Bayesian sample size for estimating the concentration of organisms in ballast water

1 Introduction

With the growth of maritime traffic and the use of large ships, ballast water has become a leading dispersing agent of harmful organisms (David et al., 2007; Souza et al., 2001). According to Ruiz et al. (2000) human pathogens are also being transported through ballast water. As an example, we mention a cholera epidemic occurred in Peru in 1991 and quickly spread through Latin America (McCarthy et al., 1992).

The D-2 Regulation of the International Convention for the Control and Management of Ship Ballast Water and Sediments adopted by the International Maritime Organization (IMO) in 2004 sets upper limits on the concentrations of living organisms in ballast water discharges to which ships must comply with. In particular, this regulation requires that ballast water discharged by ships contain fewer than 10 viable organisms with minimum dimension $\geq 50 \mu m$ by $m^3$.

Because of cost and time restrictions, examination of the entire volume of discharge is not feasible and sampling must be considered. Different approaches to this problem have been considered in the literature. Basurko & Mesbahi (2011) adopted a binomial model, but their results suggest that large volumes of discharged ballast water should be sampled to assure compliance with reasonable accuracy. Miller et al. (2011) employed Poisson distribution models and their approach produced more realistic results. Both approaches, however, are based on the assumption that the organisms concentration is homogeneous in the ballast water tank. This may not be reasonable in practice, as pointed out by Miller et al. (2011) or Carney et al. (2013), among others. More recently, Bierman et al. (2012) and Costa (2013), working independently, suggested negative binomial models that take the expected heterogeneity of the concentration into account. In all cases, issues concerning the implications of statistical “representativeness” of the sample are discussed, without a clear definition of what this means. For the purpose of compliance with the D-2 Regulation, a “representative sample” is one for which the analysis allows an estimation of the true concentration of viable organisms with a given accuracy. Given that accuracy is
directly related to the underlying statistical model and that different models are based on different assumptions on the distribution of the concentration of organisms in the tank, heterogeneity plays a major role in the quest for statistical representativeness.

Costa et al. (2015a,b) considered specific models to accommodate different assumptions on the heterogeneity of the concentration distribution and derive threshold values, minimum number of aliquots (or sample volume) for determining whether a ship complies with the D-2 regulation under different sampling schemes and precision requirements. We use the term “organisms” *in lieu* of “viable organisms” throughout the manuscript. The parameter we are concerned with is the concentration $\lambda$ of organisms per $m^3$ (or $mL$) in the ballast water tank.

1.1 Hypothesis testing approach

In statistical terms, deciding whether a ship complies or not with the D-2 regulation is equivalent to testing the null hypothesis $\lambda \leq 10$ *versus* the alternative hypothesis $\lambda > 10$ based on a sample of $v$ $m^3$ (or $mL$) of ballast water obtained from one or more aliquots collected during the discharge. Costa et al. (2015a) approach this problem by considering different assumptions on the distribution of organisms in the tank. The authors started by assuming an homogeneous distribution, in which case, a Poisson distribution might be adopted. More specifically, denoting by $X$ the number of organisms in the sample, the probability that $X = k$, is

$$P[X = k|\lambda] = \exp(-\lambda v)(\lambda v)^k/k!,$$

for $k = 0, 1, \ldots$. If the concentration of organisms in the ballast tank is $\lambda$, then we expect to find a concentration of $v\lambda$ organisms in a sample $v$. Under this model, the decision rule is to indicate non-compliance if $X > c$ where $c$ and the minimum sample volume $v$ are determined from

$$\alpha = P[X \geq c|\lambda = 10] \quad \text{and} \quad \beta(\lambda_0) = P[X < c|\lambda = \lambda_0 > 10],$$

where $\alpha$ denotes the probability of Type I error and $\beta(\lambda_0)$ denotes the probability of Type II error for a specified value $\lambda_0 > 10$, or indicate compliance otherwise.

