Container and line mode

Christopher M. Baker Howard Bondell Nathaniel Bloomfield
Andrew P. Robinson

October 27, 2021

Abstract

In the Australian border biosecurity system, data about shipping containers is recorded in one of two modes: *container* mode or *line* mode. The key difference between the modes is how the *directions* are recorded, that is, data about whether entries were inspected or found to be non-compliant. In general, an entry contains multiple lines of data, where each line is a single type of item. The challenge comes when an entry is in container mode, which results in any directions being listed against every line within the entry, rather than only the [specific??] lines. Container mode data therefore creates a challenge when we try to estimate the probability that certain items are non-compliant, because we can't be sure which records of non-compliance match up with which line. We develop a statistical model to use container mode data to help inform biosecurity risk of items. We use asymptotic analysis to estimate the value of container mode data compared to line mode data, do a simulation study to verify that we can accurately estimate parameters in a large dataset, and we finally apply our methods to a real dataset.

1 Introduction

Biosecurity is important.

Border biosecurity is important - pathways

Strategic analysis requires interception data.

Cargo (volumes, activity, etc)

Data capture - line and container mode

Our biosecurity is closely related to pooled testing, which is often used for disease surveillance. Within the pooled testing literature there is generally two branches: one aims to identify positives within a pool, while the other seeks to use pooled data to estimate quantities about the population. It is the latter – estimating quantities – that we are interested in. The fundamental problem is estimating a prevalence, p, within a population, when only pooled data is available [Thompson, 1962]. More recent work has focussed on improving estimates by reducing bias, either through altering the sampling strategy [Schaarschmidt, 2007, Hepworth and Watson, 2009] or by incorporating bias correction into models [Hepworth and Biggerstaff, 2017, 2021]. There have also been extensions of the problem where p is not a constant, but it is estimated using linear regression using only the pooled data [Delaigle and Hall, 2015, Chatterjee and Bandyopadhyay, 2020, McMahan et al., 2017, Liu et al., 2020]. These papers have made significant progress in fitting increasingly complex models.

In practice, pooled testing is manifestly different from the biosecurity scenario. Pooled testing exists by design, as a way to gather information about a population while reducing testing. In biosecurity, the searching is done to every individual line, and the results are only pooled at the point of data capture. Hence, we are interested in how much information we are losing due to aggregating results as container mode. Or,

equivalently, how much better could we be at managing bio-security risk if we stopped using container mode when an entry contains multiple lines?

2 Asymptotic analysis

We use asymptotic analysis to investigate how the precision of estimates depends on entry size, the number of entries, the probability of interception and whether item types are mixed. This analysis is broken into two parts. The first part assumes that all items are a single type, which allows us to see how the amount of data, probability of interception and entry size affects precision. The second part assumes that there are two items and explores how changing the proportion of entries with both item types mixed affects precision.

Throughout this asymptotic analysis we make two simplifications. Firstly, we do not separate line mode and container mode because container mode data with entry size 1 is mathematically equivalent to line mode data. Hence, throughout these analysis, entry size 1 means line mode and entry size greater than 1 implies container mode. Secondly, we assume that each item has a fixed probability of interception. Therefore, we consider each line a Bernoulli trial, which only depends on the item type. When the entry size is greater than one, the relevant probability is whether at least one line was intercepted.

We estimate precision via calculation the Fisher information matrix, $\mathcal{I}$. The Fisher information matrix is the expected value of the negative of the Hessian matrix of the log-likelihood evaluated at the maximum likelihood estimate. The standard error estimates are the square roots of the diagonal elements of the inverse of $\mathcal{I}$.

2.1 Single item type

For the single item type case, we set the probability of interception to be p, and define N as the total number of entries, I as the number of entries intercepted and S as the size (i.e. the number of lines) in each entry. The likelihood is a binomial distribution, where the outcome is the an entry being intercepted. The probability that an entry is not intercepted is

$$\mathbb{P}(\text{entry not intercepted}) = (1-p)^S, \tag{1}$$

meaning that the binomial likelihood for a set entry size, S, is

$$\mathcal{L}_S = \left(1-(1-p)^S\right)^I \left((1-p)^S\right)^{N-I}. \tag{2}$$

To generalise Eq. (2) to arbitrary entry sizes, we need the product over entry size:

$$\mathcal{L} = \prod_{S\in\mathbb{N}} \left(1-(1-p)^S\right)^{I_{E,S}} \left((1-p)^S\right)^{N_{E,S}-I_{E,S}}, \tag{3}$$

where S is the entry size, $I_{E,S}$ is the number of inteceptions of entry size S and $N_{E,S}$ is the number of entries of size S. The log-likelihood is

$$\log\mathcal{L} = \sum_{S\in\mathbb{N}} I_{E,S}\log\left(1-(1-p)^S\right) + (N_{E,S}-I_{E,S})\log\left((1-p)^S\right). \tag{4}$$

