Analysis of Multi-Laboratory Validation Experiments using Logistic Mixed Effects Models

Daniel Grant

July 27, 2021

Master of Mathematics Research Essay
Department of Statistics and Actuarial Science
University of Waterloo

Abstract

Binary-outcome tests may be used to detect the presence of a substance of interest within a larger sample. As the concentration of the substance of interest within the larger sample varies, so too does the ability of a test to reliably detect the substance. This relationship between concentration and test outcome is often characterized through the use of a collaborative experiment whereby testing of samples at a number of different concentrations is performed at numerous laboratories. Through these studies, conclusions may be made about the range of concentrations on which the test may be applied to achieve reliable results, and also how individual lab circumstances may affect this range. ISO/TS 16393 (2019) provides guidelines for the allocation of samples and labs for such multi-laboratory studies. This essay will analyse the effect of these guidelines with an alternative method of analysis that is based on mixed effects logistic regression. Application of this method to real data is provided, and a simulation study examining the effects of ISO/TS 16393 guidelines on the proposed model under a known data generation mechanism will be investigated.

1 Introduction

The use of a binary ("qualitative") test to determine the presence (or lack thereof) of a substance of interest in a sample can be seen in a variety of disciplines (Scherf et al., 2016; Yamamura & Sugimoto, 1995; Department of Defense, 2009). One well-known application of these tests occurs in food sciences, where tests may be performed on samples of food to check for allergens (peanuts, gluten, etc.) contained within food (Scherf et al., 2016; Hengel et al., 2006). Given that these tests return only positive or negative results, the quantification of the substance in a tested sample is not of interest, but rather whether the substance of interest is contained in the sample at relevant levels. These binary tests are often subject to uncertainty with regard to the concentration of a substance they are able to detect. This uncertainty may be magnified when testing is performed in the real world, as testing laboratories may perform testing procedures under conditions that differ from each other. This uncertainty surrounding a test's result may be described through the probability of detection (POD_i), the probability that a sample returns a positive test result for a given concentration of the substance of interest at a given lab i,

$$\text{POD}_i(x) = P(\text{Positive test}|\text{Concentration} = x, \text{Lab} = i). \tag{1}$$

In this way, we assume that the probability that a test returns positive forms a curve, a continuous function of the concentration of the sample being tested. Under these tests, very low concentrations may not be detected unless a test is repeated numerous times at a given lab and, conversely, sufficiently high concentrations will almost always result in a positive test regardless of the testing laboratory. In using this measure, one may wish to assess how POD_i curves differ by each lab i (as seen in Scherf et al. (2016)), or compare shapes of POD_i curves for competing testing methods (as seen in Hengel et al. (2006)).

In discussing how likely a random test is to return positive in general, it is possible to consider all laboratories in order to construct a mean POD curve (denoted "LPOD

curve" in ISO/TS 16393 (2019)). This measure seeks to make statements about the a probability of detection at a given concentration in an "average" sense. What "average" means may vary by author. Authors, including those of ISO/TS 16393, use the definition

$$\text{LPOD}(x) = P(\text{Positive test}|\text{Concentration} = x) \tag{2}$$

which is marginalized over the effect of labs. Another approach seen in Scherf et al. (2016) is to condition on some sense of an "average lab", resulting in a probability of detection that is not marginalized across labs, but is a claim about what POD the average lab may have, i.e.

$$\text{LPOD}(x) = P(\text{Positive test}|\text{Concentration} = x, \text{Lab} = \text{"Average"}). \tag{3}$$

The choice of using (2) or (3) depends on the conclusions one wishes to make through the analysis.

In either situation, it may be of interest to quantify the minimum concentration required to return a positive result most (e.g. 95%) of the time. This concentration, sometimes called "$\text{LOD}_{95\%}$" as per Uhlig (2015), is a concentration $x_{95\%}$ such that

$$\text{LPOD}(x_{95\%}) = 0.95. \tag{4}$$

1.1 ISO/TS 16393

ISO/TS 16393 (2019) seeks to standardize how to reliably assess a given testing method or compare multiple methods by providing guidelines for the design of tests that should be performed by laboratories. In this standard, samples of measurand are to be collected with at least 5 fixed and known concentrations of the substance of interest. Each concentration is then split into a minimum of 96 identical subsamples, which are then divided among at least 8 laboratories, so that each laboratory will perform at least 12 replicate tests.

After performing these tests, one of the main goals of analysis in ISO/TS 16393 is to characterize the mean POD (LPOD) curve. ISO/TS 16393 prescribes that tests done at each concentration be analysed independently, where the goal at each concentration studied is to estimate a single point along the theoretical LPOD curve, along with a confidence interval for this LPOD estimate (see Annex A-D of ISO/TS 16393 (2019)). Using these methods of analysis, data from other concentrations tested do not contribute any information to the analysis at given concentration. One problem with this approach is that it does not characterize the LPOD curve explicitly, but rather estimates a discrete number of points along it. In this way, inference about LPOD or POD_i values at untested concentrations, in particular those such as (4), may be unreliable as there exists no precise way to connect studied points to form an LPOD curve. Additionally, if concentrations are chosen poorly (too extreme to give meaningful information about the shape of individual POD_i curves), it may be impossible to fit the proposed models at a given concentration (see Table 1 at concentration 47.1mg/kg and discussion in Section 2).

Scherf et al. (2016) eliminate many of the above concerns by considering all studied concentrations simultaneously in their attempt to characterize the full mean POD (LPOD) curve (defined as (3) instead of (2) seen in ISO/TS 16393 (2019)). Scherf et al. proposed a 4-parameter log-logistic curve with random lab effect, given as

$$\text{POD}_i(x) = \frac{A - D}{1 + (x/C\gamma_i)^B} + D. \tag{5}$$

Where A and D are parameters representing the minimum obtainable POD_i value (≥ 0) and the maximum obtainable POD_i value (≤ 1), respectively (Finney, 1976). These parameters represent inherent false negative and false positive rates of a testing method which are assumed to be present for all labs. The parameter C represents the inflection point of the POD_i curve, and B represents the slope of the curve at this inflection point (Finney, 1976). A, B, C, and D are assumed to be identical for all labs, with each POD_i curve differing only in the random lab effect γ_i. The random lab effect γ_i was assumed to

be such that

$$\log \gamma_i \sim N(0, \sigma^2).$$

In this case, the mean POD (LPOD) curve was given by setting the random effect to its median value of 1. This gave an LPOD curve of

$$\text{LPOD}(x) = \frac{A - D}{1 + (x/C)^B} + D. \tag{6}$$

Scherf et al. propose a prediction interval for the POD_i curves which seeks to make claims about the range on which a new lab's curves are likely to fall. This was constructed by replacing $C\gamma_i$ with a 95% prediction interval for this term in (5)(Scherf et al., 2016, p. 315-16). The details surrounding the construction of this interval were not provided.

