UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Sampling Plan for Microbial Testing of Natural Origin Products

Maria Merezhko

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :
dr. Stijn JASPERS

SUPERVISOR :
Dr. Chella ENSOY-MUSORO

2023
2024

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Sampling Plan for Microbial Testing of Natural Origin Products

Maria Merezhko

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :
dr. Stijn JASPERS

SUPERVISOR :
Dr. Chella ENSOY-MUSORO

Abstract

Pre-gelatinized starch, a natural origin product used in pharmaceutical manufacturing, undergoes microbiological testing to ensure its quality and safety. This testing not only safeguards public health but also the reputation and credibility of the pharmaceutical industry. Acceptance sampling is a statistical quality control procedure used to determine whether a specific quantity of goods or materials should be accepted or rejected. The current sampling plan under consideration is based on the WHO's n-plan, which employs a composite sampling approach that uses the $\sqrt{N} + 1$ equation to calculate the sample size, which is a non-risk-based "rule of thumb" approach. Consequently, the efficacy of this sampling approach varies depending on the range of conditions. Additionally, when a large number of primary samples are combined into a composite sample, it is possible to dilute highly contaminated samples, resulting in failure to detect microbial contamination. To evaluate the effectiveness of the sampling plans, the current study used a simulation approach to draw Operating Characteristic (OC) curves, which provide the probability of acceptance for different mean bacterial counts. The n-sampling has been found to be satisfactory for the majority of conditions characterized by low variability. For conditions that provide unsatisfactory results with n-plan, two alternatives are suggested: WHO's r-plan and alternative composite plan with more than one composite sample.

Introduction

Pre-gelatinized starch is a natural origin product that serves as a binder in pharmaceutical manufacturing to improve the cohesion of the powder mixture. Testing for bacterial contamination in products of natural origin used in the pharmaceutical industry is essential to ensure the quality and safety of these products. Microbiological contamination testing not only safeguards public health but also the reputation and credibility of the pharmaceutical industry.

Therefore, upon arrival of a shipment of pregelatinized starch at a production site, the product undergoes microbiological testing before it is formally accepted or used to ensure the safety of the product. Microbiological testing aims to determine whether a shipment meets the recommended specifications for microbiological quality. Various strategies can be used to test whether shipments are of sufficient quality. To unambiguously check the quality of interest, one may inspect every item in the lot if testing is not destructive. However, such an effective approach is costly and often unnecessary for ensuring product quality. A useful approach that can be used instead is the acceptance sampling (Montgomery, 2009).

Acceptance sampling is a statistical quality control procedure used to determine whether a specific quantity of goods or materials should be accepted or rejected. The basic procedure involves testing a random sample from a large quantity of items relative to the quality characteristic of interest, such as the microorganism concentration. If a sample passed the test, the entire lot was considered acceptable. Alternatively, if the sample failed to pass the test, the entire lot was returned to the supplier. Such inspection should be controlled by a sampling plan that specifies the number of units to be inspected, depending on the lot size and the criteria for accepting or rejecting the lot. Different types of acceptance sampling approaches, such as single, double, and sequential sampling, are used to determine the best plan for both the producer and customer. A good plan would minimize the risk of rejecting a

high-quality lot (producer risk) and/or the risk of accepting a low-quality lot (consumer risk), while keeping inspection costs low.

The performance of a sampling plan is influenced by the parameters of microorganism distribution, such as within- and between-lot variability, and the parameters of the sampling plan itself, including the number of samples collected, specification limit (the microbial concentrations that distinguish acceptable samples from unacceptable ones), and acceptance number (the maximum number of contamination-positive samples allowed for a lot to be accepted). Modifying any of these characteristics necessitates alteration of the sample plan.