Ideally, $\alpha$ and $\beta$ should be very small, but this might require larger sample volumes and their choice must be based on cost considerations. Note that
if the sampled volume is composed of \( n \) aliquots and \( X_i \) denote the number of organisms observed in the \( i \)-th aliquot, then \( X = \sum_{i=1}^{n} X_i \), and the decision rule is similar. In practice, the homogeneity assumption underlying the Poisson model is seldom sustainable. The distribution of organisms in the ballast water tank depends on its configuration, on the discharge flow, on the upload location, on the lighting conditions, changes in water density etc., as discussed in Murphy et al. (2002), among others. The exact distribution is difficult if not impossible to determine. Furthermore, in practice, over-dispersion, \( i.e., \) variance larger than the mean, is usually observed, as indicated in Miller et al. (2011) or Bierman et al. (2012), for example. This characteristic thus invalidates the Poisson model. To bypass this problem, Costa et al. (2015a) considered different settings, identified appropriate models and indicated how the decision rule is in each case.

When there is little information about how the concentration varies along the deballasting process, an alternative is to consider it as a random variable \( L \) following some convenient probability distribution. A possible approach is to assume that \( L \) follows a gamma distribution, for which the probability density function is

\[
    f(\ell|\lambda, \phi) = \frac{1}{\Gamma(\phi)} \left( \frac{\ell \phi}{\lambda} \right)^{\phi} \exp\left( -\phi \ell / \lambda \right) / \ell, \quad \ell > 0,
\]

where \( \Gamma(x) = \int_{0}^{\infty} t^{x-1} \exp(-t)dt \) is the gamma function. This implies that its expected value is \( \mathbb{E}(L) = \lambda \) and the corresponding variance is \( \mathbb{V}(L) = \lambda^2 / \phi \). Different values of its two parameters \( (\lambda, \phi) \) correspond to different ways in which the concentration of organisms vary in the ballast tank.

Conditionally on a specific value \( L = \ell \) for the concentration and on the shape parameter \( \phi \), assume that the number of organisms \( (X) \) in a ballast water aliquot of \( w \) m\(^3\) (or \( mL \)) follows a Poisson distribution with mean \( w\ell \). This implies that (unconditionally) \( X \) follows a negative binomial distribution with mean \( w\lambda \) and dispersion parameter \( \phi \), \( i.e., \) with probability function

\[
    \mathbb{P}[X = k|\phi, w, \lambda] = \frac{\Gamma(\phi + k)}{\Gamma(k + 1)\Gamma(\phi)} \left( \frac{w\lambda}{w\lambda + \phi} \right)^k \left( \frac{\phi}{w\lambda + \phi} \right)^\phi, \quad \lambda, \phi > 0, \quad (1)
\]

for \( k = 0, 1, 2, \ldots, \) where \( \Gamma(x) = \int_{0}^{\infty} t^{x-1} e^{-t}dt \) is the gamma function. The parameter \( \phi \) is also known in the literature as aggregation, clustering or heterogeneity parameter. In the ballast water set-up, small values of \( \phi \) correspond to more heterogeneous distribution of the organisms in the tank. Large
values of $\phi$, on the other hand, correspond to homogeneous distributions and in this case, the negative binomial distribution may be approximated by a Poisson distribution. In this context, $\mathbb{E}(X) = w\lambda$ and $\mathbb{V}(X) = w\lambda + (w\lambda)^2/\phi$, indicating that this model may accommodate a possible over-dispersion.

Suppose that $n$ aliquots of $w m^3$ have been randomly collected from the discharged ballast water and that the corresponding number of organisms, $X_1, \ldots, X_n$ have independent Poisson distributions conditionally on their means that follow a gamma distribution. As a consequence, $X_i, i = 1, \ldots, n$ follow negative binomial distributions and $X = \sum_{i=1}^n X_i$ follows a negative binomial distribution with mean $nw\lambda$ and dispersion parameter $n\phi$, i.e., $\mathbb{E}(X) = nw\lambda = v\lambda$ and $\mathbb{V}(X) = nw\lambda + (nw\lambda)^2/n\phi = v\lambda + (v\lambda)^2/n\phi$. In this case, the compliance threshold $c$ and the minimum number of aliquots $n$ are determined from