Given the number of replicates and number of participating labs that were available for testing, questions arise as to whether such a curve may be reliably fitted. This problem of fitting will be evident even in a more parsimonious model and under more ideal sampling concentrations used in the simulation study of Section 3.

1.2 Mixed Effects Logistic Model

An alternative to (5) from Scherf et al. will be proposed in this essay. Like Scherf et al., it is possible to assume that the probability of detection may be described by specifying a continuous function of concentration (as opposed to the discrete analyses proposed by ISO/TS 16393). This proposed model is given as

$$\ln\left(\frac{\text{POD}_i(x)}{1 - \text{POD}_i(x)}\right) = (\beta_0 + b_{0,i}) + \beta_1 x \tag{7}$$

where

$$b_{0,i} \sim N\left(0, \sigma^2_{b_0}\right).$$

This model assumes that each lab has random intercept component that represents lab conditions that may systematically effect the probability of detection at a given lab. This model also benefits from only having 3 parameters $(\beta_0, \beta_1, \sigma^2_{b_0})$ to estimate, which is reduced from 5 (A, B, C, D, σ^2) as required by Scherf et al.. It should be noted that because it is assumed that all labs share a common slope, the change in log odds for a 1 unit increase in concentration will be the same for each lab's POD_i curve. This does not imply that changes in POD_i will be the same, however, due each lab's unique random effect $b_{0,i}$. This model is in contrast to a model that includes both a random slope and random intercept for each lab, which requires larger large sample sizes than the proposed model which may not be supported by some multi-laboratory studies. This random intercept model is also desirable due to its mathematical simplicity.

As one of the primary goals of an analysis is to estimate the concentration associated with $\text{POD}_i = 0.95$, it is possible to treat this particular concentration as a random quantity itself. For a given lab i this value is given as

$$x_{95\%} = \frac{\ln\left(\frac{0.95}{1-0.95}\right) - (\beta_0 + b_{0,i})}{\beta_1}. \tag{8}$$

Without conditioning on the lab, the concentration (8) is a random quantity which varies from lab to lab. Using the normality of $b_{0,i}$,

$$X_{95\%} \sim N\left(\frac{\ln\left(\frac{0.95}{1-0.95}\right) - \beta_0}{\beta_1}, \frac{\sigma^2_{b_0}}{\beta_1^2}\right). \tag{9}$$

Interest lies in the mean and variance of the 95^{th} percentile concentration, which will be denoted θ_1 and θ_2, respectively, so that

$$\theta_1 \equiv E(X_{95\%}) = \frac{\ln\left(\frac{0.95}{1-0.95}\right) - \beta_0}{\beta_1} \qquad \theta_2 \equiv \text{Var}(X_{95\%}) = \frac{\sigma^2_{b_0}}{\beta_1^2}. \tag{10}$$

1.2.1 Interval for realizations from $X_{95\%}$

It may also be of interest to estimate which range of concentrations will contain most individual lab 95^{th} percentile concentrations. In terms of drawing conclusions using a single sample, as would be the case using real data, it is possible to use the assumed normality of the random lab effect, along with θ_1 and θ_2 to construct an appropriate 95% confidence interval of the form

$$\hat{\theta}_1 \pm 1.96\sqrt{\hat{\theta}_2} \tag{11}$$

Particular interest should be taken in the upper bound of this interval, as it estimates an upper bound on the 95^{th} percentile concentrations across labs. In other words, such a concentration is the smallest concentration which will consistently yield a positive result for vast majority of labs; if the concentration in a sample is less than this value, it becomes less certain that a substance will be detected by a reasonably performing lab. It should be noted that (11) does not take into account the sampling uncertainty of the estimates of θ_1 or θ_2 or $\sigma^2_{b_0}$. By treating (11) as a function of β_0, β_1 and $\sigma^2_{b_0}$, a confidence interval about this upper bound is given

$$\left(\frac{\ln\left(\frac{0.95}{1-0.95}\right) - \hat{\beta}_0}{\hat{\beta}_1} + 1.96\sqrt{\frac{\hat{\sigma}^2_{b_0}}{\hat{\beta}^2_1}}\right) \pm 1.96\sqrt{\text{Var}\left(\frac{\ln\left(\frac{0.95}{1-0.95}\right) - \hat{\beta}_0}{\hat{\beta}_1} + 1.96\sqrt{\frac{\hat{\sigma}^2_{b_0}}{\hat{\beta}^2_1}}\right)}. \tag{12}$$

The delta method is used to approximate the variance in this interval, with calculations provided in Appendix A. It should be noted that this method will require modification in the case that $\hat{\sigma}^2_{b_0} = 0$. In such cases, it may be possible to derive the interval including only sampling variability of β_0 and β_1 estimates.

2 Application to Gluten Detection

In order to apply this method to a real data set, a data set referenced in ISO/TS 16393, an analysis of gluten in food samples by Scherf et al. (2016) was used.

The testing configuration used by Scherf et al. differs from that of ISO/TS 16393 in numerous key ways: The first is that the number of participating laboratories is 17, more than double the minimum of 8 recommended by ISO/TS 16393. Given that more participating laboratories allows for more accurate estimates of the lab effect in addition to increasing the total sample size, this has the potential of resulting in estimates of θ_1 and θ_2 closer to the true values than might be seen using the minimum guidelines of ISO/TS 16393.

A second key difference between this data set and the recommendations of ISO/TS 16393 relates to the sampled concentrations; only 4 concentrations were sampled in this example, rather than the minimum recommended 5. Furthermore, only 10 tests were performed at each lab, as opposed to the recommended minimum 12. This testing configuration results in 40 tests being performed per lab, as opposed to 60 under the minimum recommendations of ISO/TS 16393. Given that the number of labs is large, however, it is uncertain as to how the submininum tests per lab may affect the analyses using this data.