There are two broad groups of microorganisms whose presence or count has been tested in the natural origin products used in pharmaceutical production. The first group comprises microorganisms that cause disease, which are commonly referred to as pathogens, and includes microorganisms such as Salmonella, Listeria monocytogenes, Cronobacter spp., and Escherichia coli. When these substances are identified in samples, they typically result in complete rejection of the product. Microbiological tests for pathogens aim to identify their presence or absence rather than count their number. These tests typically involve enrichment to allow pathogen multiplication and require a long time to complete. The second group of microorganisms is the hygiene indicator, which is a non-pathogenic microorganism that is not harmful in low quantities but, when present in large amounts, may suggest considerable contamination by pathogens. Therefore, microbiological tests for hygiene indicators usually focus on a specific group of microorganisms, such as the Total Aerobic Microbial Count (TAMC), and rely on quantifying the number of colony forming units per gram (CFU/g) and comparing it with a predefined limit.

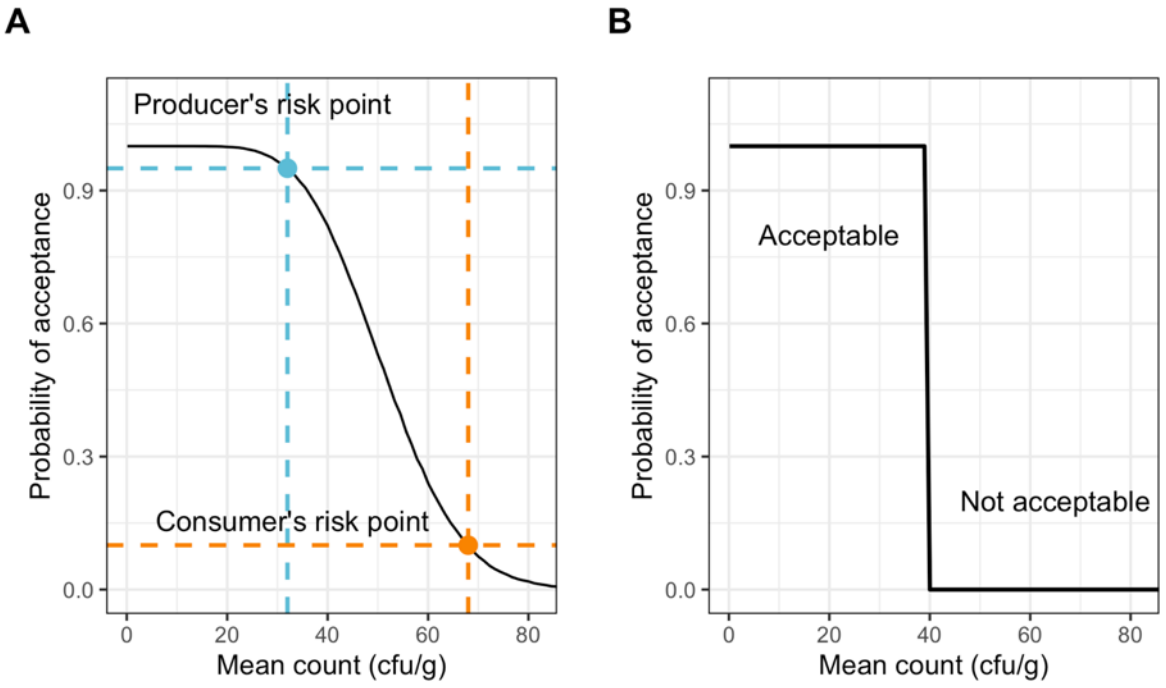

Figure 1. Real (A) and idealized (B) OC curves. The Producer's and Consumer's risk points are shown in blue and orange on the real OC curve, respectively.

The effectiveness of the sample strategy was evaluated by analyzing its Operating Characteristic (OC) curve. The OC curve provides the probability of acceptance for different levels of lot quality, which is commonly represented either as the proportion of non-conforming units in the population (representing the fraction in the lot that does not meet the microbiological limit) or as the mean concentration of microorganisms in the lot. The OC curve graphically illustrates how the chance of accepting a lot decreases as the quality decreases. Figure 1A shows an example of the OC curve.

As shown in the graph, any specific level of microbial contamination corresponds to a certain probability of accepting a lot, and a sampling plan can be designed to control this probability. Normally, there are two points on the lot quality axis whose probabilities we want to control, resulting in two fixed points on the OC curve: the producer's and the consumer's risk points. The producer's risk point defines the probability of rejecting a lot of acceptable quality and is characterized by the Acceptance Quality Limit (AQL) and producer's risk (α). AQL is the worst lot quality that is still considered acceptable, and the producer's risk is the probability of rejecting a lot of this quality. Similarly, the consumer's risk point defines the probability of accepting a lot of poor quality and is characterized by the Limiting Quality Level (LQL) and the consumer's risk (β). LQL is a lot quality that is expected to be rejected with a high probability by consumers (see Figure 1A). A sampling plan can be designed by defining two risk points on the OC curve.