$$\alpha = \mathbb{P}[X > c|\phi, w, \lambda = 10] \quad \text{and} \quad \beta(\lambda_A) = \mathbb{P}[X \leq c|\phi, w, \lambda = \lambda_A], \quad (2)$$

with

$$\mathbb{P}[X = k|\phi, w, \lambda] = \frac{\Gamma(n\phi + k)}{\Gamma(k + 1)\Gamma(n\phi)} \left( \frac{w\lambda}{w\lambda + \phi} \right)^k \left( \frac{\phi}{w\lambda + \phi} \right)^{n\phi}, \quad \lambda, \phi > 0, \quad (3)$$

for $k = 0, 1, 2, \ldots$. For large $\phi$ or $n$ the distribution of $X$ may be approximated by a Poisson distribution with mean $v\lambda$. More details may be obtained in Costa et al. (2015a).

### 1.2 Confidence interval approach

The hypothesis testing approach is convenient for decision making, but does not provide the magnitude of the concentration estimates. To incorporate this feature, Costa et al. (2015b) propose a solution based on confidence intervals.

Suppose that $n$ aliquots with fixed volume $w$ have been randomly collected from the tank and that the corresponding number of organisms $X_1, \ldots, X_n$ follow independent negative binomial distributions. Then, we may use the maximum likelihood estimator $\hat{\lambda} = (nw)^{-1} \sum_{i=1}^n X_i$, to estimate the concentration $\lambda$. Under this set-up, we want to determine the minimum number of aliquots such that the upper bounds on the absolute estimation error $(|\hat{\lambda} - \lambda|)$ or on the relative estimation error $(|\hat{\lambda} - \lambda|/\lambda)$ are $\epsilon_a > 0$ and $\epsilon_r \in (0, 1)$, respectively, with minimum pre-specified probability (confidence level) $1 - \delta$,
\( \delta \in (0, 1), \) \textit{i.e.}, to determine \( n \) such that for all \( \lambda \)

\[
\mathbb{P} \left[ |\hat{\lambda} - \lambda| < \epsilon_a \right] > 1 - \delta
\]

(4)

or

\[
\mathbb{P} \left[ |\hat{\lambda} - \lambda| < \epsilon_r \lambda \right] > 1 - \delta.
\]

(5)

Chen (2007, 2011) establish a method to compute the minimum sample size for constructing confidence intervals for the mean of a Poisson or of a binomial distribution with fixed confidence levels, controlling the absolute estimation error, the relative estimation error or both. The only requirement is the specification of a lower and an upper bound on the concentration, \textit{i.e.,} that \( \lambda \) belongs to a interval \([a, b]\). In this setting, Chen (2007, 2011) show that the minimum \( 1 - \delta \) for \( \mathbb{P} \left[ |\hat{\lambda} - \lambda| < \epsilon_a \right] \) or for \( \mathbb{P} \left[ |\hat{\lambda} - \lambda| < \epsilon_r \lambda \right] \), considered as functions of \( \lambda \), may be determined by computing these probabilities a finite number of times. Costa et al. (2015b) extend such results to the negative binomial distribution and show how they may be used to construct algorithms to compute the required sample sizes.

When there is no prior information about the concentration range, \textit{i.e.,} there is no \textit{a priori} information that \( \lambda \) belongs to a interval \([a, b]\), Costa et al. (2015b) propose a simple formula to compute the minimum number of aliquots required to estimate the concentration \( \lambda \), adapting the results of Chen (2008) to the negative binomial distribution. In this case, however, computation of the minimum number of aliquots requires that upper bounds be set for both the absolute and relative errors simultaneously, \textit{i.e.,} it relates to the determination of the minimum number of aliquots \( n \) such that

\[
\mathbb{P} \left[ |\hat{\lambda} - \lambda| < \epsilon_a \text{ or } |\hat{\lambda} - \lambda| < \epsilon_r \lambda \right] > 1 - \delta,
\]

(6)

where \( 1 - \delta, \delta \in (0, 1) \) is the minimum confidence level.