As there is only one parameter, we calculate its second derivative rather than there being a Hessian matrix:

$$\left[\frac{\partial^2\log\mathcal{L}}{\partial p^2}\right] = \sum_{S\in\mathbb{N}} \frac{S\left(N_{E,S} + \frac{I_{E,S}((1+S)(1-p)^S-1)}{((1-p)^S-1)^2}\right)}{(1-p)^2}. \tag{5}$$

To calculate the Fisher information, we need the expected value of the number of interceptions, which depends on the size of the entry:

$$\mathbb{E}\left[I_{E,S}\right] = N_{E,S}(1-(1-p)^S). \tag{6}$$

Hence the Fisher information is

$$\mathcal{I} = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}}{\partial p^2}\right] = -\sum_{S\in\mathbb{N}} \frac{N_{E,S}S^2(1-p)^{S-2}}{(1-p)^S - 1}, \tag{7}$$

and the standard error estimate is

$$SE = \sqrt{\frac{1}{-\sum_{S\in\mathbb{N}} \frac{N_{E,S}S^2(1-p)^{S-2}}{(1-p)^S-1}}}. \tag{8}$$

Using Eq. (8) we can understand how the probability of interception, the entry size and the number of lines affect the standard error, and we plot these relationships in Figure 1. The left plot shows that the standard error depends on the probability of interception and that the relationship depends on the entry size. For all entry sizes, the standard error is small when the probability of interception is small (below ~ 0.3). However, for larger values of the probability of interception, the standard error increase significantly if the entry size is three or greater. The large p behaviour is driven by the $(1-p)^{S-2}$ term in Eq. (8), which means SE goes to 0 if $S = 1$, while it diverges if $S \geq 3$. Figure 1 also shows how the standard error decreases as the number of lines of data increases. The lower the entry size is, the lower the standard error, and, as $SE \sim \sqrt{1/N_{E,S}}$, container mode data with larger entry sizes require a large amount of data to reach the same standard error.

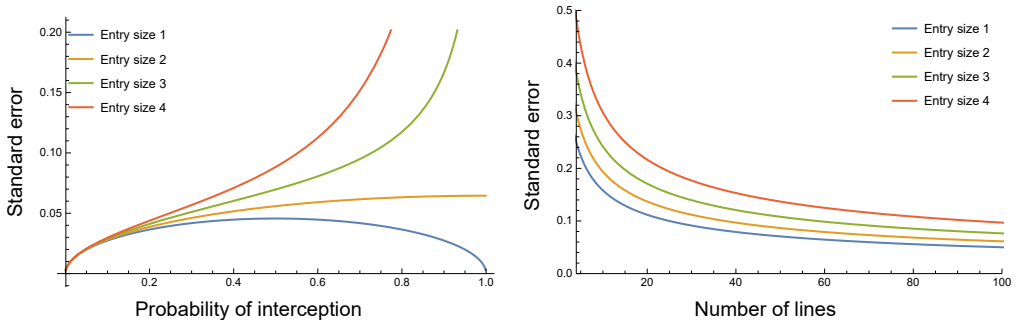

Figure 1: The standard error (Eq. (8)) as the probability of interception, p, is varied (left) and as the number of lines are varied (right). For the left plot the number of lines is held constant at 120. For the right plot, the probability of interception is held constant at 0.5

2.2 Two item types

In this section we consider a situation where there are two items with probabilities of interception of p_1 and p_2. We focus on a scenario where every entry is of size two, meaning there are three types of entries: only type 1; only type 2; or mixed, with one line of type 1 and one of type 2. We denote the number of lines within a single entry of type 1 and 2 as S_1 and S_2 respectively, and $I(S_1, S_2$ and $N(S_1, S_2)$ are the number of interceptions and total number of entries with S_1 type 1 lines and S_2 type 2 lines. For

our case, we can have $S_1 = 2, S_2 = 0$, $S_1 = 1, S_2 = 1$ or $S_1 = 0, S_2 = 2$. Rewriting the log-likelihood from Eq. (4), we get

$$\log \mathcal{L} = \sum_{S_1,S_2} I(S_1,S_2) \log\left(1-(1-p_1)^{S_1}(1-p_2)^{S_2}\right) + (N(S_1,S_2) - I(S_1,S_2)) \log\left((1-p_1)^{S_1}(1-p_2)^{S_2}\right). \tag{9}$$