An ideal sampling plan should clearly distinguish between acceptable and unacceptable lots. Acceptable lots should always be accepted, whereas unacceptable lots should not. An example of this ideal plan is shown in Figure 1B. However, in practice, no sampling plan can achieve a perfect OC curve. The transition between the 100% and 0% acceptance probabilities is gradual. The steeper the OC curve, the more discriminating the sampling plan, allowing us to differentiate between acceptable and unacceptable lots more effectively. The chance of incorrectly accepting or rejecting a lot decreases as the sample size increases; however, the expenses also increase. Therefore, the risks and costs should be balanced when designing the sampling plan.

Acceptance sampling techniques can be classified in diverse ways. One way to classify sampling plans is based on the measured quality characteristics. There are two main types of sample plans in this regard: attribute sampling plans and variable sampling plans. In attribute sampling plans, the measured characteristics are classified on a pass or fail basis using a specification limit. For example, a sample containing less than 100 CFU/g of microorganisms is considered acceptable. A concentration of 100 CFU/g or higher is considered unacceptable. Variable sampling plans, on the other hand, evaluate non-categorized quantitative characteristics.

Description of the problem(s) or research questions

The World Health Organization (WHO) outlines three schemes for sampling pharmaceutical goods and associated materials (World Health Organization, 2005):

1. n-plan: This sampling plan is recommended when the material is consistent and the source is reputable and dependable. The n-plan is a composite sampling plan; thus, prior to testing, the samples are first combined to form a single composite sample (Figure 2B). The

number of primary samples taken can be determined using the formula n=$\sqrt{N}+1$, where n represents the number of samples and N represents the lot size or total number of sampling units.

2. p-plan: This sampling plan can be used when the material is uniform and received from a recognized source, and the main purpose is to test for identity. The formula n=0.4*$\sqrt{N}$ was used to determine the sample size. The p-plan is not of interest in this study.

3. r-plan: This plan is suitable for sampling non-uniform materials or materials supplied by an unknown source. The r-plan is an individual testing plan (Figure 2A). The sample size was determined using the formula n=1.5*$\sqrt{N}$.

The current sampling plan under consideration is based on the WHO 's n-plan. This plan employs a composite sampling approach with only one composite sample (Figure 2B), in which the quantity of bags sampled is based on a formula $\sqrt{N}+1$. Subsequently, individual samples were combined into a single composite sample for microbiological analysis, which was performed in duplicates. There are two main issues with the WHO n-plan: (1) the use of composite sampling, and (2) the validity of $\sqrt{N}+1$ the rule to establish the sample size.

The composite sampling procedure is often considered controversial. The advantage of composite sampling is the reduced inspection cost. Instead of conducting multiple individual tests, compositing allows the material to be combined from these bags into a single or several composite samples, and therefore, conduct fewer tests. Testing a composite sample is advantageous for testing a sample prepared from a single bag, as compositing is a physical averaging process, and therefore, more accurately represents the lot mean value of interest. However, information regarding this variability has been partially or fully lost. When the sampling plan involves testing several composite samples, it is possible to estimate the variability of the lot at the individual sample level by weighing the measured variance of the composite samples. However, the WHO n-plan requires the testing of a single composite sample; therefore, within-lot variability must be known to estimate the risks imposed by the plan.

