends (where weekend is defined to be Saturday, i.e., it's assumed there is no fishing on Sunday). The entire day is sampled on each weekday sampled but only one of two shifts is sampled on a weekend. Estimates are made for whole days, i.e., separate estimates are not made for first and second shift.

Every other Saturday is sampled, and the number of days sampled during the week is reduced when there is weekend and graveyard shift sampling. Thus, scheduling is done by week.

My evaluation of the sampling design is as follows.

Although the description of the sampling says days are stratified into weekdays and weekends (Saturdays), this is only approximately true since selection of a Saturday affects the scheduling of sampling during the week. Schedules are determined by week so week is really a stratum but the investigators collapse the weekly strata. One might say that sampling approximates a simple random sample of days being selected for a month (ignoring the constraints imposed by doing the scheduling on a weekly basis, and the constraints associated with reductions in weekday sampling associated with Saturday and graveyard shift sampling). Thus, for effort sampling, the design approximates a simple random sample of days with: 1) all effort observed on a sampled weekday, and 2) the day being treated as a cluster of two shifts with one shift selected on a sampled weekend day. For catch rate sampling, the design approximates a simple random sample of days, with each day being treated as a cluster of regions (for weekdays) or region x shift combinations (for weekend days).

Because sampling intensity is high, there can be substantial reductions in variance if a finite population correction is applied. This is not appropriate (at any stage) if only one sampling unit is chosen from each cluster (because the components of variability cannot be estimated and it is inappropriate to apply a finite population correction to the total instead of the components).

Boat-based design, (Commonwealth of the Northern) Mariana Islands.

The survey design and sample selection procedure is as follows.

Similar to the boat-based survey in Guam, a survey lasts one month and there are 3 main ports surveyed plus a collection of other sites where only trailer counts are obtained (for estimating fishing effort). One weekday per week is sampled for 3 of the four weeks in a month. And, one weekend day a week is sampled for 3 of the four weeks in a month. (Thus, total sampling effort is 6d/mo.) Separate estimates are made for weekends and weekdays, but the estimates are not independent since sampling restrictions prohibit Friday-and-subsequent-Saturday and Sunday-and-subsequent-Monday sampling. A further restriction is that shore-based and boat-based sampling can't occur at the same time because the same crew is used for both shore- and boat-based sampling.

Catch rate is surveyed in one port per sampling day; effort is determined from trailer and empty berth counts in a day for all ports. Counts for estimating effort are made two times in a shift. Interviews occur throughout the sampling shift. It appears (though it is not stated explicitly in the documentation) that balance is insured by constraining the randomization so that each port is sampled once during the month.

My evaluation of the sampling design is as follows.

The design is very similar to that used in Guam. For the reason given in my description of the Guam sampling, weekdays and weekend days are technically not separate strata, but they might reasonably be treated as such. Within a daytype, the design resembles multi-stage cluster sampling, as follows: Stage 1 – pick 3 weeks out of 4, Stage 2 – pick one day within the week. For effort sampling, counts are made for all 3 ports during a sampling day. For catch rate estimation, add Stage 3 – pick a port within a day. However, in practice, the selection of the port is not fully random. There is the constraint that each port is sampled once. If this constraint were not imposed, the sampling would be three-stage cluster sampling with the obvious implication that some ports may be sampled more than once and some not sampled at all in a month.

Table 1. Characteristics of the six surveys. Eff = effort survey, CR = catch rate survey. Similar design refers to my judgement that the nominal design most closely approximates the similar design, or can be made to approximate the similar design.

