Entry and line mode

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

January 27, 2022

Abstract

In the Australian border biosecurity system, data about shipping containers is recorded in one of two modes: *entry* 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. Analysis is simple when the data is recorded in line mode: the directions are recorded individually for each line. The challenge comes when data is recorded in entry mode, because the same direction is recorded each line in the entry. In other words, if at least one line in an entry is non-compliant, then all lines in that entry are recorded as non-compliant. Therefore, entry mode data creates a challenge when we try to estimate the probability that certain items are non-compliant, because we do not know which records of non-compliance match up with which line. We develop a statistical model to use entry mode data to help inform biosecurity risk of items. We use asymptotic analysis to estimate the value of entry 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 apply our methods to a real dataset.

1 Introduction

Biosecurity is important.

Border biosecurity is important - pathways [Consider including the following here: In this paper, we assume that the outcome is either compliant or non-compliant. Describing non-compliance.]

Strategic analysis requires interception data.

Cargo (volumes, activity, etc)

Data capture - line and entry mode Data recording compliance of cargo entering Australia is recorded in two main formats: line mode and entry mode. Entry mode is often referred to as container mode in biosecurity practice, but we will only use the

term entry mode in this paper for simplicity. A line is a group of the same type of item being imported. An entry is a group of lines. When cargo enters the country, each line is given *directions*, which consist of a decision whether to check the line for compliance and the outcome after checking. In line mode, the directions assigned to each line are recorded along with the outcome for that line. In entry mode, directions are only recorded per entry. This means that in entry mode, if any line in that entry is found to be non-compliant, every line in that entry is recorded to be non-compliant. If all of the lines in the entry are compliant, then they are all marked compliant, so in this case line and entry mode are equivalent.

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

In this paper we investigate how entry mode affects our ability to identify the biosecurity risk of items. We start with an asymptotic analysis, where we calculate the precision of estimates the implications of mixing different item types in entry mode. We then do a simulation estimation study, which allows us to understand how larger entries and more item types affect precision of estimates. Finally, we analyse a subset of Australian border interception data to see the real-world differences between using the line-only data and including the entry mode data.

2 Model overview

Throughout this paper, we focus on estimating the probability that a line is non-compliant, using information about it, including the type of item, country of origin and whether it has complete documentation. The full model for the probability that a

line is non-compliant, $p_{ijk\ell}$, is:

$$\text{logit}(p_{ijk\ell}) = \alpha_i + \beta\mathbb{I}_j + \delta_k + \gamma_\ell, \tag{1}$$

where the fixed effects are α_i, $\beta\mathbb{I}_j$ and δ_k: α_i represents the item type, the indicator variable $\mathbb{I}_j$ denotes whether there is correct documentation, the coefficient β is the weight given when there is correct documentation, and δ_k represents the country of provenance. The random effect γ_ℓ represents the entry effect. The indices can take values

$$i = 1, \ldots, a, \qquad a \in \mathbb{N}, \qquad a = \#\text{ items} \tag{2}$$

$$j = 1, 2, \qquad \text{without and with documentation} \tag{3}$$

$$k = 1, \ldots, d, \qquad d \in \mathbb{N}, \qquad d = \#\text{ countries} \tag{4}$$

$$\ell = 1, \ldots, g, \qquad g \in \mathbb{N}, \qquad g = \#\text{ entries.} \tag{5}$$

The values of the indicator variable are

$$\mathbb{I}_1 = 0, \qquad \text{without documentation} \tag{6}$$

$$\mathbb{I}_2 = 1, \qquad \text{with documentation.} \tag{7}$$

The random effect γ_ℓ has distribution

$$\gamma_\ell|\sigma \sim \text{Normal}(0, \sigma), \qquad \ell = 1, \ldots, g, \qquad g \in \mathbb{N}. \tag{8}$$

If all data were in line mode, the above model would be a fairly standard mixed effects logistic regression with categorical variables. However, because of entry mode, we don't observe outcomes for each line, as every line in the entry is marked as non-compliant if any line in the entry is found to be non-compliant. Therefore, the outcome is whether the entry is compliant and we need to calculate the probability that the entry is non-compliant, which is 1 minus the probability that every line in the entry is compliant:

$$\text{Entry non-compliant} = 1 - \prod[1 - p]. \tag{9}$$

Hence, for entries in entry mode, we treat the entry mode as a Bernoulli random variable with probability defined by Eq. (9), while for entries in line mode, we treat each line as a Bernoulli random variable with probability as defined in Eq. (1).