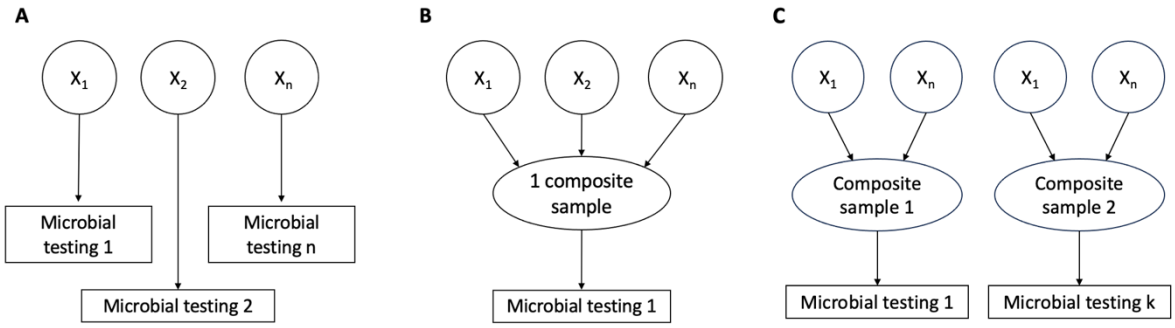

Figure 2. Individual (A) and composite (B, C) sample testing. In individual sampling plans, such as the WHO's r-plan, each sample is tested individually; therefore, the number of analytical samples is equal to the number of items in a sample. In composite sampling plans, such as the WHO's n-plan, several primary samples are combined into one composite sample (B). Alternatively, composite sampling may include several composite samples each consisting of several primary samples (C).

Another concern with composite sampling is the risk of diluting an individual contaminated sample with "clean" samples, and therefore failing to detect any contamination. However, a single uncontaminated sample has the potential to lower the average concentration of other moderately contaminated samples, leading to the approval of a substandard lot. Therefore, the only reason for choosing composite sampling instead of testing all the samples individually is that it is economical, and this methodology has many limitations.

The use of an $\sqrt{N}+1$ equation to calculate the sample size poses another problem because it is not risk-based, which leads to varying levels of effectiveness across different shipment sizes and numerous other associated parameters (Izenman, 2001; Saranadasa, 2003). The probability-based sampling plan estimates the sample size to control for the risk of rejecting a high-quality lot (producer risk) and/or accepting a low-quality lot (consumer risk). The sample size needed to maintain these risks at the desired level depends on multiple parameters, including the underlying distribution of the quality of interest, specification limit, and capability to properly mix the composite sample. The ability of an $\sqrt{N}+1$ equation-based plan to control desired producer's and/or consumer's risks also depends on these parameters. The aims of this study were as follows:

1. To evaluate the efficacy of the existing sampling plan for a range of shipment sizes, within- and between-lot variabilities, and sample mixing quality.
2. Evaluate whether alternative sampling plans are advantageous to the n-plan: (1) WHO's r-plan and alternative risk-based composite sampling plan with more than one sample (Figure 2C)

Description of the methodology

Simulation approach

If individual samples are tested and the distribution is known, the acceptance probabilities for the OC curve can be derived analytically. One can use, for instance, the World Health Organization's risk managers' guide for statistical aspects of microbiological criteria related to foods and their accompanying spreadsheets (Nations & Organization, 2019). However, when composite sampling and several sources of variability are involved, the analytical solution becomes intractable. Therefore, this study used a simulation approach to evaluate the sampling plans over a range of conditions (the detailed simulation algorithm is described below). The methodology here is similar to one used by Edgar Santos-Fernández for variable sampling plans with several composite samples (Santos-Fernández et al., 2015). In contrast to the Santos-Fernández approach, the current project (1) is mainly focused on composite sampling plans with only one composite sample, which was not considered by Edgar Santos-Fernández, (2) uses larger range of primary samples in one composite sample, (3) uses concentration-based attribute sampling, and (4) instead of simulating sample directly, first simulates a lot and then takes a sample from the lot. The later approach is taken as lot sizes typical for pregelatinized starch shipments in J&J are rather small.

The sampling distribution

Microorganisms multiply rapidly, doubling in number with each replication cycle, resulting in a right-skewed distribution, which can be modelled using different distributions.