survey	nominal design	similar design
shore-based		
Guam	Eff: stratify by daytype & shift	Eff: stratify by daytype, OR stratify by daytype and compute estimates by shift; if estimates are by shift, include covariance between early & late shifts when calculating variance of total. Do not use finite population correction.
	CR: stratify by daytype, shift, and region (for hook and line fishing)	CR: not clear, this is complicated
Samoa	stratify by daytype, time of day, and maybe region	stratify by daytype, use 2-stage cluster within daytype (first stage, day; 2nd stage, region x shift)
CNMI	Eff & CR: stratify by daytype & shift	weekdays: 2-stage cluster? weekends: not clear
boat-based		
Guam, boat	Eff & CR: stratified random sampling	lattice design; could be 2-stage cluster sampling if regions are selected randomly for each day (without constraints)
Samoa, boat	Eff & CR: stratified random sampling by daytype	stratification by daytype; within strata: Eff: simple random sample of days for weekdays; for weekends, simple random sample of 2 days at the first stage, with one shift chosen at the second stage. CR: two stage cluster sampling with days at the first stage and region or region x shift at the second stage
CNMI, boat	Eff & CR: stratified random sampling	lattice design; could be 2- or 3-stage cluster sampling if ports are selected randomly for each day. The extra stage (compared to Guam) comes from selecting 3 of 4 weeks.

Additional Considerations

Randomization in roving designs

For the roving design, there is also an issue concerning randomization of routes. The theory behind estimating fishing effort in gear usage hours is that each possible fishing location can be viewed for an instant at any time with equal probability (during the completion of a given route in a given time interval) if the starting time, starting location, and direction of travel are randomized (see Hoenig et al. 1993 and the discussion in the Background section of this report). Randomizing the starting time is generally easy. Randomizing the starting location is logistically inconvenient. Suppose the route can be thought of as being a circle, and the nearest point of the circle to home base (e.g., the fisheries department headquarters) is called Point A. It is most convenient to start the route at Point A, travel the circle, and end up back at Point A at the end of the shift. Then locations close to Point A are always sampled at the extremes of the shift and points on the far side of the circle are always sampled mid-shift. There is zero probability that locations close to Point A are sampled mid-shift. This causes bias of unknown direction and magnitude. The proper procedure is to randomly pick the starting location along the circle with equal probability. This is inconvenient because the total distance traveled is the circumference plus two times the distance from Point A to the starting location, but this is what is required by the design. (Sometimes, a route appears to be linear between two points A and B (e.g., a section of the coastline). However, a linear route can be made circular by considering the route to be travel from A to B and then back to A. Again, this can be inconvenient because, if the randomly selected starting point is close to Point B, one needs to travel almost to B, then start the count or interviewing and travel to Point B, then go back to Point A (while not counting fishers), and proceed to the starting location close to Point B counting or interviewing. Nonetheless, this is the proper procedure under the theory.) The third aspect of randomization, randomizing the direction of travel ("clockwise" vs. "counter-clockwise"), arises from a technical issue where fishers who visit more than one location during a trip may have zero probability of inclusion in the count if the speed of travel of the fisher is greater than the speed of travel of the survey agent and the direction of travel is not randomized; it is not a major concern. In Table 1 it is stated that for American Samoa and CNMI the starting point is always one of the ends of the route which is not valid; the procedure used on Guam appears to be correct.

It is stated on page 5, for the CNMI survey, that a shift is divided into 3 parts of equal length and counts are made during the first and third parts; interviewing is done during the middle third. For this to be unbiased, the interviewing part should be randomized so that the three possibilities (interview during first, second or third part of the shift) have equal probability. Table 1 suggests the same problem may apply to American Samoa. Also, it is stated (p.6) for Guam that "It is assumed that the fishing activities observed during the survey time (finishing around noon during day shift and finishing around midnight during night shift) are representative of those during daytime fishing hours (0600 – 1800 hours) or nighttime fishing hours (1800 – 0200 hours) (Amesbury et al., 1991)." This indicates the survey counts were not randomized in violation of the requirements of the method.

Also, I'm perplexed that, for CNMI, the interview time equals the count time (the parts of a shift are of equal length) (p.5). Sometimes, survey agents combine interviewing with counting fishers.