Perhaps the most important departure from ISO/TS 16393 minimum guidelines, however, is that the choices of concentrations at which to sample were poor. Concentrations chosen do not represent different points distributed about the POD curve as suggested by ISO/TS 16393, but rather represent solely the extreme low and extreme high values. Because of this, for nearly all labs, the lowest concentration tested resulted in zero detections, and the next highest concentration tested resulted in nearly 10 out of 10 concentrations, with the remaining two larger concentrations tested providing solely positive tests. In this way, for many labs the largest two concentrations tested did not provide much new information about the shape of the POD curve that was not already evident by

the second tested concentration. Furthermore, recommended analyses that first involve independent modelling for each laboratory individually, as Uhlig (2015), break down in these cases. This is because individual curves cannot be fit to each lab to make inference about slope or intercept parameters due to perfect separation occurring anywhere between the lowest two concentrations of 0.4mg/kg and 6.4mg/kg in some labs. These methods are in contrast to the method of Scherf et al., along with the mixed effects logistic model proposed in Section 1.2, since information from labs without perfect separation may be utilized in order to allow model fitting.

	Concentration (mg/kg)			
Lab	0.4	6.4	13.3	47.1
A	2	7	10	10
D	0	9	10	10
E	0	1	10	10
F	0	10	10	10
G	0	10	10	10
H	0	10	10	10
I	0	9	10	10
L	0	8	10	10
M	0	10	10	10
N	0	10	10	10
O	0	10	10	10
P	0	10	10	10
R	0	10	10	10
S	0	0	10	10
T	0	9	10	10
U	0	1	10	10
W	0	10	10	10

Table 1: Gluten detection data from Scherf et al.(2016). Note separation existing in e.g. Lab M

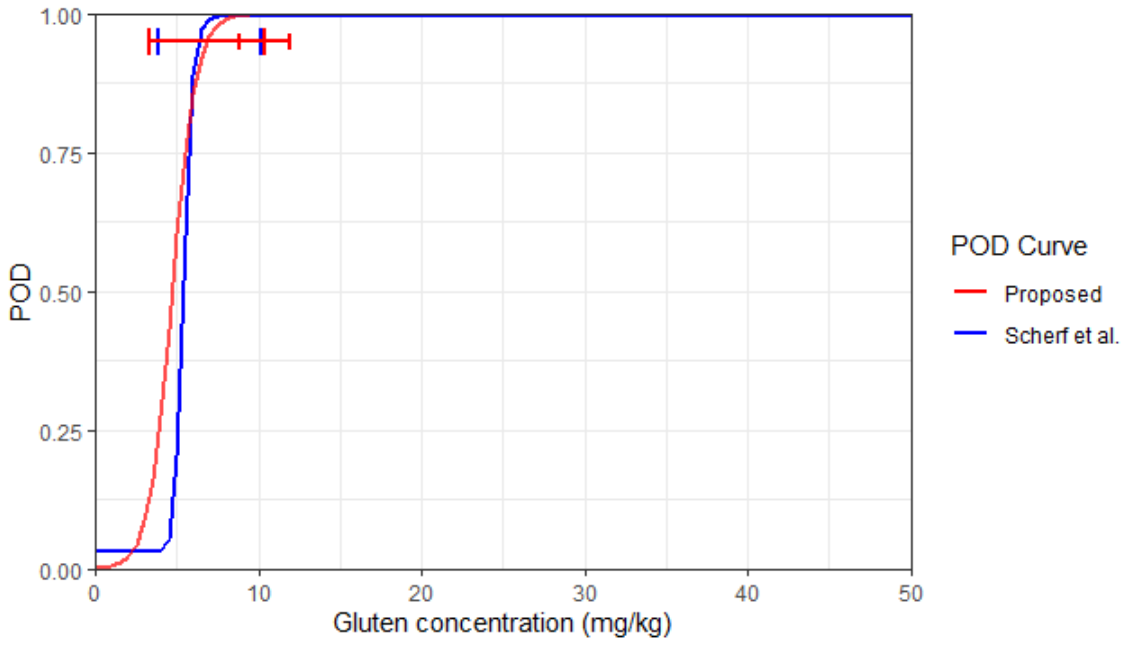

Figure 1: Estimated LPOD curve under the proposed method vs. method of Scherf et al.. Interval (11) is represented (tall red error bars). Interval (12) about the upper limit of (11) incorporates additional uncertainty in the effect estimates (shown with short red error bars). Numerical results for these intervals can be found in Appendix B

Based on the application of the proposed method to the data, it can be seen that LPOD curves appear to be similar, most notably in the region on which the LPOD values change rapidly (between 0.4mg/kg and 13.3mg/kg). It can also be noted that the interval (11) provided by the proposed method is estimated to be marginally wider than that provided by Scherf et al., with upper limits being very similar (10.37mg/kg vs. 10.20mg/kg, respectively). Taking into account the uncertainty of the effect estimates via (12) (shown using short red error bars) shows that the estimate provided by Scherf et al. falls nearly centrally within this range. This suggests an agreement between these two methods. As Scherf et al. do not appear to take uncertainty of their parameter estimates into account, their $\text{LOD}_{95\%}$ prediction interval may be incorrectly understood as a being more precise than is true in reality. This issue is solved in the proposed method, where sampling variability is considered to allow for more informed decision making. As the data generating mechanism is unknown in this example, it is not possible to make claims about the validity of either method. In order to make such claims about the proposed

method, a simulation study is performed and discussed in Section 3.

3 Simulation

3.1 Aims

The aim of this simulation is to assess the estimation of the parameters defining the distribution of the 95^{th} percentile concentration of a binary test, i.e. the concentration for which the probability of a positive test result for a lab is 95%. Data generation will occur under a grid of scenarios corresponding to different lab effect sizes and sample sizes, with the baseline case being given by the minimum recommendations of ISO/TS 16393. The parameters to be estimated are the mean and variance of the distribution of the 95^{th} percentile concentration (9), along with the variability of these estimates across simulations. An interval to describe the likely range of a randomly chosen lab's 95^{th} percentile concentration will also be constructed, along with a confidence interval about the upper bound of this interval.

3.2 Data Generation

Data will be generated according to the logistic mixed effects model proposed in (7), where each lab has a random intercept component in addition to a fixed intercept and slope. The basic form of this model, in terms of the logit of the probability of success for a test at concentration x at lab i is

$$\ln\left(\frac{\text{POD}_i(x)}{1-\text{POD}_i(x)}\right) = (-10 + b_{0,i}) + 0.5x \tag{13}$$

where

$$b_{0,i} \sim N\left(0, \sigma^2_{b_0}\right).$$

The choices $\beta_0 = -10$ and $\beta_1 = 0.5$ were arbitrary, as the sampled concentrations (discussed below) were chosen to be independent of the particular units of concentration used.