We compute the Fisher information matrix and the standard error using Mathematica. The standard error for p_1 is

$$\frac{1}{2}\sqrt{-\frac{p_1\left(p_1^2-3p_1+2\right)\left(N_{1,2}(p_1-1)(p_2-2)p_2+4N_{2,2}(p_2-1)(p_1(p_2-1)-p_2)\right)}{N_{1,1}(p_1-1)(N_{1,2}(p_1-1)(p_2-2)p_2+4N_{2,2}(p_2-1)(p_1(p_2-1)-p_2))+N_{1,2}N_{2,2}(p_1-2)p_1(p_2-1)^2}}, \tag{10}$$

where $N_{1,1}$ and $N_{2,2}$ are the number of entries with two type 1 lines and two type 2 lines respectively, while $N_{1,2}$ are the number of entries with both type 1 and type 2 lines. The standard error for p_2 is the same, with $N_{1,1}$ and $N_{2,2}$ switched and p_1 and p_2 switched.

By examining Eq. (10) we can see that the behavior of the standard errors is more complex than when we considered only one type of item. Notably, the number of entries of only type 2, $N_{2,2}$, is in the equation for the type 1 standard error, along with the probability of interception of type 2, p_2.

Figure 2 shows how the standard error varies with the proportion of mixed entries, for different interception probabilities. In each case, we hold p_1 at 0.1, and we vary p_2 for 0.05 up to 0.9 across the four plots. Interestingly, how the standard error changes as a function of the proportion of mixed entries changes markedly, depending on the value of p_2. In particular, when p_2 is 0.7 and 0.9, the standard error for the p_1 increases as the proportion of mixed entries increases, while the standard error for p_2 actually decreases, while the proportion of mixed entries is below $\sim 90\%$.

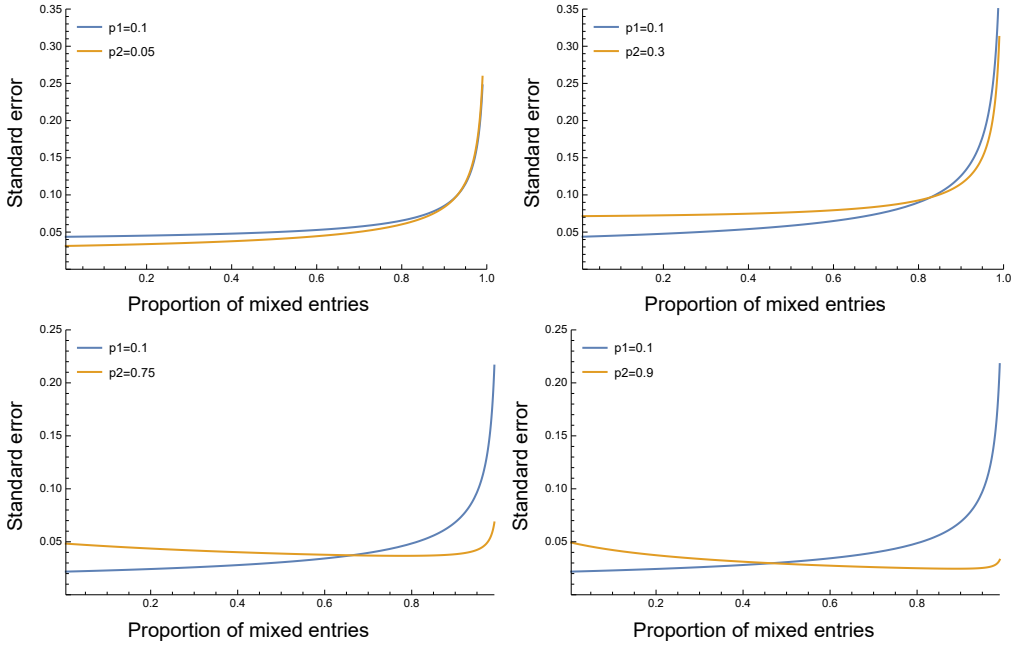

Figure 2: The standard error when for entry size 2, when the proportion of mixed entries is varied. In each plot there are 50 total entries and p_1 is held constant at 0.1, while p_2 goes from 0.05 up to 0.9.

Across the plots in Figure 2, the behavior of the standard error for p_1 looks quite similar. However, comparing the standard error for p_1 for different values of p_2 shows that the value of p_2 has a measurable impact on the standard error of the p_1 estimate (Figure 3). Here we see that the standard error for p_1 increases as the value of p_2 increases. Combing this result with the results in Figure 2, we see that if low and high probability items are mixed in container mode, we can estimate the risk of the high probability items, but it comes at the detriment for our ability to identify low-risk items.