But, it is certainly possible to separate the counting from the interviewing. In this case, the time to make a count should be much faster than the time needed to travel the route and interview everyone encountered. (That's a prime reason for separating counting from interviewing – you could make several complete (entire route) counts quickly and one complete route interviewing everyone encountered.) If the time devoted to a count is equal to the time devoted to interviewing that implies that not all fishers along a complete route are interviewed, i.e., either every kth fisher is interviewed, or a randomly selected portion of the route is covered. Exactly what was done is not specified so it is not clear if proper randomization was used. A practical way to implement the interviewing is to have the survey agent cover an entire route, and to have a series of "mileposts" where the agent has to be at specified times, e.g., one at the start/end, and one at ¼, ½, and ¾ of the distance around the route. The survey agent has to be at the ¼, ½, and ¾ mileposts at the ¼, ½, and ¾ fractions of the allocated travel time. This forces an approximate constant rate of travel and thus approximate randomization.

Number of domains

As indicated above, there are two reasons to define domains: because they are of intrinsic interest and to increase precision of a survey. It does not appear that many of the domains are of intrinsic interest. For example, if quota is not allocated by gear sector then it may not be important how much of the catch is taken by each gear type. Indeed, it is often difficult to obtain 3 observations for a domain suggesting the domain is not terribly important in itself.

By defining domains, one can increase or decrease the survey precision depending on whether the means differ greatly among domains. Precision can be reduced if domain means do not differ because then more parameters are being estimated with smaller sample sizes. For example, suppose Domains A and B have the same size, and the same mean and variance. Further suppose they are surveyed with equal intensity. Then two totals are being estimated, each from sample sizes of n. If the two domains are collapsed, then one total is being estimated with a sample of size 2n and precision is increased. On the other hand, if the two means have very different means, then collapsing domains increases the variability greatly and this may lead to decreased precision.

Another determinant of the efficacy of using domains is whether the sizes of the domains are known a priori versus the sizes of the domains need to be estimated. For example, it might be known with certainty that 25% of the trips are charter and 75% are private, or it might be that one estimates that 25% of the trips are charter. Greater precision is had when the sizes of the domains are known.

Defining too many domains can cause problems when sample sizes for the domains approach zero. The solution adopted by the investigators is to pool data over similar domains which defeats the purpose of defining domains.

I recommend that the investigators reconsider which strata and domains are important for the assessment and management of the fisheries to see if the sampling designs can be simplified and made to follow standard sampling designs more closely. This can be evaluated by estimating totals and variances with domains collapsed (combined) in various ways.

2. Evaluate the shore-based and boat-based survey expansion algorithm and assess if the expansion methods are appropriate and consistent with the survey designs.

The algorithms used to compute estimates of catch and variance of catch (in total and by subsets (domains) of the fishery are reasonable for stratified random sampling, with a number of caveats (such as it being assumed there is no fishing at times that are never sampled). However, as noted above (see Table 1), in only a couple of cases is the sampling design close to stratified random sampling. In most cases, the sampling design is close to, or could be modified to be, two- or three-stage cluster sampling. The effect of mis-specification of the sampling design likely does not induce bias (not considering various caveats about the impact of constraints on the sampling). The mis-specification of the sampling design does have a major effect on the estimates of variance. In particular, assuming random sampling within a stratum instead of multi-stage cluster sampling tends to underestimate the variance of the estimated catch.

In a few cases, sampling effort is extensive. In some of these cases, it may be possible to apply a finite population correction (fpc) which would reduce the estimated variance. The fpc is somewhat intuitive. Suppose half of all possible sampling units are completely surveyed. Then for half of the fishery there is no uncertainty and the estimated variance applies only to that part of the fishery that was not sampled. Hence, the uncertainty in the estimate is reduced by half since half was completely sampled and is known exactly. The use of the finite population correction should be investigated once proper sampling designs have been specified for each survey.

I identified a few issues with the expansion algorithms as follows.

Shore-based Expansion for estimating effort. There appears to be a problem with equations (1) through (3), as follows. The term g_i is defined to be “the number of gears counted in an effort survey” (p. 6) and $\bar{g}$ is defined as the average gear count across survey days (and across replicate counts within a shift). Thus, g_i refers to a single count, i.e., at one route/region/port and one shift. To get the daily effort, the count must be expanded by the number of shifts in the day and the number of routes/regions/ports. Similarly, the variance of the mean daily effort needs to include the expansion factors used to convert counts to daily estimates. Furthermore, for Guam, there are two counts per shift so that there is within day and among day components of the variance that should be accounted for using the theory for two-stage cluster sampling. The description of the method does not indicate that expansion from counts to daily estimates and proper computation of variance was not done.