This paper includes three analyses: an asymptotic analysis, a simulation study, and a case study of Australian biosecurity data. For the asymptotic analysis we only consider different item types, so we rather than using Eq. (21), we just consider the probability that a line of item type i is non-compliant, p_i. The simulation study and the case study both use the full model, as defined above.

3 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 quantify how the amount of data, probability of interception and entry size affects precision. The second part assumes that there are two different item types, and it explores how changing the proportion of entries with both item types mixed affects precision.

Throughout this section we make two simplifications. Firstly, we do not separate line mode and entry mode because entry 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 entry 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 value of the parameter. We use the standard error estimate as our measurement of precision, which are the square roots of the diagonal elements of the inverse of $\mathcal{I}$.

3.1 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

$$\mathbb{P}(\text{entry not intercepted}) = (1-p)^S, \tag{10}$$

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

$$\mathcal{L}_S = \left(1-(1-p)^S\right)^I \left(1-p)^S\right)^{N-I}. \tag{11}$$

To generalise Eq. (11) to arbitrary entry sizes, we need the product over entry size:

$$\mathcal{L} = \prod_{S\in\mathbb{N}} \left(1-(1-p)^S\right)^{I_{E,S}} \left((1-p)^S\right)^{N_{E,S}-I_{E,S}}, \tag{12}$$

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

$$\log\mathcal{L} = \sum_{S\in\mathbb{N}} I_{E,S}\log\left(1-(1-p)^S\right) + (N_{E,S}-I_{E,S})\log\left((1-p)^S\right). \tag{13}$$

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

$$\left[\frac{\partial^2\log\mathcal{L}}{\partial p^2}\right] = \sum_{S\in\mathbb{N}} \frac{S\left(N_{E,S} + \frac{I_{E,S}((1+S)(1-p)^S-1)}{((1-p)^S-1)^2}\right)}{(1-p)^2}. \tag{14}$$

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

$$\mathbb{E}\left[I_{E,S}\right] = N_{E,S}(1-(1-p)^S). \tag{15}$$

Hence the Fisher information is

$$\mathcal{I} = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}}{\partial p^2}\right] = -\sum_{S\in\mathbb{N}} \frac{N_{E,S}S^2(1-p)^{S-2}}{(1-p)^S - 1}, \tag{16}$$

and the standard error estimate is

$$SE = \sqrt{\frac{1}{-\sum_{S\in\mathbb{N}} \frac{N_{E,S}S^2(1-p)^{S-2}}{(1-p)^S-1}}}. \tag{17}$$

Using Eq. (17) 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. (17), 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 $SE \sim \sqrt{1/N_{E,S}}$, entry mode data with larger entry sizes require a large amount of data to reach the same standard error.