The random lab effect will take on 2 different values in the simulation. The first ($\sigma_{b_0} = 0.6604$) corresponds to a lab effect with standard deviation such that 95% of 95^{th} percentile concentrations will fall within $\pm 5\%$ of the mean (θ_1 as defined in (9)) of $25.88\bar{8}$. In the alternative case ($\sigma_{b_0} = 1.3208$), a larger lab effect was chosen so that 95% of labs will have 95^{th} percentile concentrations fall within $\pm 10\%$ of the mean of $25.88\bar{8}$. In cases where lab effects are smaller than these, estimation of the lab effect variability becomes increasingly difficult, with many estimates being 0. Here it was desired that the lab effect be large enough to be discernable from 0 in most cases, but this may not always be true in practical applications.

According to ISO/TS 16393, 12 samples should be distributed to each of a minimum of 8 labs, resulting in 96 total tests to be performed for a given concentration. Additionally, a minimum of 5 concentrations are to be tested, resulting in a total of 60 tests per lab and 480 total tests. This allocation of samples to labs defines the baseline case to be investigated by the simulation. In order to investigate the effects of increases to either the number of labs or replicates, each of these will be increased by a common multiple of 3 to create 4 sets of conditions for the simulation. While in reality there may exist an imbalance between the costs of increasing the number of replicates versus the number of participating labs, this simulation is designed to illustrate the effects the distribution of tests to labs has on analysis in the case where one may freely allocate replicates or labs to a study. It was believed that increasing the labs or replicates by a factor of 3 was the largest multiple of the minimum requirements that could still be reasonable in many real world scenarios.

			Labs (n_{lab})	
			8	24
$\sigma^2_{b_0} = 0.6604^2$	Replicates per lab (n_{rep})	12	480	1440
		36	1440	4320
			8	24
$\sigma^2_{b_0} = 1.3208^2$	Replicates per Lab (n_{rep})	12	480	1440
		36	1440	4320

Table 2: Total tests performed under each combination of lab sample size, replicates per lab, and lab effect size that are to be examined via simulation. Note: 5 concentrations are tested per lab (e.g. Case 1: $12 \times 8 \times 5 = 480$ tests)

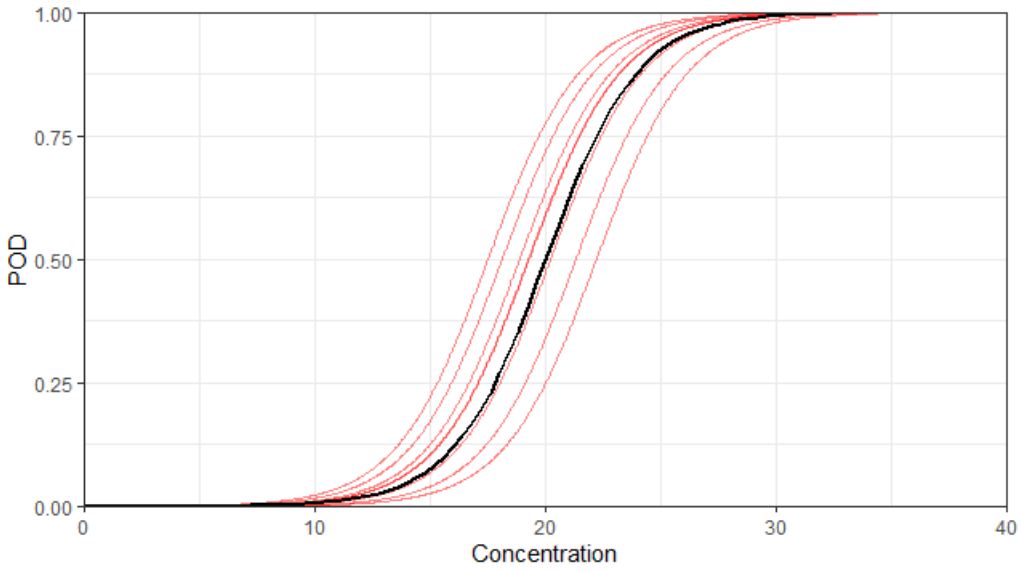

Figure 2: Typical realization of POD_i curves for an 8-lab setup with $\sigma_{b_0} = 0.6604$ (red) with typical lab ($b_{0,i} = 0$, black)

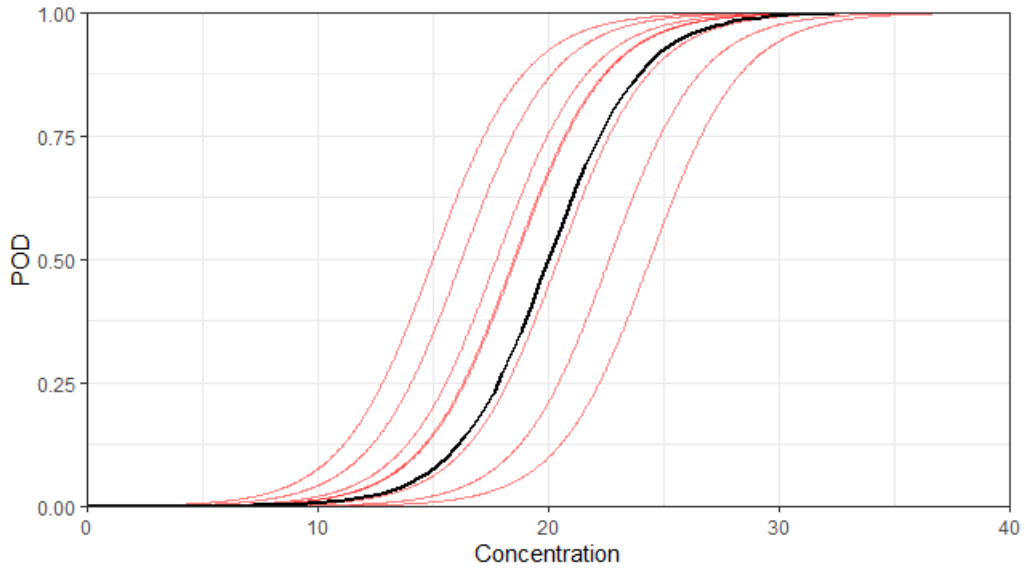

Figure 3: Typical realization of POD_i curves for an 8-lab setup with $\sigma_{b_0} = 1.3208$ (red) with typical lab ($b_{0,i} = 0$, black)

The concentrations x chosen to satisfy the recommendations of ISO/TS 16393 will represent the lower extreme, the lower, the middle, the upper, and the upper extreme of the logistic curve, respectively. The concentrations chosen to satisfy this requirement are the 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} percentile concentrations as defined using (13) with $b_{0,i} = 0$. These concentrations are $14.11, 17.80, 20.00, 22.20$, and 25.89, respectively, and are fixed throughout all scenarios in the simulation. As Figure 1 and Figure 2 show, however, the chosen concentrations will not necessarily correspond to the true 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} percentile concentrations for a particular lab's POD_i curve. Figure 1 and Figure 2 also illustrate that because the random lab effect $b_{0,i}$ affects the location of the POD_i curve; growth in σ_{b_0} also enlarges the range of lab-specific 95^{th} percentile concentrations. Although the choice of concentrations may be a problem worthy of its own research, the selection of concentrations above was justified through the belief that prior to any analysis of this form, preliminary testing of a testing method would be performed to allow for some understanding of which concentrations meet ISO/TS 16393 guidelines.