This requires the estimation of distribution parameters such as prevalence, mean microbial count, standard deviation, and shape parameters. Determining the most suitable distribution for a given situation is often a complex task that requires data collection and the estimation of distribution parameters. In the absence of data, the default approach considers a $\log_{10}$-normal distribution because microbiological populations in foods are often described using a $\log_{10}$-normal distribution. The $\log_{10}$-normal distribution being right-skewed is particularly suitable when the contamination is high and therefore the microorganisms may form clusters. If there is evidence against the $\log_{10}$-normal distribution, one can use alternative distributions, such as Gamma, Poisson-lognormal, or negative-binomial distributions (Jongenburger et al., 2012). The current projects assume that the mean count of microorganisms in pregelatinized starch follows a $\log_{10}$-normal distribution.

When using a lognormal distribution, one should be careful when converting the calculated statistics on a logarithmic scale back to an arithmetic scale. Therefore, it is important to differentiate between arithmetic and geometric means in the original scale. When the average log count is transformed to the original scale via direct exponentiation, it provides a geometric mean, resulting in underestimation of the microorganism count and risks. Therefore, in the current study, the arithmetic mean, calculated as the average microorganism count in the incoming lot, was used to evaluate the risks, as recommended by the World Health Organization's risk managers' guide for statistical aspects of microbiological criteria related to foods (Nations & Organization, 2019). All plots use arithmetic mean counts if not mentioned otherwise.

Sources of variability

In addition to (1) within-lot variability, the current study also analyzed how (2) lot-to-lot variability and (3) additional variability due to imperfect mixing impacts the performance of sampling plans.

Modeling lot-lot variability

Lot-to-lot variability was added to the model via hierarchical modeling. Thus, each lot's mean was first computed at the $\log_{10}$-scale as a sample of size one from the normal distribution centered around the theoretical mean value on the $\log_{10}$-scale with a standard deviation equal to 0.2, 0.4, or 0.8, corresponding to low, medium, and high lot-to-lot variability. If no lot variability is assumed, the lot values are exactly equal to the theoretical mean value.

Modeling variability due to imperfect mixing

The process of mixing/blending is another source of variability in sampling plans, as imperfect physical averaging in composite sample preparation can lead to less representative final products. The quality of mixing during composite sample preparation varies depending on both the type of material and sample preparation method. For instance, liquids are generally mixed better than solid materials such as starch powder. When homogenization is performed manually, the mixing quality is worse than that of automatic sample preparation. Therefore, mixing depends not only on the material, but also on the equipment used to prepare a composite sample.

The mean count of microorganisms in the composite sample can be seen as a weighted average of the primary samples (Elder et al., 1980). When the contributions of primary samples are well controlled and, therefore, equal, the weights become fixed and can be described by a discrete uniform probability distribution (Patil et al., 2011). Otherwise, when the contributions of primary samples are unequal (imperfect mixing), the weights become random and can be described by non-uniform probability distributions, such as the Dirichlet distribution, multivariate hypergeometric distribution, and negative binomial (Brown & Fisher, 1972; Rohde, 1976; Santos-Fernández et al., 2015). The current project models mixing quality with the Dirichlet distribution with the shape parameter equal to 0.1, 1, or 5 to model the poor, moderate, or good mixing quality, respectively (Santos-Fernández et al., 2015).

Summary of the range of parameters modelled:

The performance of the acceptance sampling plans using composite sampling and $\sqrt{N} + 1$ equation to determine the sample size was compared across a range of four parameters:

- Three lot [sample] sizes (10 [4], 50 [8], and 100 [11]) were chosen to represent three categories of lot sizes common for pre-gelatinized starch inspection at J&J: small (10 items), medium (50 items), and large (100 items).
- Three within-lot variability levels (low: sdlog = 0.2; medium: sdlog = 0.4; high: sdlog = 0.8).
- Four between-lot variability levels (no variability: sdlog=0; low: sdlog=0.2; medium: sdlog = 0.4; and high: sdlog = 0.8);
- Four mixing qualities (perfect, good, medium, and poor).

Importantly, high between-lot variability was included for comparison purposes, as with high lot-to-lot variability, it is recommended to improve the manufacturing process rather than increase the stringency of the sampling plan. Even with a discriminating sampling plan, the sampling costs will be high owing to the constant additional sampling of replacement lots.