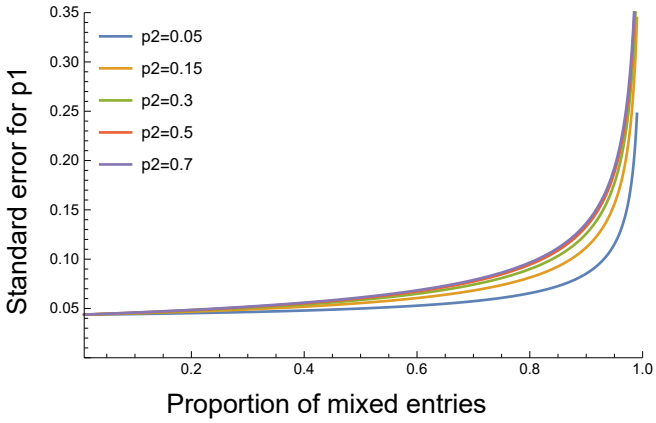

Figure 3: The stand error for p_1 as the proportion of mixed entries changes for varying p_2. The number of total entries is held constant at 50.

2.3 Asymptotic analysis conclusions

From the asymptotic analysis we can conclude two key lessons:

1. Container mode data is most useful when the probability of interception, p, is small. In the extreme case where $p \to 0$, container mode data is equivalent to line mode data because whenever an entry is *not* intercepted, we know every line was not intercepted, regardless of whether it is container mode or line mode.

2. When p is not small (e.g. greater than 0.5) it can still be useful to analyse container mode data, but only if the amount of line mode data is small. However, if there is sufficient line mode data to get acceptable precise parameter estimates, there would be little to gain by using container mode data.

3. Entries in container mode with mixed entries can increase or decrease the precision of estimates, depending on the true probability of interception. If the probability of interception is low, mixed entries tend to lead to lower precision, but item types with high probability of interception may lead to lower standard errors.

3 Simulation study

As the first step in developing a method to analyse real-world data sets, we simulate a data set that contains key complexities including multiple item types, fixed effects and random effects. Once we have simulated data, we can fit a model to estimate parameters. The advantage of our approach is that we can (1) verify that our model behaves correctly and (2) explore how model precision varies in a more complex setting.

3.1 Data simulation

In our simulation model, we use a logistic model to generate the probability p_i that a line is intercepted. We assume that each item type has its own baseline fixed effect, which is $\alpha_{\text{type}[i]}$ and that there is a fixed effect of having compliant documentation, β. Finally we include a random effect which is the entry that line i is in. The full equation is:

$$\text{logit}(p_i) = \alpha_{\text{type}[i]} + \beta \times \text{document}[i] + \text{entry}[i]. \tag{11}$$

For our simulations, we set $\beta = -1$ and we include 8 item types with α values of -4.60 -2.94 -2.20, -1.39, 0, 0.85, 2.20, and 2.94 (corresponding to interception probabilities of 0.01, 0.05, 0.1, 0.2, 0.5, 0.7, 0.9 and 0.95). We run simulations with and without an entry effect, but when we include entry

$$\text{entry}[i] = \text{Normal}(0, \sigma), \tag{12}$$

where $\sigma = 0.25$. For each entry we draw whether it is in line mode or not with probability 0.25, and for each line we set document to be compliant with probability 0.2. We show an example of the format of the data in Table 1.

Table 1: Example simulated data. Lines 1-3 are all marked as intercepted because entry 1 is in container mode, even though only 1 or 2 of the lines were actually intercepted

Line	Entry	Type	Document	Entry correlation	Mode	Container	Intercepted
1	1	3	1	0.023	Container	1	1
2	1	2	0	0.023	Container	1	1
3	1	1	0	0.023	Container	1	1
4	2	5	0	0.08	Line	2	0
5	3	8	1	-0.37	Line	2	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

3.2 STAN model

We use the RSTAN package to model the system. The basic structure for fitting the simulated data is to define the probability of interception for each line (following Eq. (11))

$$\text{logit}(p_i) = \alpha_{\text{type}[i]} + \beta \times \text{document}[i] + \text{entry}[i], \tag{13}$$

where

$$\text{entry}[i] \sim \text{Normal}(0, \sigma). \tag{14}$$

As we fit our model in a Bayesian framework using STAN, we set priors on each parameter:

$$\alpha_{\text{type}[i]} \sim \text{Normal}(0, 100) \tag{15}$$
$$\beta \sim \text{Uniform}(-2, 2) \tag{16}$$
$$\sigma \sim \text{Uniform}(0, 1). \tag{17}$$