Test results will be sampled from a $\text{Bin}(n_{rep}, \text{POD}_i(x))$ distribution, where $\text{POD}_i(x)$ is given through (7) for each sampled concentration x. The process outlined above will be

repeated independently to result in a total of $N = 2000$ simulation runs per combination of sample sizes and lab effects.

3.2.1 The 95^{th} Percentile Concentration

In this simulation, the mean and variance of the $95th$ percentile concentrations is given via (14) and (15)

$$\theta_1 = \frac{\ln\left(\frac{0.95}{1-0.95}\right) + 10}{0.5} = 25.88\bar{8} \tag{14}$$

With θ_2 changing depending on the lab variability scenario, giving

$$\theta_{2,low} = \frac{0.6604^2}{0.5^2} = 1.7445 \qquad \text{or} \qquad \theta_{2,high} = \frac{1.3208^2}{0.5^2} = 6.9781 \tag{15}$$

The properties of the estimates of θ_1 and θ_2 of this distribution are to be estimated via the simulation study.

3.3 Methods

For each combination of sample sizes and lab effects, $N = 2000$ runs of the simulation will be performed, whereby a mixed effects model which assumes (13) will be fitted to the data to yield estimates of β_0 and β_1, and $\sigma^2_{b_0}$. Estimation of model parameters is performed through the use of Laplace approximation in R through the package lme4.

3.3.1 Mean (θ_1)

Using these estimates, an estimate of the 95^{th} percentile concentration for each simulation run will be calculated as

$$\widehat{\theta_1} = \frac{\ln\left(\frac{0.95}{1-0.95}\right) - \hat{\beta}_0}{\hat{\beta}_1} \tag{16}$$

It should be noted that these estimates are conditional on the random effects, meaning that this quantity does not reflect an estimate of the concentration for which a randomly performed test returns positive with probability 95%. Instead, this estimate reflects the concentration for which a lab with median random effect ($b_{0,i} = 0$) returns positive with probability 95%.

In order to assess any bias that is present, the average of the $N = 2000$ estimates (14) will be calculated and compared to the known value of $25.88\bar{8}$.

Interest also lies in how these estimates vary. As such, when the estimates for each of the $N = 2000$ runs have been constructed, the variance of these estimates may be investigated. This measure gives an indication as to how sensitive the estimates of θ_1 are to variations in the random lab effect and binomial samples, which is of interest for making conclusions about the appropriateness of the minimum requirements as defined by ISO/TS 16393.

3.3.2 Variance (θ_2)

In estimating the variance of the 95^{th} percentile concentration, the estimate of θ_2 may be given as

$$\hat{\theta}_2 = \frac{\hat{\sigma}^2_{b_0}}{\hat{\beta}^2_1} \tag{17}$$

As with $\hat{\theta}_1$, properties of this estimator are to be investigated. The average of this estimate across all simulation runs may be constructed to evaluate deviation from the known values described in (15). The variance of these estimates may be used to assess how variable estimates of θ_2 are under the minimum guidelines provided by ISO/TS 16393. This estimate will then be combined with the estimate of θ_1 to give intervals (11) or (12). Coverages of the theoretical uppermost 95^{th} percentile concentration by interval (12) will be examined.

3.4 Results

3.4.1 Estimation of θ_1

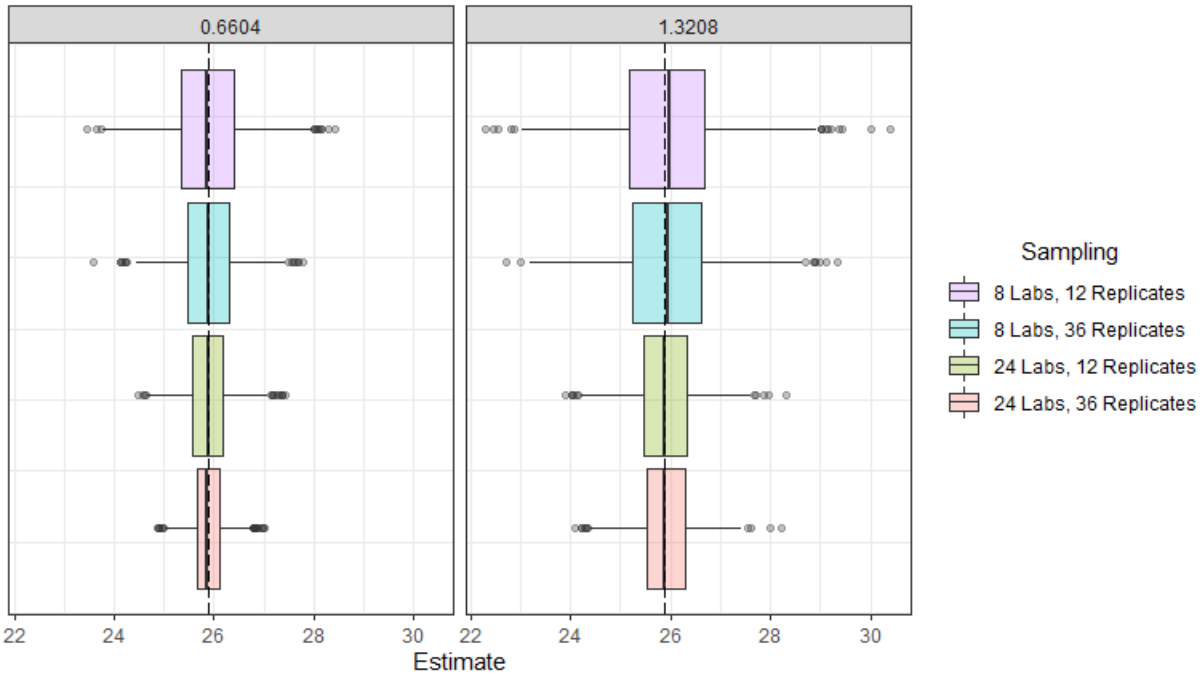

Figure 4: Distribution of estimates for θ_1 across $N = 2000$ simulation runs for each of 2 values of σ_{b_0}.

