Data review and sampling verification

IOTC Species identification and sampling workshop

Kochi, India, September 29th to October 3rd, 2025

Food and Agriculture Organization of the United Nations

INDIAN OCEAN Tuna Commission

Outline

- Objectives of the session
- Introduction

1. Sample information
2. Dispersion parameters
3. Accuray of the estimation

Objective of the session

- Identify sampling issues
- Have a critic eyes on your sampling
- Be aware of the estimation quality based on the sample's quality

Introduction

Different way to control the quality of samples (and the estimations)

1. Summarize sample information:
2. Dispersion parameters: Standard deviation, Intervale of Confidence, Coefficient of Variation
3. Sampling uniformity index (SUI)
4. Accuracy for CPUE and for AC
   - Spatial and temporal accuracy
   - 2 approaches: probabilistic and algebraic
   - Overall accuracy

1. Sample information

- List your sample sites
  - They must be represented the fishing activity: when you miss some major sites your estimation will probably be biased
- Verify all strata has been sampled
  - Otherwise, you won't be able to produce the statistics
- Number of sample per month per fishing unit (per site…)
  - For landings
  - For effort survey
- Compare with what you expected
  - If number of samples is fine (sufficient or even bigger)
  - If number of samples is much lower -> your sample will probably not give you accurate estimation (unless the variability inside the stratum is very low)

2. Dispersion parameters

1. Summarise your sample information:
   - Mean (arithmetical)
   - Median
   - Quartiles

$$A = \frac{1}{n}\sum_{i=1}^{n} x_i$$

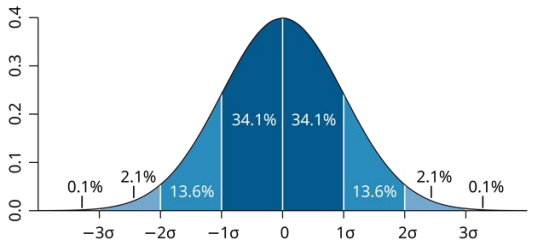
- CPUE and AC dispersion parameters:
  1. Standard deviation (SD)
  2. Intervale of Confidence (IC)
  3. Standard error of the mean (SEM)
  4. Coefficient of Variation (CV)
  5. Relative standard error (RSE)

Standard deviation

- Standard deviation also measures the dispersion of a dataset relative to its mean

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- The meaning of the standard deviation is that 95% of the data are included in the interval $[\bar{x} - 2\sigma; \bar{x} + 2\sigma]$



Coefficient of variation

- **Coefficient of variation** (CV) is a standardized measure of dispersion of a probability distribution or frequency distribution. It is defined as the ratio of the standard deviation to the mean and often expressed as a percentage. The higher the CV, the higher the dispersion around the mean. On the contrary the lower the CV, the more precise the estimation.

$$CV = \frac{\sigma}{\bar{x}} \ (\times 100)$$

Standard error of the mean

- **Standard error of the mean** (SEM) indicates how different the population mean is likely to be from a sample mean. Standard error matters because it helps you estimate how well your sample data represents the whole population.

$$SEM = (\sigma_{\bar{x}}) = \frac{\sigma}{\sqrt{n}}$$

Interval of confidence

- The best way to report the standard error is in a **confidence interval.** A confidence interval is a range of values where an unknown parameter is expected to lie most of the time. It provides a meaningful interval easier to interpret.

$$CI = \bar{x} \pm Z^*(\sigma_{\bar{x}})$$

- Where: Z* is the critical value of the Z distribution. For a 95% confidence interval Z*=1.96.

Relative standard error

- Relative standard error (RSE) is the coefficient of variation of the standard error of the mean (SEM). This formula calculates the relative variability of the standard error about the mean, which gives an idea of the relative uncertainty in the estimate of the sample mean. It is also called the relative standard error rate or coefficient of variation of the standard error. This coefficient is useful for assessing the precision of the estimated mean in statistical studies, especially when working with samples.

$$RSE = \frac{\sigma_{\bar{x}}}{\bar{x}} \ (\times 100)$$

3. Sampling uniformity index (SUI)

SUI for FAC

$$Average\ interviews\ per\ day = \frac{Total\ number\ of\ fishers\ effort\ interviews}{Number\ of\ calendar\ days\ sampled}$$

For each sampled day, $Ratio1 = \frac{Number\ of\ interviews}{Average\ interviews\ per\ day}$ and if Ratio1>1, Ratio1<-1.

$$SUI\ FAC = Arithmetical\ mean(Ratio1)$$

SUI for CPUI

$$Average\ sampled\ landings\ per\ day = \frac{Total\ number\ of\ landings\ sampled}{Number\ of\ calendar\ days\ sampled}$$

For each sampled day, $Ratio2 = \frac{Number\ of\ landings\ sampled}{AverageAverage\ sampled\ landings\ per\ day}$ and if Ratio2>1, Ratio2<-1.

=> Interpretation: SUI is considered good if SUI > 0.6 => Sample is well distributed

3. Accuracy: Probalistic approach

Formula:

$$A = 1 - 1.96 \frac{\sigma_R}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

with

$$\sigma_R = \sqrt{\frac{2N - 1}{6(N - 1)} - \frac{1}{4}}$$

This formula arises from estimating the accuracy of a measurement or proportion under probability sampling with correction for a finite population.

3. Accuracy: Algebraic approach

On algebraic approach for large populations, the accuracy A is defined by the equation (Stamatopoulos, 2014):

$$A = a_1 + a_2 N^{-kx}$$

The parameters of the equation are computed as follow:

$$x = \frac{\ln(n)}{\ln(N)}$$

$$W = 0.75(1 - \frac{1}{N})$$

$$a = \frac{2WN^2}{(N-1)^2} - \frac{N+1}{N-1}$$

$$g = a + \frac{1-a}{N}$$

$$S = (1-a)(\frac{1}{\ln N} - \frac{1}{N \ln N} - \frac{1}{N})$$

$$k = \frac{-2}{\ln N} \ln\left(\frac{S}{1-S-g}\right)$$

$$a_2 = \frac{(1-S-g)^2}{2S+g-1}$$

$$a_1 = g - a_2$$

3. Exemple: Estimate the sample size

2 options interesting to compare:

- With tables: in case no previous data are available

- Calculations:
  - Probabilistic approach

$$n = \left[\frac{1}{N} + \left(\frac{1-A}{1.96\sigma_R}\right)^2\right]^{-1}$$

  - Algebraic approach

$$n = \left(\frac{A - a_1}{a_2}\right)^{-\frac{1}{k}}$$

  - n retained = smaller result from the 2 approaches

Thank you for your attention

Questions