There is a correction factor, p_2, for areas unsampled by the survey agent. This is sometimes assumed and sometimes derived from aerial counts in the unsampled areas during the day time. No correction is made for areas unsampled at night because there are no aerial counts at night. Thus, it appears p_2 is set to zero for the night; it is unclear to me why the day time p_2 is not applied to night fishing as well.

In equation (4), p_2 is treated as a known constant when, in fact, it is measured with error. Since there are multiple observations of p_2 in some cases it is possible to estimate the variance of the estimate of p_2 and this should be included in the variance of *ghr*.

Catch rate estimator for shore-based (roving) design. There are two general approaches to estimating catch rate, the ratio of means and the mean of ratios estimators. This project utilizes the ratio of means but the literature suggests the mean of ratios estimator with truncation (elimination) of trips of short duration (say, less than 30 minutes duration) is better because it is more precise (Hoenig et al. 1997, Pollock et al. 1997).

Pooling of observations across similar sampling situations (domains). It is not uncommon that sample sizes are too small to obtain reliable estimates of variance for some subsets of the fishery, e.g., for certain combinations of gear, location, daytype, etc. This leads to pooling of data across similar sampling situations. This is not unreasonable. For example, one sometimes has a large number N of strata and one takes a random sample of n = 1 from each stratum. One cannot calculate a variance from one number. However, one can use the method of collapsed strata to obtain samples of size n = 2. As an example, suppose there is a fish kill in a lake and all the fish are blown by the wind to the northern shore of the lake. One could divide the north shore into N strata, each one km long and then take a random sample of one quadrat per stratum (per km) to count the dead fish. We might combine the first and second strata, the third and fourth strata, etc. so that each of the new "collapsed" strata has two observations, thus affording the opportunity to calculate a variance. To the extent that adjacent strata are "similar" in terms of the number of dead fish this provides a reasonable estimate of the variance. However, to the extent that the true means in the adjacent strata differ, the estimated variance will be biased high because the estimate is the average deviation from the overall mean, not the stratum means. In the case of port sampling in the three island groups, a minimum sample size of 3 has been specified by the researchers, a not unreasonable requirement. Suppose one sampling domain, A, has 2 observations and another, B, has 3. I can envision two ways to compute the variance. One way is to calculate the average squared deviation from the mean of the 5 observations. Another way is to calculate the sum of the squared deviations from the domain mean separately for each domain, add the two sum of squares, and divide by the degrees of freedom. In symbols, the two approaches would be:

$$S_1^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

$$S_2^2 = \frac{\sum_{i=1}^{2}(y_{iA} - \bar{y}_A)^2 + \sum_{i=1}^{3}(y_{iB} - \bar{y}_B)^2}{n-2}$$

Here, n is the total number of observations (= 5), y_i is an observation from the pooled sample, y_{iA} is an observation from sample A, and y_{iB} is an observation from sample B.

There may be guidance in the literature about the appropriate method to use but I am not familiar with such work. Clearly, if the true means in the two domains are the same, then the first formula is to be preferred. To the extent the means differ, the second estimator is likely to provide better results. But, the whole point of pooling data is to combine data from similar domains and, if they're similar, then the first formula should be preferred.

The above procedure is for estimating the variability of the observations, S^2. This, then, is used to estimate the variance of the mean as S^2/n. I believe that, here, n should be the number of observations in the domain for which pooling was needed, i.e., in the example considered the n should be 2. I didn't see this discussed in the report and I'm afraid I didn't spot this in the R code.

3. Review the boat-based expansion script (in R) for Guam and assess if the expansion methods in the R script are consistent with the expansion algorithm described the PIFSC tech memo.