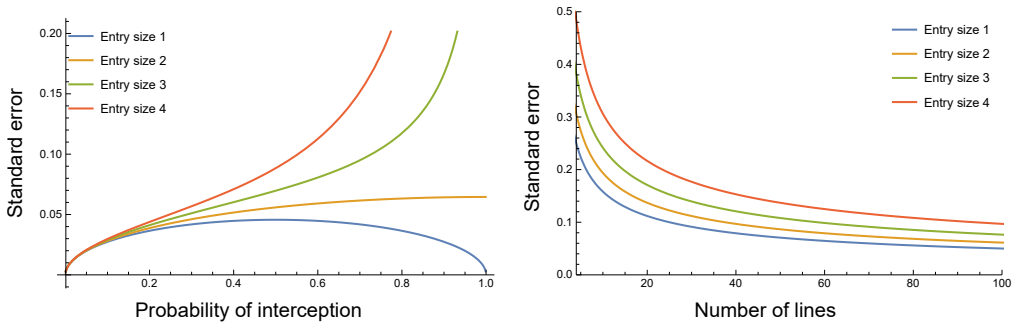

Figure 1: The standard error (Eq. (17)) as the probability of interception, p, is varied (left) and as the number of lines are varied (right). For the left plot the number of lines is held constant at 120. For the right plot, the probability of interception is held constant at 0.5

3.2 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, and we examine how these different items type probabilities

interact. 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. (13), we get

$$\log \mathcal{L} = \sum_{S_1,S_2} I(S_1, S_2) \log \left(1 - (1 - p_1)^{S_1}(1 - p_2)^{S_2}\right) + (N(S_1, S_2) - I(S_1, S_2)) \log \left((1 - p_1)^{S_1}(1 - p_2)^{S_2}\right). \tag{18}$$

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

$$\frac{1}{2}\sqrt{-\frac{p_1\left(p_1^2 - 3p_1 + 2\right)\left(N_{1,2}(p_1 - 1)(p_2 - 2)p_2 + 4N_{2,2}(p_2 - 1)(p_1(p_2 - 1) - p_2)\right)}{N_{1,1}(p_1 - 1)(N_{1,2}(p_1 - 1)(p_2 - 2)p_2 + 4N_{2,2}(p_2 - 1)(p_1(p_2 - 1) - p_2)) + N_{1,2}N_{2,2}(p_1 - 2)p_1(p_2 - 1)^2}}, \tag{19}$$

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. (20) 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 $\sim 90\%$.

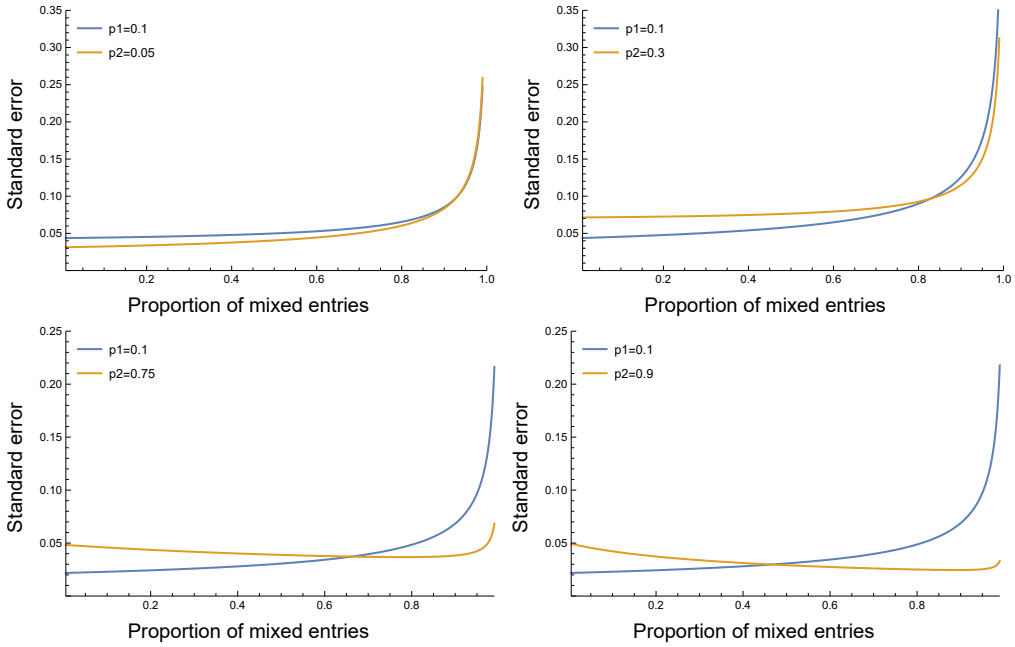

Figure 2: The standard error when for entry size 2, when the proportion of mixed entries is varied. In each plot there are 50 total entries and p_1 is held constant at 0.1, while p_2 goes from 0.05 up to 0.9.