The proposed approach does not require asymptotic approximations; it suffices to fix the confidence level, \( 1 - \delta \), the aliquot volume \( w \), and the upper bounds for the absolute and relative estimation errors. The required number of aliquots is the smallest integer \( n \) satisfying

\[
n > \frac{\epsilon_r}{\epsilon_a} \times \frac{\log(2/\delta)}{w(1 + \epsilon_r) \log(1 + \epsilon_r) - \left[ \frac{\epsilon_r}{\epsilon_a} \phi + w(1 + \epsilon_r) \right] \log \left( 1 + \frac{w \epsilon_a \epsilon_r}{\phi \epsilon_r + w \epsilon_a} \right)}
\]

(7)
The difference between (7) and the corresponding result in Chen (2008) lies in the denominator, where in the latter, the term $\epsilon_r$ is replaced by

$$\left[ \frac{\epsilon_r}{\epsilon_a} \phi + w(1 + \epsilon_r) \right] \log \left( 1 + \frac{w\epsilon_a\epsilon_r}{\phi\epsilon_r + w\epsilon_a} \right).$$

(8)

Since, as expected, this term is greater than $w\epsilon_r$, it is clear that the number of aliquots obtained under the negative binomial distribution are larger than those computed under the Poisson distribution with the same values for $w$, $\epsilon_a$, $\epsilon_r$ and $\delta$. Note also that only the second term in the denominator of (7) depends on $\phi$. The limit of (8) as $\phi$ approaches infinity i.e., as the concentration distribution becomes more homogeneous, is $w\epsilon_r$, which suggests that for large values of the dispersion parameter $\phi$, the required number of aliquots may be obtained from the Poisson distribution.

Costa et al. (2015b) suggest that when the aggregation parameter is large, i.e., when the distribution of organisms in the tank is homogeneous, the total sample volume is constant irrespectively of the aliquot volume, suggesting that the choice of the aliquot volume is a matter of convenience. On the other hand, when the aggregation parameter is small i.e., when the distribution of organisms in the tank is heterogeneous, the total sample volume increases, suggesting that sampling more aliquots with small volume will capture the heterogeneity of the organisms distribution with a smaller sample volume. Their simulation study suggests that when the absolute error is considered, the estimated coverage probability is greater than the minimum confidence level, even when the concentration $\lambda$ is smaller than the lower bound. The situation is reversed when $\lambda$ is greater than the upper bound. In the relative error case, the nominal confidence level is not attained only when $\lambda$ is smaller than the lower bound. In the absolute error case it is important to specify the upper bound ($b$) correctly to guarantee the minimum coverage probability $1 - \delta$, whereas in the relative error case it is important to specify the lower bound ($a$).

These authors also note that the coverage probabilities obtained in our simulation study are generally larger than the nominal level (adopted as 95%), suggesting that their results are conservative, in the sense that smaller sample volumes may be enough to ensure the required limits on the estimation errors are attained.

According to the sampling scheme suggested by Costa et al. (2015b), the decision with respect to compliance with the D-2 standards must be taken
after the entire sample is collected. This may be a problem when the concentration is large, since non-compliance will possibly be detected only after the entire deballasting process is completed. To bypass this problem the authors consider a sequential procedure in which non-compliance could be evaluated as the ballast water is discharged. A naive approach to this alternative based on the proposed sample size determination is considered as follows. Suppose that $n$ aliquots are required to guarantee that the estimation error falls within the prescribed limit. We may consider a decision based on the first $n/K$ aliquots where $K$ is a given constant. We give the ship owner the benefit of the doubt by supposing that no organisms are detected in the remaining $n - n/K$ aliquots and use these pseudo observations along with those obtained in the first $n/K$ aliquots to estimate the concentration. If the lower limit of the corresponding confidence interval for the concentration is larger than the D-2 standards, we declare non-compliance. Otherwise, we abandon the pseudo observations, take an extra $n/K$ aliquots assuming that no organisms are detected in the remaining $n - 2n/K$ aliquots and repeat the decision process.