As can be seen in Figure 4 above, the biases of estimates of θ_1 were minimal, with simulation estimates appearing to be symmetrically distributed around the theoretical value (dashed line). The tightness of this distribution around the true value increased as the number of tests performed increased, with an increase in labs seemingly providing less variable estimates than an increase in replicates. While θ_1 does not explicitly involve the lab variability parameter σ_{b_0}, it is also evident that the lab variability has a large impact on the ability to estimate θ_1, with estimates varying more under the higher lab variability scenario. Depending on the desired certainty in the fixed effect estimates, the minimum recommendations of ISO/TS 16939 may be insufficient to estimate θ_1, as estimates of θ_1 range by $\pm 15\%$ from the true value in the case when only the minimum sample size recommendations are met in a high lab variability scenario. This is in contrast to the case

of 24 participating labs with 12 replicates each, where estimates ranged less than ±10% from the true value in the high lab variability case ($\sigma_{b_0} = 1.3208$).

3.4.2 Estimation of θ_2

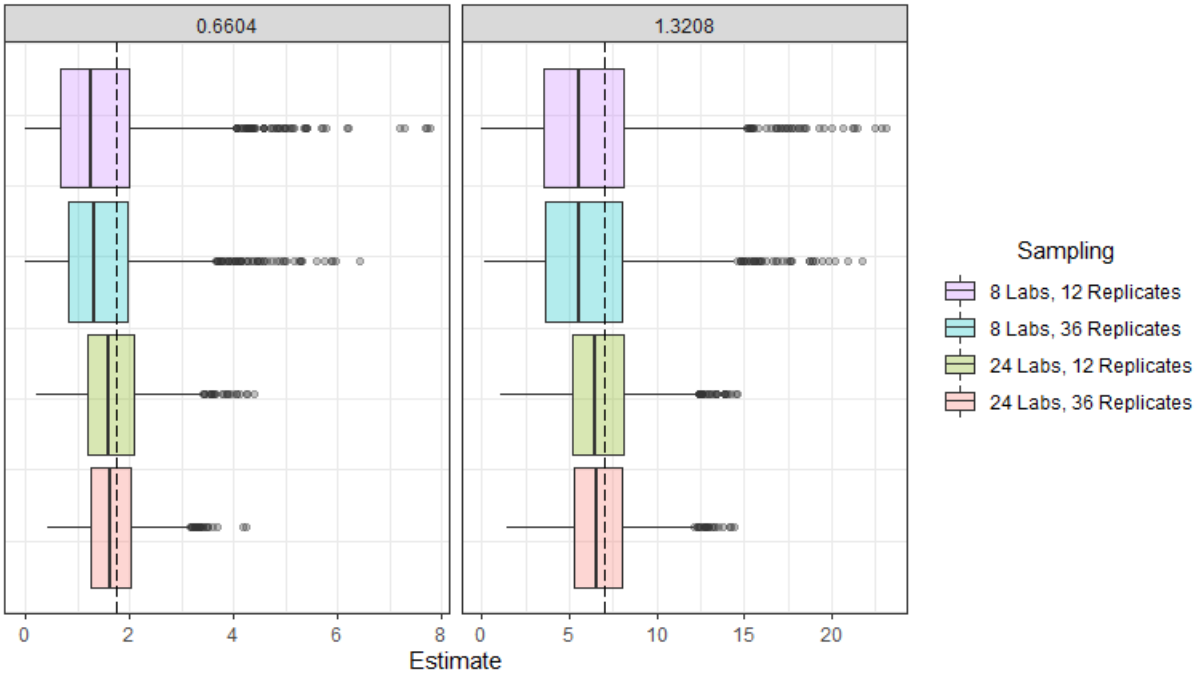

Figure 5: Distribution of estimates for θ_2 across $N = 2000$ simulation runs for each of 2 values of σ_{b_0}.

The distribution of estimates of θ_2 differs strongly from that seen with θ_1. Most notably, these estimates tend to be downwardly biased, with some very large estimates resulting in a right-skewed distribution. As the sample size increased (either by increasing the number of labs or replicates), estimates of θ_2 tended to form a more symmetric distribution that was centred closer to the theoretical value. Notably, the allocation of tests to more labs resulted in less skewed and less variable estimates of θ_2, despite having the same number of tests performed as the increased replicate scenario. Estimates of θ_2 in the low lab variability case ($\sigma_{b_0} = 0.6604$) tended to be quite poor in the case of 8 participating labs, with either 12 or 36 replicates. In these cases, nearly 70% of estimates were lower than the true value, which downwardly affects the width of associated intervals

(11) and (12). This is expected, as Diaz (2007) showed that estimates of σ_{b_0} tend to be downwardly biased under Laplace approximation (used by lme4 in R). Interestingly, Gelman & Hill (2007)[p.275] suggest that in cases with low numbers of random effect levels, variance components are often overestimated, which is seems to be in contrast to what is seen in our simulation.

3.4.3 Interval for the 95^{th} Percentile Concentration

Intervals based on (12) constructed via delta method are shown in Figure 6. Intervals shown are sorted in descending order by value of lower limit for clarity. This interval should ideally cover the upper bound concentration given by the quantile function of (9) (dashed line) at the nominal level of 95%. While almost all intervals were reasonably-well estimated for 6 of the 8 examined scenarios, scenario 1 (8 Labs, 12 replicates, $\sigma_{b_0} = 0.6604$) and 2 (8 Labs, 36 replicates, $\sigma_{b_0} = 0.6604$) resulted in a total of 60 exceptionally poor intervals. Of these, 49 were inestimable due to an inability to obtain standard errors via delta method, arising from an estimate of $\hat{\sigma}^2_{b_0} = 0$. The remaining 11 were a result of exceptionally small variance estimates $\hat{\sigma}^2_{b_0} \approx 0$. These estimates of lab variance resulted in extremely large standard error estimates ($> 1,000,000$) which yielded uninformative intervals. Of these 60 problematic intervals, 59 occurred in scenario 1, while only 1 occurred in scenario 2. Less extreme examples of this phenomena are evident in the case of 8 labs and 12 replicates under $\sigma_{b_0} = 0.6604$ in Figure 6, where interval width covers the entire range of sampled concentrations. This result is an expected consequence of small sample sizes, as lack of random effect levels and replicate count may result in estimates of 0 or nearly 0 for lab effect variance (Gelman & Hill, 2006, p.275). In cases where the lab effect is small and sample sizes are not increased from the minimum provided by ISO/TS 16393, intervals increasingly become inestimable using this method.