3.3 Asymptotic analysis conclusions

From the asymptotic analysis we can conclude two key lessons:

1. Entry mode data is most useful when the probability of interception, p, is small. In the extreme case where $p \to 0$, entry 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 entry mode or line mode.

2. When p is not small (e.g. greater than 0.5) it can still be useful to analyse entry 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 entry mode data.

3. Entries in entry 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.

4 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.

4.1 Data simulation

We simulate data from the full model (as described in Section 2). For our simulations, chose parameters which we expect to be similar to their values in real-world data. We set $\beta = -1$, choosing an negative effect of having correct documentation, because we expect lines with correct documentation have a higher chance of being compliant. We include $a = 5$ item types with α_i taking values of -6.91, -4.60, -3.89, -2.94, -1.39 (corresponding to interception probabilities of 0.001, 0.01, 0.02, 0.05 and .2, if all other effects were 0). We chose negative values for the item effect α_i, since the probability of detection is expected to be low for any item in real-world data. We used $d = 3$ countries and choose their weights to be .5, -1 and .25, because... We run simulations with and without an entry effect, but when we include entry (8) it takes a normal distribution with mean zero and standard deviation $\sigma = 0.25$. For each entry we draw whether it is in line mode or not with probability 0.25, and for each line we set the probability of having correct documentation to be 0.2.

Do we need to specify how we generated values for countries and entries?

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 entry mode, even though only 1 or 2 of the lines were actually intercepted.

Line	Entry	Type	Documentation	Mode	Intercepted
1	1	3	1	Entry	1
2	1	2	0	Entry	1
3	1	1	0	Entry	1
4	2	5	0	Line	0
5	3	4	1	Line	1
⋮	⋮	⋮	⋮	⋮	⋮

4.2 STAN model

We model the system in a Bayesian framework, using the RSTAN package in R. The basic structure for fitting the simulated data is to define the probability of interception for each line (following Eq. (1))

$$\text{logit}(p_{ijk\ell}) = \alpha_i + \beta \mathbb{I}_j + \delta_k + \gamma_\ell, \tag{20}$$

where

$$\gamma_\ell|\sigma \sim \text{Normal}(0, \sigma), \qquad \ell = 1, \ldots, g, \qquad g \in \mathbb{N}. \tag{21}$$

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

$$\alpha_i \sim \text{Normal}(-4, 4), \qquad i = 1, \ldots, a, \qquad a \in \mathbb{N}, \tag{22}$$

$$\beta \sim \text{Normal}(0, 0.5), \tag{23}$$

$$\delta_k \sim \text{Normal}(0, 0.5), \qquad k = 1, \ldots, d, \qquad d \in \mathbb{N} \tag{24}$$

$$\sigma \sim \text{Normal}(0, 0.5) \tag{25}$$

The probability $p_{ijk\ell}$ is the probability of non-compliance of a line that has characteristics (i, j, k, ℓ). We denote the probability that line n is non-compliant as $p_{(n)}$, where $p_{(n)} = p_{ijk\ell}$ if that line has characteristics (i, j, k, ℓ).