Accuracy

Although originally acceptance sampling was not viewed via a hypothesis testing framework, such an approach can provide a better understanding of the sampling plans (Hund, 2014; Samohyl, 2018). Hence, sampling plans can be considered hypothesis tests regarding the quality of the lot. As the consumer faces more severe consequences for incorrectly accepting the lot of poor quality than for incorrectly rejecting the lot of good quality, the null hypothesis in this project is that the lot is of unacceptable quality, and the alternative hypothesis is that the lot is of acceptable quality (Samohyl, 2018).

A true positive result (TP) is one in which the test correctly identifies that the lot is not of acceptable quality. Similarly, a true negative (TN) result is one in which the test correctly identifies that the lot is of acceptable quality. A false negative (FN) result is one where the test

Table 1. Acceptance sampling decision classification

		Lot	
		acceptable	unacceptable
Sample	acceptable	TN	FN
	unacceptable	FP	TP

Table 2. Simulated parameters and their description		
Parameter	**Description**	**Range**
means	True geometric means of the lot-generating process	0.1 - 1500
p.accept	Probability of acceptance for each value of means; used to construct OS curve	0 - 1
concentration.incoming	Bacterial (TAMS) concentration in the lots prior to testing (actual arithmetic mean bacterial count)	0 - +inf
concentration.accepted	Bacterial (TAMS) concentration in the accepted lots (arithmetic)	0 - +inf
outliers.undetected	Probability of accepting the composite sample with at least one individual item outside the specification limit	0 - p.accept
TP	Sample is correctly rejected, where correctly means that the mean lot concentration of the accepted lot is above the specification limit	0 - 1
FP	Sample is falsely rejected, where correctly means that the mean lot concentration of the accepted lot is above the specification limit	0 - 1
TN	Sample is correctly accepted, where correctly means that the mean lot concentration of the accepted lot is not above the specification limit	0 - 1
FN	Sample is falsely accepted, where correctly means that the mean lot concentration of the accepted lot is not above the specification limit	0 - 1

incorrectly concludes that the lot is of acceptable quality, while in reality the lot is of unacceptable quality. Finally, a false positive result (FP) is one where the test incorrectly concludes that the lot is of unacceptable quality, whereas in reality, the lot is of acceptable quality. Consumers are most interested in maintaining a low false negative rate. The false negative rate depends on the prior probability of the lot being acceptable, and therefore on the true mean count of the lot. See Table 1 for the acceptance sampling decision classifications. In this study, H_0 stated that the lot mean count is 100 CFU/g or above, but the same approach can be used for testing H_0 with different mean count values or hypotheses involving prevalence (proportion of items of unacceptable quality in a lot) instead of lot mean count.

The proportion of false negative decisions was evaluated graphically and as a percentage of the area under the OC curve that resulted from false negative decisions (relative FN AUC). To calculate relative FN AUC, first area under OC curve was estimated with numerical integration using the trapezoidal rule (only the curve before LQL was analyzed). Then the area under curve of absolute probability of false negative decisions was calculated in the same way. Relative FN AUC was derived by dividing the FN AUC are by the OC AUC (AUC of acceptance probability):

$$\%\, FN\ AUC\ = \frac{AUC\ of\ FN\ probability\ prior\ to\ LQL}{AUC\ of\ acceptance\ probability\ prior\ LQL}$$

The % FN AUC is a false negative rate measure (1 – Sensitivity) for cases when no prior information on the distribution of mean microorganism count is available. % FN AUC presents an assessment of averaged false negative rate. When data on the average quality of lots is accessible, Figures 8-19 can be utilized to see the potential risk of making a false negative decision in proximity to the average quality value. For cases when the distribution of mean count in shipment is known, Bayesian approach to calculate false negative rate can be used to receive false negative rate tuned by prior distribution. Graphical evaluation of the

probabilities of false negative and false positive decisions helps to further evaluate the effectiveness of sampling plans.

Additionally, the difference between the incoming lot quality and outgoing lot quality (quality of accepted lots) was evaluated graphically to estimate the effectiveness of sampling plans. The mean microbial of incoming and accepted lots was plotted against mean microbial count of incoming lots, therefore the first represent a reference line for the second.

Simulation algorithm

Because the analytical solution is intractable when composite sampling and lot-to-lot variability are involved, a simulation approach was used to evaluate the sampling plans. Table 2 lists the parameters computed using the proposed algorithm.