1.2.1 Stratification case

Consider the situation where the tank may be divided into $H$ strata with volumes $V_h$ and concentrations $\lambda_h$, $h = 1, \ldots, H$ so that the overall volume is $V = \sum_{h=1}^{H} V_h$ and the corresponding concentration is $\lambda = \sum_{h=1}^{H} W_h \lambda_h$, where $W_h = V_h / V$. The strata may be defined according to different water levels or considering different tanks, for example. Note that the D-2 regulation requires that $\lambda \leq 10$ and not that each $\lambda_h \leq 10$.

Assume now that the sample volume $v$ is obtained by collecting a ballast water aliquot with volume $v_h = n_h w$ in stratum $h$, $h = 1, \ldots, H$, where $w$ is a pre-specified aliquot volume, $n_h$ is the number of aliquots sampled in stratum $h$, and that the corresponding number of organisms in an aliquot with volume $w$, $X_h$, follows a negative binomial model with mean $w \lambda_h$ and known dispersion parameter $\phi_h$. Consider the estimator $\hat{\lambda}_h = (n_h w)^{-1} \sum_{i=1}^{n_h} X_{hi}$ for $\lambda_h$, $h = 1, \ldots, H$, and $\hat{\lambda} = \sum_{h=1}^{H} W_h \hat{\lambda}_h$ for $\lambda$. The objective is to determine the sample volume such that $\mathbb{P} \left[ |\hat{\lambda} - \lambda| < \epsilon \right] > 1 - \alpha$ where $\epsilon$ denotes the maximum acceptable difference between the estimator and the value of $\lambda$. For such purpose, it suffices to establish $n_h$ such that $\mathbb{P} \left[ |\hat{\lambda}_h - \lambda_h| < \epsilon_h \right] > 1 - \alpha_h$ with $\epsilon_h = \epsilon / W_h H$ and $\alpha = \sum_{h=1}^{H} \alpha_h$, provided that the minimum ($a_h$) and
maximum \((b_h)\) values for the concentration in stratum \(h\) are specified, \(i.e.,\), that \(\lambda_h \in [a_h, b_h]\). This may be accomplished via the results of Costa et al. (2015b).

2 Proposed research plan

As discussed in Costa et al. (2015a,b), sample size determination is highly dependent on the heterogeneity of the distribution of organisms in the ballast water tank. This heterogeneity may be summarized by the aggregation parameter \(\phi\) determining the form of the negative binomial distribution adopted as a model for the number of organisms in the sample. If we have prior information that the organisms distribution in the tank is homogeneous, then the value of the aggregation parameter \(\phi\) is large; otherwise, if we know that that the corresponding distribution is heterogeneous, the aggregation parameter \(\phi\) must be small. Information on \(\phi\) must be obtained from historical or experimental data, such as in simulations of water sloshing patterns in ballast water tanks (Arai et al., 2002).

2.1 Bayesian sample size

A problem in the classical approach is the setting of the parameter \(\phi\) required to compute the sample size. A way to model the uncertainty about \(\phi\) is to set a probability distribution for the parameter and deal with it as if it were a random variable, this is possible in the Bayesian approach.

A Bayesian optimal sample size may be determined for any problem by specifying four components: (i) a likelihood function for the experiment; (ii) a prior distribution on the unknown parameters; (iii) a loss function for the decision problem; and (iv) a cost of experimentation. We define a loss function \(L(\lambda, a)\) for making decision \(a\) when the true value of the parameter is \(\lambda\), and the decision is based on a sample \(X_1, \ldots, X_n\) drawn for \(f(x|\lambda)\).