4.3 Simulation results

5 Case study

We use border interception data from Australia to compare the diff

Table 2: Parameter estimates that are only possible when using all data, as opposed to the line-only fits

	param_name	week	summary_mean_all	summary_sd_all	summary_rhat_all
1	p_intercept[72]	1.00	-1.70	1.68	1.00
2	p_intercept[59]	2.00	-1.34	1.71	1.01
3	p_intercept[60]	2.00	-2.07	1.63	1.04
4	p_intercept[61]	2.00	-1.18	1.92	1.00
5	country_effect[62]	2.00	-0.01	0.49	1.00
6	country_effect[63]	2.00	-0.01	0.47	1.00
7	country_effect[64]	2.00	-0.01	0.49	1.01
8	country_effect[65]	2.00	0.02	0.48	1.02
9	country_effect[66]	2.00	0.01	0.48	1.00
10	country_effect[67]	2.00	-0.01	0.48	1.00
11	p_intercept[72]	3.00	-1.43	1.78	1.00

6 Appendix

References

A. Chatterjee and T. Bandyopadhyay. Regression models for group testing: Identifiability and asymptotics. *Journal of Statistical Planning and Inference*, 204:141–

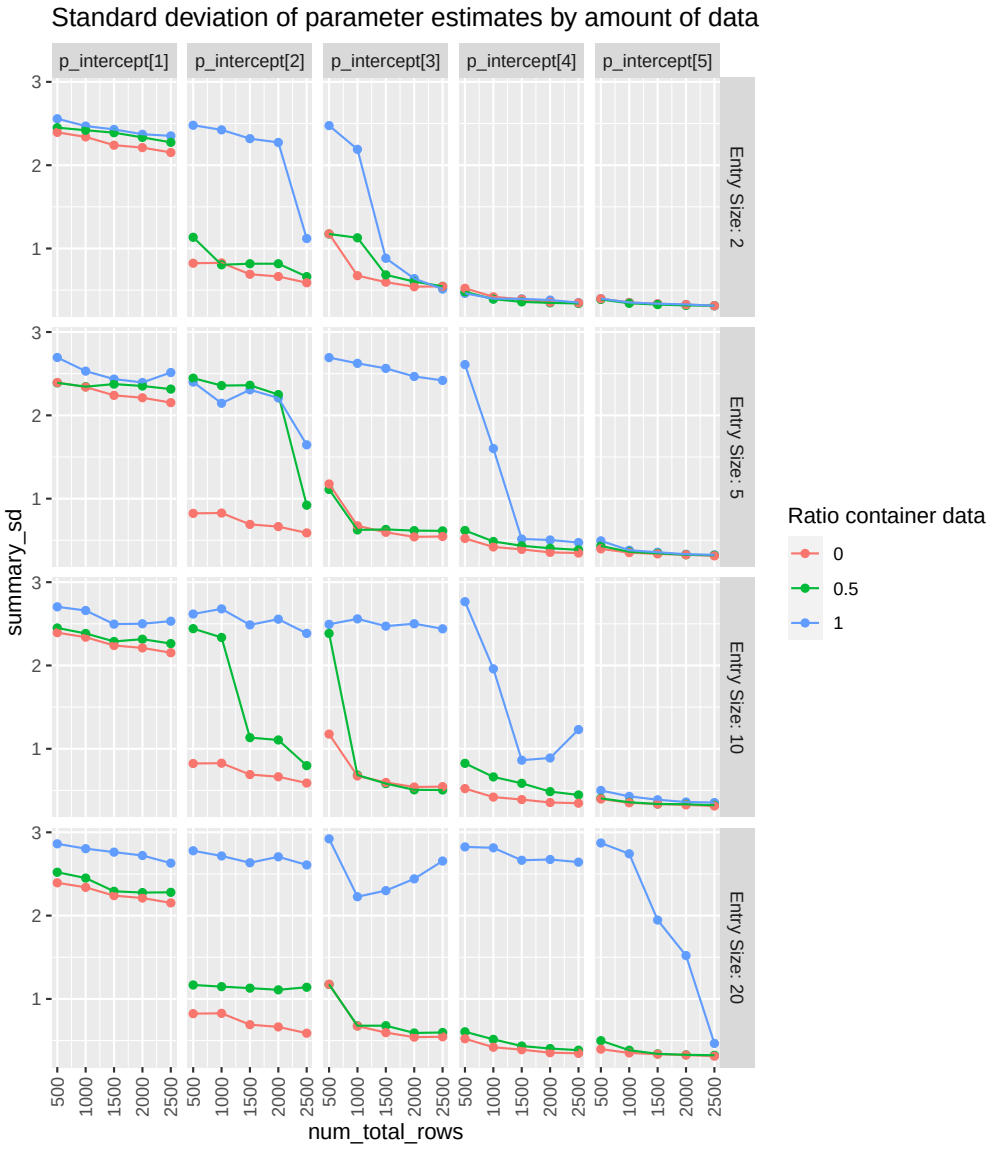

Figure 3: SD as amount of data increases, depending on p and the amount of entry data