As indicated above, I believe that the sampling design has been mis-specified by the investigators. This should not cause appreciable bias in the estimates of catch rate, effort and catch. But, the estimates of variance are not appropriate in my opinion. It did not seem necessary to review the code in detail given that the code will have to be rewritten. I therefore offer more general suggestions about the computer code.

The code (in three files) was written to handle all of the surveys. This is consistent with an important principle for trying to achieve the goal of error-free programming called DRY: Don't Repeat Yourself. Thus, any desired change in the computer code need be made only once (in one program), greatly reducing the chances of making an error (when implementing the change six times, once for each survey). But, because the surveys vary in their design, some thought should be given to writing separate programs for each survey design. (Some functions will be usable by all programs.) This is consistent with a different principle, KISS: Keep it Simple, Stupid. Essentially, the expansion procedure is necessarily complicated to deal with variable data formats, complicated data pooling procedures and other features of the sampling programs. The program or programs will need to be rewritten to account for the fact that some surveys use stratified random sampling and some use two- or three-stage cluster sampling (and some, possibly, use lattice sampling).

Programs could include additional features. It may be helpful to compute estimates for various definitions of sampling domains (i.e., to allow for one to explore the effects of collapsing sampling domains into a smaller number of domains). Also, it would be helpful to include sensitivity analysis of the effect of changing the value of various assumed parameters.

4. <u>Review the sensitivity analyses conducted for the boat-based survey in Guam regarding the impact of the interview pooling algorithm and selection of representative ports and provide suggestions if available.</u>

The sensitivity analyses focused on two aspects of the expansion algorithm: choice of pooling algorithm and selection of representative ports. These analyses are good, as far as they go.

With respect to pooling algorithms, I wish the authors had additionally considered collapsing (combining) domains as another option. If one domain has 2 observations and another has one observation, then pooling results in 3 observations for each domain. Collapsing the two domains would result in 3 observations for the collapsed (combined) domains. If one domain has 2 observations and another has 5, then their pooling results in 7 observations for one domain (2 + 5 = 7) while the other domain has 5 observations (since 5 is greater than the minimum requirement of 3). Collapsing the two domains would result in 7 observations for the collapsed stratum.

With respect to representative ports, I think the sensitivity analyses are clear, appropriate and appropriately interpreted.

5. <u>Provided recommendations for future research/improvements for the survey design and expansion methods.</u>

Conclusions

1) The overall designs look reasonable in the sense that sampling is spread out over time and space and the amount of resources devoted to sampling are modest to extensive.
2) The designs were described as stratified random sampling but in most cases they actually resemble two- or multi-stage cluster sampling (sometimes within strata) or lattice sampling within strata.
3) The mis-specification of the sampling designs probably results in little or no bias in the point estimates but likely causes the variances to be underestimated, possibly by a significant amount.
4) There are numerous constraints imposed on the randomization of sampling schedules. This induces uncertainty in the properties of the sampling design.
5) Despite criticisms of the sampling designs, I believe best available science is being used. That does not mean that improvements cannot be incorporated into the sampling designs.

Recommendations

1) The technical report should be revised to describe the designs, randomization procedures, and data analysis procedures clearly and unambiguously for each survey.
2) Each of the study designs needs to be evaluated to see if it can be related to a standard design such as two-stage cluster sampling; some modifications of survey design (scheduling) might be needed to make the design fit a standard design. Examples of these modifications include eliminating overlapping shifts, reducing the number of constraints imposed on the randomization of sampling schedules, reducing the forcing of balance in the sampling, and reducing the number of domains.
3) Randomization of start points and start times in is needed in some of the effort surveys.
4) In the expansion algorithm for the roving design, consideration of the mean of ratios estimator with truncation of short trips for calculation of cpue should be considered, and justification needs to be provided for choice of catch rate.
5) Use of finite population corrections should be considered.
6) Consideration should be given to reducing the number of domains. This would reduce the need for pooling data. If domain means do not vary significantly, having many domains can reduce precision by requiring more parameters to be estimated with fewer observations per parameter.
7) Sensitivity analyses should be performed to see the effect of mis-specifying assumed constants (p_1, p_2, reference table consisting of method-specific CPUE values).