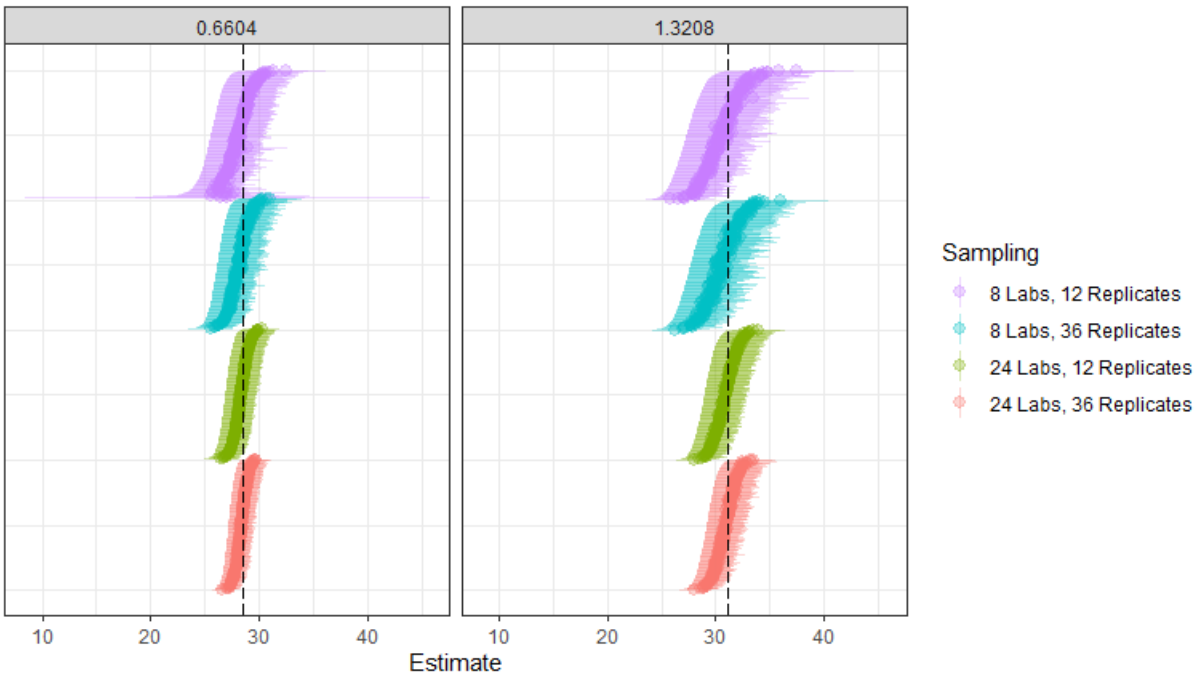

Figure 6: Subset of intervals constructed under each sampling scheme (sorted descending by lower limit). True 95^{th} percentile concentration of an average performing lab (θ_1) shown in solid black. 2 standard deviation limit from this concentration ($\theta_1 + 2\sqrt{\theta_2}$) shown in dashed black.

		θ_1 Bias	θ_1 Variability	θ_2 Bias	θ_2 Variability	Interval Coverage (%)
$\sigma_{b_0} = 0.6604$	8 Labs, 12 Replicates	-0.0021	0.5987	-0.2751	1.1995	95.50
	8 Labs, 36 Replicates	0.0023	0.3480	-0.2514	0.8343	94.10
	24 Labs, 12 Replicates	-0.0053	0.1935	-0.0734	0.4271	96.35
	24 Labs, 36 Replicates	-0.0056	0.1141	-0.0649	0.3151	95.60
$\sigma_{b_0} = 1.3208$	8 Labs, 12 Replicates	0.0537	1.2592	-0.830	12.7002	88.20
	8 Labs, 36 Replicates	0.0489	0.9893	-0.8664	10.9021	87.95
	24 Labs, 12 Replicates	0.0112	0.4091	-0.2308	4.6608	92.10
	24 Labs, 36 Replicates	0.0142	0.3298	-0.2346	4.0500	91.80

Table 3: Results from simulation study

As can be seen by Table 3, increasing the number of tests performed tended to result in more accurate estimates of θ_1 and θ_2, as would be expected. In general, allocating tests to more labs or more replicates did not result in equal increases in estimate quality. Increasing the number of participating labs by a factor of 3 resulted in much less biased estimates of θ_2 (with a slight increase in bias in the case of θ_1) when compared to increasing the number of replicates at each lab. In addition to problems estimating small values of $\hat{\sigma}_{b_0}$ discussed previously, estimation also seemed to be poor under the minimum lab recommendations of ISO/TS 16393 when $\hat{\sigma}_{b_0}$ was large. While increasing the number of replicates did not seem to considerably improve estimation, this was not true of increasing only the number of labs from the minimum of 8. While it may be more beneficial to increase labs rather than replicates given the choice, it is unlikely that these two choices would result in equal costs, however. Increases in replicates are likely to be more feasible than increasing the number of participating labs.

Intervals about the uppermost 95^{th} percentile concentration, i.e. the concentration given as

$$\frac{\ln\left(\frac{0.95}{1-0.95}\right)+10}{0.5}+1.96\sqrt{\frac{\sigma_{b_0}^2}{0.5^2}} \tag{18}$$

appeared to cover the true theoretical concentration at nearly the nominal level of 95%. Interval coverage under the low lab variability scenario ($\sigma_{b_0} = 0.6604$) was much closer to the nominal coverage of 95% than was true of intervals in the high lab variability case ($\sigma_{b_0} = 1.3208$). This discrepancy between nominal coverage and observed coverage may illustrate a need for estimating large lab effects to be supported by increased sample sizes much beyond what is recommended by the minimum of ISO/TS 16393.

4 Limitations & Future Work

Despite the ability to often reasonably estimate fixed effects under the sample size scenarios shown in the simulation study, along with interval coverages nearing nominal levels when using the proposed method, there are some limitations of this approach.