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.

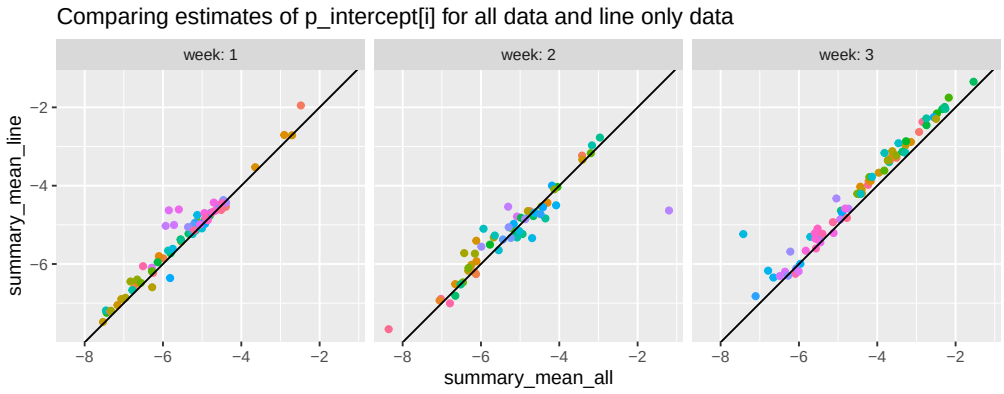

Figure 4: SD as amount of data increases, depending on p and the amount of entry data