Literature Cited

Cochran, W.G. 1977. Sampling Techniques, third edition. John Wiley.

Hoenig, J.M., D.S. Robson, C.M. Jones, and K.H. Pollock. 1993. Scheduling Counts in the Instantaneous and Progressive Count Methods for Estimating Sport Fishing Effort. N. Am. J. Fish. Manage. 13:723-736.

Hoenig, J.M., C.M. Jones, K.H. Pollock, D.S. Robson and D.L. Wade. 1997. Calculation of Catch Rate and Total Catch in Roving Surveys of Anglers. Biometrics 53:372-382.

Pollock, K.H., J.M. Hoenig, C.M. Jones, D.S. Robson, and C.J. Greene. 1997. Catch rate estimation for roving and access point surveys. North American Journal of Fisheries Management 17, 11-19.

Appendix 1: Bibliography of materials provided for review

Jasper, W., T. Matthews, J. Gutierrez, T. Flores, B. Tibbatts, N. Martin, J. Bass, S. Wusstig, R. Franquex, F. Manibusan, J. Ducusin, A. Regis, M.K. Lowe, M. Quach. 2016. DAWR Creel Survey Methodology. NOAA Fisheries PIFSC-JIMAR TSI. 109 p.

Ma, H., T. Matthews, M. Nadon and F. Carvalho. 2021. Shore-based and boat-based fishing surveys in Guam, the CNMI, and American Samoa: survey design, expansion algorithm, and a case study. NOAA Tech. Mem. NMFS-PIFSC ###.

Oram et al. Guam boat-based creel survey. PIFSC administrative report.

Oram et al. Saipan boat-based creel survey. Unpublished.

Oram et al. American Samoa boat-based creel survey. Unpublished.

R script for boat-based survey expansion in Guam.

In addition, we (the reviewers) posed certain questions to the NMFS staff and received the emails listed below in reply which I reviewed.

Email from T. Matthews dated 16 September 2021. This was in response to an email from J. Volstad dated 16 September asking for clarification of procedures used for shore-based sampling in American Samoa.

Email from T. Matthews dated 20 September 2021. This was in response to an email from J. Hoenig dated 19 September asking for clarification of procedures used for shore-based sampling in American Samoa.

Word document on boat-based surveys sent by T. Matthews on 20 September. 6 p.

Word document on shore-based surveys sent by T. Mathews on 20 September. 11 p.

Email on sampling in Guam (CPUE during the graveyard shift, scheduling sampling on weekends) sent by T. Mathews on 4 October 2021.

Email on standard errors in Hawaii sent by H. Ma on 8 October 2021.

Email exchanges between J. Volstad and T. Matthews copied to Hoenig and NMFS staff dealing with shore-based survey in CNMI, 18-19 October. (Volstad, 18 October; Matthews, 19 October; Volstad 2 emails on 19 October)

Appendix 2: A copy of the CIE Statement of Work

Statement of Work
National Oceanic and Atmospheric Administration (NOAA)
National Marine Fisheries Service (NMFS)
Center for Independent Experts (CIE) Program
External Independent Peer Review

Shore-based and boat-based fishing surveys in Guam, CNMI, and American Samoa: Survey design, expansion algorithm, and a case study

Background

The National Marine Fisheries Service (NMFS) is mandated by the Magnuson-Stevens Fishery Conservation and Management Act, Endangered Species Act, and Marine Mammal Protection Act to conserve, protect, and manage our nation's marine living resources based upon the best scientific information available (BSIA). NMFS science products, including scientific advice, are often controversial and may require timely scientific peer reviews that are strictly independent of all outside influences. A formal external process for independent expert reviews of the agency's scientific products and programs ensures their credibility. Therefore, external scientific peer reviews have been and continue to be essential to strengthening scientific quality assurance for fishery conservation and management actions.