Step 0: Initialize Parameters

The simulated parameters were set as follows: number of bags in the lot (N), primary sample size (n), number of composite samples (k, used only in custom plans), and standard deviation on the log scale for within-lot variability (sdlog), lot-to-lot variability (sdlog.batch), and mixing quality (mixing) as perfect, good, moderate, or poor. In addition, we set the out-of-specification limit (limit), number of simulated lots (n.sim), and seed for reproducible results (seed). For each mean count value (ranges from 0.1 to 1500 on the geometric mean scale, but the increase steps are not equal (0.1, below 1; 1 below 500, and 10 for the rest), repeat the following steps specified number of simulation times (n.sim). The recommended number of simulation cycles is 50 000.

Step 1: Generate lot contamination levels

For each mean concentration value, the log lot bacterial concentration is generated using a normal distribution centered at the mean concentration value, with a standard deviation equal to the specified lot-to-lot variability (sdlog.batch). If there is no lot-to-lot variability, the log bacterial concentration is exactly equal to the mean concentration value.

Step 2: Generate the lot and the sample

Simulate many size NN using a log-normal distribution with base 10 (mean on log scale is equal to lot contamination level generated in step 2, and standard deviation on the log scale is equal to the specified within-lot variability (sdlog). For each lot, draw a sample or samples of size n from the generated lot without replacement.

Step 3. Calculate the sample mean and incoming concentration

The sample mean is calculated using either the arithmetic mean (perfect mixing) or a weighted mean based on Dirichlet weights, depending on the mixing parameter. The shape parameter used to calculate the Dirichlet weights is equal to 0.1, 1, or 5 if the mixing is poor, moderate, or good, respectively. The incoming bacterial concentration is calculated as the mean bacterial concentration of all items in the lot.

Step 4: Determine acceptance or rejection of the lot

If the sampling plan implies only one composite sample, the mean of this sample is compared to a predefined limit (100 CFO/g). If the sample mean count is below the limit, the lot is accepted; otherwise, it is rejected. If the sampling plan implies several composite

samples, the mean count of each sample is compared to the limit. The lot is accepted only if all samples are below the limit. If individual testing plan is used, when value of each item in a sample has to be below the limit for the positive acceptance decision. Track the number of accepted and rejected lots.

Step 5: Calculate mean bacterial count and presence of outliers in the accepted lots

For the accepted lotes, the mean bacterial count has to be calculated. Additionally, track the number of accepted lots whose sample contains outliers (outlier here is the item with bacterial concentration above the specification limit).

Step 6: Compute True and False Positives/Negatives

Calculate the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) based on the lot acceptance decisions for H_0: lot mean count of >= 100 CFU/g. See Table 1 for decision classification.

Step 7: Aggregate results

For each mean concentration value, aggregate the results across all simulations. The acceptance probability is calculated as the number of accepted lots divided by the number of simulation cycles. Calculate the probability of accepting a sample with an outlier as the number of accepted lots containing an outlier divided by the number of simulation cycles. Calculate the concentration in incoming and accepted lots as means of the corresponding values across all cycles of simulations. Calculate the proportion of TP, FP, TN, and FN for each mean lot concentration and prevalence as the proportion of the corresponding decisions across all simulation cycles. Calculate the probability of outlier detection as the number of accepted samples containing outliers divided by the number of samples containing outliers. Find AQL and LQL values (values of mean count corresponding to producer's and consumer's probabilities.

Discussion and interpretation of results

AQL and LQL values

The performance of acceptance sampling plans was compared across a range of four parameters. Figures 3-5 show OC curves of n-sampling plans, while Figure 6 show OC curves of r-plans. Prior to visual assessment of the curves, the AQL and LQL values are presented for producer's risk α=0.05 and consumer's risk β=0.1 at Tables 3.1-4 for the n-plan with perfect, good, moderate, and poor mixing. Table 3.5 shows the AQL and LQL values for the r-plan. As the LQL point is more important for the consumer, it will be pivotal for the choice of the sampling plan. A consumer may pursue various target LQL values, but here 1000 CFU/g will be used as an example threshold, as it is the accepted microbiological limit for non-sterile substances used in pharmaceuticals according to both the United States Pharmacopeia (USP) and the European pharmacopoeia (European Directorate for the Quality of Medicines & HealthCare, 2023; United States Pharmacopeial Convention, 2009). Hence, plans with LQL > 1000 CFU/g are considered unsatisfactory. However, the OC curves and all values in the AQL-

Table 3.1. AQL and LQL values for sampling plans with perfect mixing. AQL and LQL values of various sampling plans are CFU/g. LQL values > 1000 CFU/g are shown in red.