As mentioned in Section 2, the proposed model enforces limits of 0 and 1 on the probability of detection, unlike the model of Scherf et al. (2016) which includes two additional fixed effect terms to alter these. While the proposed model benefits from having fewer estimated parameters because of this, it may be poor in cases where a large false-negative or false-positive rate exists. In particular, for tests with large false negative rates ($> 5\%$), there may exist no concentration that satisfies the requirement that its associated probability of detection is greater than 95%.

Beyond the fixed effect specification, the inclusion of random effects in this model additionally requires certain assumptions. As mentioned previously, this approach assumes only single random intercept, and thus assumes that all labs have the same slope parameter on the linear predictor scale. This assumption may not always be justified, as labs may also differ in their slope parameters. Random effect structures that incorporate other sources of uncertainty such as test operators or testing devices may also be of use, but were similarly not examined here. As the data generation process and proposed model were of the same (single random effect) form in the simulation study of Section 3, the effect of a discrepancy between the data generation process and the assumed model is unknown and would require further research to evaluate.

In addition to the random effect structure, the distribution of the random effects is also based on assumptions. In the proposed model, the distribution of the random effect is assumed to be normal. In many cases, it is conceivable that lab conditions would affect the ability of a test to detect a substance in an asymmetric manner which is inconsistent with the assumption of normality. In these cases, Litière (2008) showed that the misspecification of the random effects distribution can potentially severely affect both the

fixed effect estimates and the random effect variance estimates. Extensions of this work could consider the robustness of this approach to misspecification of the data generating process. As an alternative, if inference about the population is desired (e.g. under LPOD definition (2)), an approach utilizing generalized estimating equations (GEEs) may be advantageous, as no distributional assumption is imposed upon the random effect (Hubbard et al., 2010).

References

[1] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[2] Rafael Diaz. Comparison of pql and laplace 6 estimates of hierarchical linear models when comparing groups of small incident rates in cluster randomised trials. *Computational Statistics & Data Analysis*, 51:2871–2888, 03 2007.

[3] D. J. Finney. Radioligand assay. *Biometrics*, 32(4):721–740, 1976.

[4] International Organization for Standardization. Molecular biomarker analysis — determination of the performance characteristics of qualitative measurement methods and validation of methods. ISO 16393:2019, International Organization for Standardization, Geneva, Switzerland, 2019.

[5] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press, 2006.

[6] Arjon Hengel, Claudia Capelletti, Marcel Brohee, and Elke Anklam. Validation of two commercial lateral flow devices for the detection of peanut proteins in cookies: interlaboratory study. *Journal of AOAC International*, 89:462–8, 03 2006.

[7] Alan E. Hubbard, Jennifer Ahern, Nancy L. Fleischer, Mark Van der Laan, Sheri A. Satariano, Nicholas Jewell, Tim Bruckner, and William A. Satariano. To gee or not to gee: Comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, 21(4):467–474, 2010.

[8] S. Litière, A. Alonso, and G. Molenberghs. The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Statistics in Medicine*, 27(16):3125–3144, 2008.

[9] Department of Defense. Nondestructive evaluation system reliability measurement. MIL-HDBK 1823A, 2009.

[10] K.A. Scherf, S. Uhlig, K. Simon, K. Frost, P. Koehler, T. Weiss, and M. Lacorn. Validation of a qualitative r5 dip-stick for gluten detection with a new mathematical-statistical approach. *Quality Assurance and Safety of Crops & Foods*, 8(2):309–318, 2016.

[11] Steffen Uhlig, Kirstin Frost, Bertrand Colson, Kirsten Simon, Dietrich Mäde, Ralf Reiting, Petra Gowik, and Lutz Grohmann. Validation of qualitative pcr methods on the basis of mathematical–statistical modelling of the probability of detection. *Accreditation and Quality Assurance*, 20:1–9, 04 2015.

[12] Kohji Yamamura and Tamio Sugimoto. Estimation of the pest prevention ability of the import plant quarantine in japan. *Biometrics*, 51(2):482–490, 1995.

Appendices

A Delta Method for interval (12)

$$g(\beta_0, \beta_1, \sigma^2_{b_0}) = \frac{\ln\left(\frac{0.95}{1-0.95}\right) - \beta_0}{\beta_1} + 1.96\sqrt{\frac{\sigma^2_{b_0}}{\beta_1^2}} \tag{19}$$

$$\nabla g(\beta_0, \beta_1, \sigma^2_{b_0}) = \begin{bmatrix} -\frac{1}{\beta_1} \\ \frac{-\left(\log\left(\frac{0.95}{0.05}\right)-\beta_0\right)}{\beta_1^2} - 1.96\frac{\sqrt{\frac{\sigma^2_{b_0}}{\beta_1^2}}}{\beta_1} \\ 1.96\frac{\sqrt{\frac{\sigma^2_{b_0}}{\beta_1^2}}}{2\sigma^2_{b_0}} \end{bmatrix} \tag{20}$$

Then using the variance-covariance matrix $\mathbf{\Sigma}$ for all 3 parameters of this model, as given with `merDeriv::vcov()`, the approximate variance of this function is

$$\text{Var}(g(\beta_0, \beta_1, \sigma^2_{b_0})) \approx \nabla g(\beta_0, \beta_1, \sigma^2_{b_0})^T \mathbf{\Sigma} \nabla g(\beta_0, \beta_1, \sigma^2_{b_0}) \tag{21}$$

B Fitted Models from Section 2

B.1 Model of Scherf et al.

Scherf et al. (2016)[p. 316] state that their fitted LPOD curve is given as

$$\text{LPOD}(x) = \frac{0.031 - 0.996}{1 + (x/5.40)^{19.75}} + 0.996 \tag{22}$$

Where the 95% prediction interval for $C\gamma_i$ is given as (3.33mg/kg,8.76mg/kg) resulting in a $\text{LOD}_{95\%}$ prediction interval of (3.88mg/kg,10.20mg/kg).

B.2 Proposed Model

The fitted LPOD curve under (7) is given as

$$\ln\left(\frac{\mathrm{POD}_i(x)}{1-\mathrm{POD}_i(x)}\right) = (-6.464 + b_{0,i}) + 1.376x \tag{23}$$

where

$$b_{0,i} \sim N(0, 2.485^2). \tag{24}$$

Interval (11) corresponding to the $\mathrm{LOD}_{95\%}$ interval given by Scherf et al. was thus (3.29mg/kg,10.37mg/kg) with a confidence interval (12) about this upper bound given as (8.81mg/kg,11.94mg/kg).