Scientific peer review is defined as the organized review process where one or more qualified experts review scientific information to ensure quality and credibility. These expert(s) must conduct their peer review impartially, objectively, and without conflicts of interest. Each reviewer must also be independent from the development of the science, without influence from any position that the agency or constituent groups may have. Furthermore, the Office of Management and Budget (OMB), authorized by the Information Quality Act, requires all federal agencies to conduct peer reviews of highly influential and controversial science before dissemination, and that peer reviewers must be deemed qualified based on the OMB Peer Review Bulletin standards[1]. Further information on the Center for Independent Experts (CIE) program may be obtained from www.ciereviews.org.

Scope

The National Marine Fisheries Service is seeking a desk review to evaluate the survey design and expansion algorithm of shore-based and boat-based fishing surveys in American Samoa, Guam, and the Commonwealth of the Northern Marina Islands (CNMI).

The shore-based effort and catch surveys each utilize a roving survey design. During the roving catch survey, encountered fishers are interviewed to gather data on fishing methods, hours fished, and fish caught (or released). In the roving effort survey, accessible shoreline is visited to

[1] http://www.cio.noaa.gov/services_programs/pdfs/OMB_Peer_Review_Bulletin_m05-03.pdf

record active fishing methods and gear counts. The shore-based survey is stratified by day type (weekday vs weekend) and shift (different periods in a day) in all three territories. The boat-based survey is mainly an access point survey by design. Catch and effort surveys are conducted at major ports and the surveys are stratified by day type and port (except for American Samoa). For both shore-based and boat-based surveys, expanded catch is estimated as the product of catch rate (from catch survey) and fishing effort (from effort survey).

Local departments in each territory conduct boat-based and shore-based creel surveys: the Guam Department of Agriculture, Division of Aquatic and Wildlife Resources (DAWR); the CNMI Department of Lands and Natural Resources, Division of Fish and Wildlife (DFW); and the American Samoa Government Department of Marine and Wildlife Resources (DMWR). Data management and programming support for the creel surveys are provided by NOAA's Western Pacific Fishery Information Network (WPacFIN). The creel survey data have been used to estimate and report total catch, catch-per-unit-effort (CPUE), and fishing effort, mostly at annual levels across island areas. The statistical method used to compute these estimates is referred to as the expansion algorithm.

Expanded catch and effort estimates (and other estimates such as CPUE) are used by the territorial agencies in their fiscal year reports, and in calendar year reports to the various plan teams and committees under the Western Pacific Regional Fisheries Management Council ("the Council"). The Council and NOAA Fisheries evaluate estimated landings with respect to Annual Catch Limits established under the Revised Magnuson-Stevens Fisheries Management Act for the US Pacific Islands Region (PIR). Both raw and expanded data are for various uses, including for annual reports on US PIR fisheries such as *Fisheries of the US* and *Annual Stock Assessment and Fishery Evaluation (SAFE) Reports*.

The goals and objectives specific to the review of the Pacific Islands territorial fishing surveys are to:

1) Evaluate the survey design for the shore-based and boat-based surveys to estimate catch rate and fishing effort estimates.
2) Evaluate survey expansion algorithm for catch rate, fishing effort, and expanded catch estimates (point estimators and variance estimators included in the document and the expansion script in R)
3) Suggest future research priorities to improve the existing survey design and expansion algorithm.

The specified format and contents of the individual peer review reports are found in **Annex 1**. The Terms of Reference (ToRs) for the review of the surveys are listed in **Annex 2**.

Requirements

NMFS requires two reviewers to conduct an impartial and independent peer review in accordance with the SOW, OMB Guidelines, and the ToRs below. The reviewers shall have working knowledge and recent experience in survey design and catch/effort expansion for fishery-dependent surveys. The reviewers should be a survey statistician with comprehensive knowledge of both theoretical and applied sampling design and analysis. Furthermore, the reviewers should have a proficient understanding of the R programming language, familiarity with the R package “dplyr”, and experience with data manipulation using the R function “array”. Experience in marine fisheries is beneficial. Each CIE reviewer’s duties shall not exceed a maximum of 10 days to complete all work tasks of the peer review described herein.

Tasks for reviewers

Each CIE reviewer shall complete the following tasks in accordance with the SOW and Schedule of Milestones and Deliverables herein.