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=`

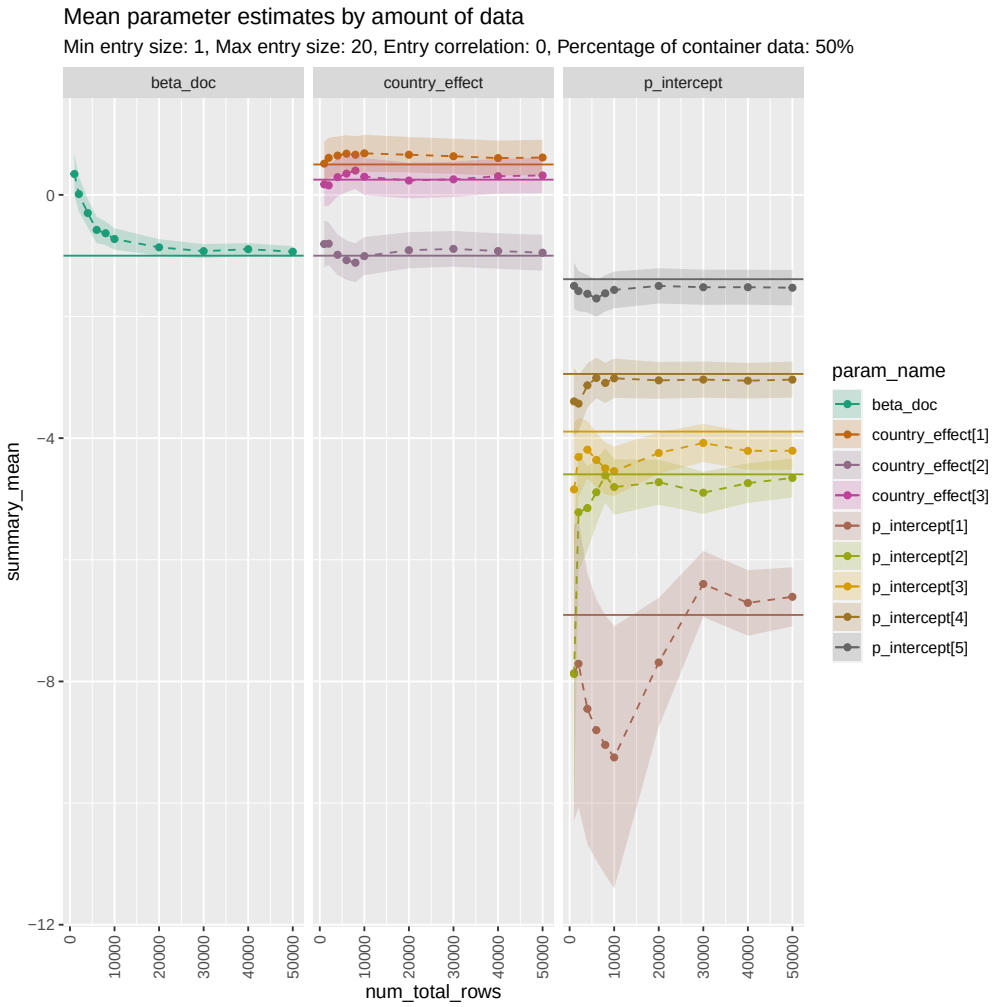

Figure 5: SUPPLEMENT FIGURE

`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].

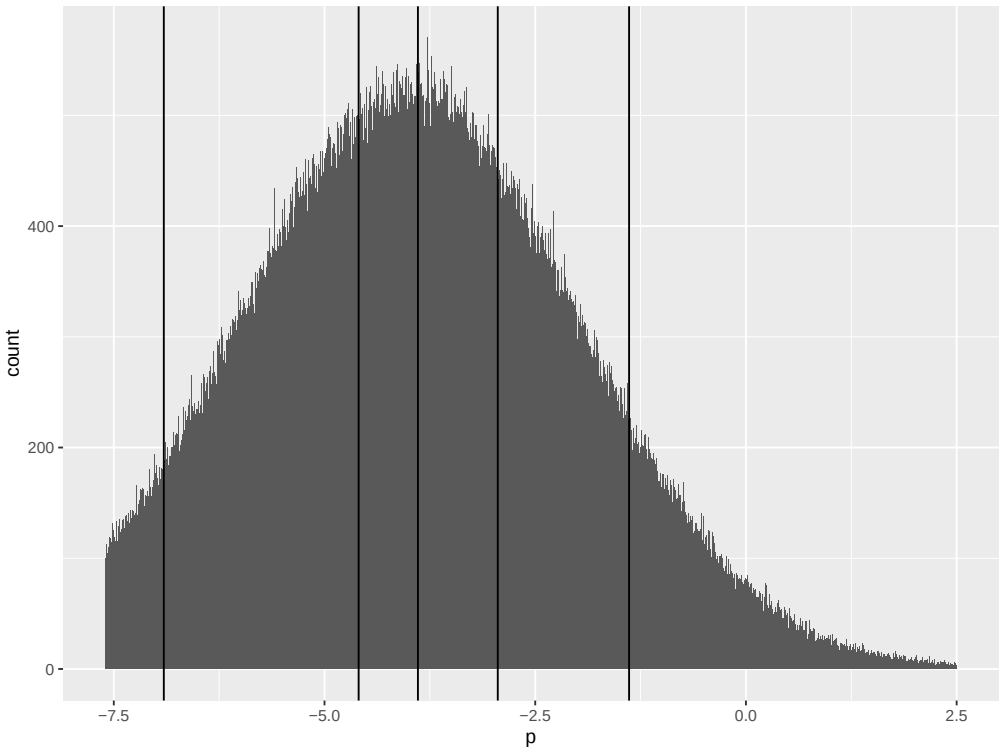

Figure 6: SUPPLEMENT FIGURE