If the data were exclusively, fitting the model would be relatively standard, as it is logistic regression. For the line mode data we can write

$$\text{Intercepted}_{i,\text{line}} \sim \text{Bernoulli}(p_i). \tag{18}$$

However, the container mode data complicates the model, because for each entry there is only one response variable (the interception), irrespective of how many lines are contained within the entry. Hence, as we did in the asymptotic analysis for the container data (Eq. (2)), we need to calculate the probability that at least one line in the entry is intercepted:

$$\mathbb{P}(\text{Entry } j \text{ intercepted}) = q_j = 1 - \prod_{i=\text{Lines}\in j} (1 - p_i). \tag{19}$$

3.3 Model verification

The primary aim of our work is to investigate the relative contribution of container mode data to precision of parameter estimates. However, before analysing precision it is important to ensure that we are able to accurately estimate parameters with our model and that precision estimates using STAN match the asymptotic analysis.

3.3.1 Model precision

To check model precision, we simulate a simplified dataset, with a single item type and no other fixed or random effects, and compare the STAN estimates with the asymptotic approximation, from Eq. (8). As STAN is a Bayesian approach, it is apropriate to comapare the standard error estiamte, Eq. (8), to the standard deviation of the STAN posterior draws. The simulation uses $p = 0.05$ and we vary the number of entries N_E and the entry size, S and the comparison is shown in Figure **??**.

References

A. Chatterjee and T. Bandyopadhyay. Regression models for group testing: Identifiability and asymptotics. *Journal of Statistical Planning and Inference*, 204:141–152, Jan. 2020. ISSN 0378-3758. doi: 10.1016/j.jspi.2019.05.003. URL `https://www.sciencedirect.com/science/article/pii/S0378375818301423`.

A. Delaigle and P. Hall. Nonparametric methods for group testing data, taking dilution into account. *Biometrika*, 102(4):871–887, Dec. 2015. ISSN 0006-3444. doi: 10.1093/biomet/asv049. URL `https://doi.org/10.1093/biomet/asv049`.

G. Hepworth and B. J. Biggerstaff. Bias Correction in Estimating Proportions by Pooled Testing. *Journal of Agricultural, Biological and Environmental Statistics*, 22(4):602–614, Dec. 2017. ISSN 1537-2693. doi: 10.1007/s13253-017-0297-2. URL `https://doi.org/10.1007/s13253-017-0297-2`.

G. Hepworth and B. J. Biggerstaff. Bias Correction in Estimating Proportions by Imperfect Pooled Testing. *Journal of Agricultural, Biological and Environmental Statistics*, 26(1):90–104, Mar. 2021. ISSN 1537-2693. doi: 10.1007/s13253-020-00411-5. URL `https://doi.org/10.1007/s13253-020-00411-5`.

G. Hepworth and R. Watson. Debiased estimation of proportions in group testing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(1):105–121, 2009. ISSN 1467-9876. doi: 10.1111/j.1467-9876.2008.00639.x. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2008.00639.x`. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9876.2008.00639.x.

Y. Liu, C. S. McMahan, J. M. Tebbs, C. M. Gallagher, and C. R. Bilder. Generalized additive regression for group testing data. *Biostatistics*, (kxaa003), Feb. 2020. ISSN 1465-4644. doi: 10.1093/biostatistics/kxaa003. URL https://doi.org/10.1093/biostatistics/kxaa003.

C. S. McMahan, J. M. Tebbs, T. E. Hanson, and C. R. Bilder. Bayesian regression for group testing data. *Biometrics*, 73(4):1443–1452, 2017. ISSN 1541-0420. doi: 10.1111/biom.12704. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12704. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12704.

F. Schaarschmidt. Experimental design for one-sided confidence intervals or hypothesis tests in binomial group testing. *Communications in Biometry and Crop Science*, 2007. ISSN 1896-0782. URL http://agrobiol.sggw.waw.pl/~cbcs/pobierz.php?plik=CBCS_2_1_5.pdf. Publisher: Faculty of Agriculture and Biology, Warsaw Agricultural University, Poland.

K. H. Thompson. Estimation of the Proportion of Vectors in a Natural Population of Insects. *Biometrics*, 18(4):568–578, 1962. ISSN 0006-341X. doi: 10.2307/2527902. URL https://www.jstor.org/stable/2527902. Publisher: [Wiley, International Biometric Society].