Parmigiani & Inoue (2009) consider the choice of a minimum sample size as a decision problem and specify a more detailed function, the loss function \(L(\lambda, a, n)\) for observing a sample size \(n\) and making decision \(a\) when the value of the parameter is \(\lambda\). Commonly it is defined as

\[
L(\lambda, a, n) = L(\lambda, a) + C(n),
\]

where \(C(n)\) is the cost of experimentation.
Consider that $\pi$ represents a priori knowledge about the unknown parameters, and $\delta^*_n$ is the Bayes rule with respect to that prior. The Bayes risk is defined as
\[
r(\pi, \delta_n) = \int_{\Theta} R(\lambda, \delta_n) \pi(\lambda) d\lambda, \tag{10}
\]
where $R(\lambda, \delta_n)$ is the expected loss and the Bayes rule $\delta^*_n$ minimizes $r(\pi, \delta_n)$ among all the possible $\delta_n$ when the sample size is $n$. Thus the optimal Bayesian sample size $n^*$ is the one that minimizes
\[
r(\pi, n) = r(\pi, \delta^*_n) + C(n), \tag{11}
\]
with this approach we may compute the minimum (or optimum) Bayesian sample size for estimating the concentration of organisms in ballast water for hypothesis and confidence interval approaches. For more details, see Parmigiani & Inoue (2009), for example.

We intend to consider the Poisson-gamma and other mixtures or even generalized distributions and choose the most appropriate model for the problem.

2.2 Dirichlet process

Although other parametric mixtures, such as the Poisson-lognormal, may be considered, the lack of knowledge about the distribution of organisms in the ballast water tank complicates the choice. We propose to follow a way based upon considering more comprehensive families for the mixing distribution of mixed Poisson laws that, namely, encompass standard models as special cases and account for longer tails than those concerning the first ones. We think of applying both parametric and semi-parametric models. The latter, well exemplified by a Dirichlet process mixture (see Ishwaran & James, 2002, for example), pursues the requirement of a broad flexibility.

Let $X_i, i = 1, \ldots, n,$ be a random sample of size $n$ from a distribution function $F$. In the parametric approach, we assume a known form for $F$, indexed by a finite-dimensional parameter specified a priori, but generally unknown. To allow greater flexibility in modelling and robustness against misspecification of $F$, we consider non-parametric models; paradoxically this does not mean that the corresponding models are completely free from parameters, but rather indicates that the number and nature of parameters are variable and determined somehow by the data, potentially reaching infinity.
- hence the name of massively parametric models. Reviews of some non-parametric methods in the Bayesian and classical paradigms are presented, respectively, by Müller & Quintana (2004) and Scott (1992).

One way to avoid the specification of the form of $F$ is to employ random probability measures (RPM), which are probability distributions over the space of probability measures. Ferguson (1973) introduced the Dirichlet process (DP) as an RPM. Admitting that $F$ follows a DP, symbolically, $F \sim \text{DP}(\alpha, F_0)$, means that for any measurable partition $A_1, \ldots, A_M$ of the sample space, the probability vector $[F(A_1), \ldots, F(A_M)]$ follows a Dirichlet distribution with parameter vector $[\alpha F_0(A_1), \ldots, \alpha F_0(A_M)]$, where $\alpha$ is a precision parameter and $F_0$ is a reference distribution measured on the space of the variables. Under this parametrization, $F_0$ is the prior expectation of the distribution $F$ and as $\alpha$ increases there is a greater concentration of $F$ around $F_0$, up to the extreme case where $\alpha \to \infty$ indicates that $F$ is assumed to be equal to $F_0$; on the other hand, small values of $\alpha$ (e.g., < 5) allow, in general, $F$ to deviate considerably from $F_0$ (see Congdon, 2006, for example). Given $n$ independent and identically distributed variables, the posterior distribution is $F(x_1, \ldots, x_n) \sim \text{DP}(\alpha + n, F_1)$, where $F_1 = (\alpha F_0 + n F_n) / (\alpha + n)$ and $F_n$ is the empirical distribution function of the observations. The simplicity of the properties of the DP and the ease with which the posterior distribution is obtained highlight why this model is so attractive when analyzed under a Bayesian perspective. Note that in our setting the Dirichlet process is the prior assigned to the mean concentrations of the conditionally Poissonian observations.

We intend to apply such Bayesian methods to the problem of sampling ballast water so that the posterior distribution is updated during the deballasting process and establish decision rules to identify whether compliance of the D-2 regulation is satisfied. Computational implementation of the results will be tested on an optical system developed at the Oceanographic Institute of the University of São Paulo (see Schultes & Lopes, 2009, for example).

References