Pre-review Background Documents: Review the following background materials and reports prior to the review:

a) Guam creel survey manual (Jasper et al.)
b) Guam boat-based creel survey (Oram et al., packaged as PIFSC administrative report but not officially released)
c) Saipan boat-based creel survey (Oram et al., not officially released)
d) American Samoa boat-bases creel survey (Oram et al., not officially released)

Desk Review: Each CIE reviewer shall conduct the independent peer review in accordance with the SOW and ToRs, and shall not serve in any other role unless specified herein. Modifications to the SOW and ToRs cannot be made during the peer review, and any SOW or ToRs modifications prior to the peer review shall be approved by the Contracting Officer’s Representative (COR) and the CIE contractor.

a) PIFSC tech memo “Shore-based and boat-based fishing surveys in Guam, CNMI, and American Samoa: Survey design, expansion algorithm, and a case study”
b) R script for boat-based survey expansion in Guam

Contract Deliverables - Independent CIE Peer Review Reports: Each CIE reviewer shall complete an independent peer review report in accordance with the SOW. Each CIE reviewer shall complete the independent peer review according to required format and content as described in Annex 1. Each CIE reviewer shall complete the independent peer review addressing each ToR as described in Annex 2.

Place of Performance

Each CIE reviewer shall conduct an independent peer review as a desk review, therefore no travel is required.

Period of Performance

The period of performance shall be from the time of award through July 2021. Each reviewer’s duties shall not exceed 10 days to complete all required tasks.

Schedule of Milestones and Deliverables: The contractor shall complete the tasks and deliverables in accordance with the following schedule.

Within two weeks of award	Contractor selects and confirms reviewers (2)
Within four weeks of award	Contractor provides the pre-review documents to the reviewers
August 2021	Each reviewer conducts an independent peer review as a desk review
Within two weeks after review	Contractor receives draft reports
Within two weeks of receiving draft reports	Contractor submits final reports to the Government

Applicable Performance Standards
The acceptance of the contract deliverables shall be based on three performance standards: (1) The reports shall be completed in accordance with the required formatting and content (2) The reports shall address each ToR as specified (3) The reports shall be delivered as specified in the schedule of milestones and deliverables.

Travel
Since this is a desk review travel is neither required nor authorized for this contract.

Restricted or Limited Use of Data
The contractors may be required to sign and adhere to a non-disclosure agreement.

NMFS Project Contact
Felipe Carvalho
Supervisory Research Mathematical Statistician
Fisheries Research and Monitoring Division
NMFS | Pacific Islands Fisheries Science Center
1845 Wasp Blvd, Bldg #176 | Honolulu, HI 96818
Felipe.Carvalho@noaa.gov
808-725-5605

Annex 1: Peer Review Report Requirements

1. The report must be prefaced with an Executive Summary providing a concise summary of the findings and recommendations, and specify whether or not the science reviewed is the best scientific information available.

2. The main body of the reviewer report shall consist of a Background, Description of the Individual Reviewer's Role in the Review Activities, Summary of Findings for each ToR in which the weaknesses and strengths are described, and Conclusions and Recommendations in accordance with the ToRs.

3. The reviewer report shall include the following appendices:
    a. Appendix 1: Bibliography of materials provided for review
    b. Appendix 2: A copy of the CIE Statement of Work

Annex 2: Terms of Reference for the Peer Review

Shore-based and boat-based fishing surveys in Guam, CNMI, and American Samoa: Survey design, expansion algorithm, and a case study

6. Evaluate the shore-based and boat-based fishing survey designs and determine if the survey designs are appropriate.
7. Evaluate the shore-based and boat-based survey expansion algorithm and assess if the expansion methods are appropriate and consistent with the survey designs.
8. Review the boat-based expansion script (in R) for Guam and assess if the expansion methods in the R script are consistent with the expansion algorithm described the PIFSC tech memo.
9. Review the sensitivity analyses conducted for the boat-based survey in Guam regarding the impact of the interview pooling algorithm and selection of representative ports and provide suggestions if available.
10. Provided recommendations for future research/improvements for the survey design and expansion methods.