Variability (sdlog)		Lot (sample) size					
Within-lot	Between-lot	10 (4)		50 (8)		100 (11)	
		AQL	LQL	AQL	LQL	AQL	LQL
0.2	0	69	140	77	126	79	122
	0.2	49	222	50	210	51	208
	0.4	33	535	33	516	33	509
	0.8	24	5991	25	5791	25	5679
0.4	0	33	219	57	172	61	159
	0.2	24	307	45	257	46	241
	0.4	49	679	32	586	32	567
	0.8	40	7091	26	6489	28	6161
0.8	0	31	869	33	487	35	399
	0.2	27	1082	32	629	33	523
	0.4	30	2045	29	1279	28	1087
	0.8	29	18736	30	12131	29	10714

Table 3.2. AQL and LQL values for sampling plans with good mixing. AQL and LQL values of various sampling plans are CFU/g. LQL values > 1000 CFU/g are shown in red.

Variability (sdlog)		Lot (sample) size					
Within-lot	Between-lot	10 (4)		50 (8)		100 (11)	
		AQL	LQL	AQL	LQL	AQL	LQL
0.2	0	68	143	76	128	78	123
	0.2	48	227	50	213	51	211
	0.4	32	537	34	519	34	512
	0.8	23	6205	24	5975	24	5768
0.4	0	32	227	55	179	60	164
	0.2	23	314	44	261	46	248
	0.4	47	699	33	601	33	577
	0.8	39	7152	25	6695	25	6406
0.8	0	30	918	33	513	33	415
	0.2	28	1133	30	663	30	544
	0.4	27	2157	25	1339	25	1125
	0.8	31	19350	31	12446	29	10841

Table 3.3. AQL and LQL values for sampling plans with moderate mixing. AQL and LQL values of various sampling plans are CFU/g. LQL values > 1000 CFU/g are shown in red.

Variability (sdlog)		Lot (sample) size					
Within-lot	Between-lot	10 (4)		50 (8)		100 (11)	
		AQL	LQL	AQL	LQL	AQL	LQL
0.2	0	64	151	71	135	73	130
	0.2	47	230	50	220	50	217
	0.4	33	545	33	529	33	509
	0.8	27	5996	26	5995	27	5893
0.4	0	33	254	51	199	54	181
	0.2	27	346	42	285	44	265
	0.4	44	750	31	627	32	600
	0.8	38	7605	30	7202	24	6720
0.8	0	30	1128	33	619	33	499
	0.2	29	1385	30	800	33	646
	0.4	30	2594	29	1558	30	1307
	0.8	27	21576	30	14431	29	12239

Table 3.4. AQL and LQL values for sampling plans with poor mixing. AQL and LQL values of various sampling plans are CFU/g. LQL values > 1000 CFU/g are shown in red.

Variability (sdlog)		Lot (sample) size					
Within-lot	Between-lot	10 (4)		50 (8)		100 (11)	
		AQL	LQL	AQL	LQL	AQL	LQL
0.2	0	56	182	59	167	61	159
	0.2	44	265	46	251	46	244
	0.4	32	601	32	576	32	570
	0.8	24	6760	23	6233	23	6403
0.4	0	32	392	40	318	41	286
	0.2	24	514	36	429	37	383
	0.4	37	1025	30	880	31	808
	0.8	34	9847	26	9127	25	8356
0.8	0	28	2961	27	1705	27	1343
	0.2	25	3524	30	2111	30	1653
	0.4	27	6321	25	3844	25	2990
	0.8	24	44449	29	31980	37	24313

Table 3.5. AQL and LQL values for r sampling plans (noncomposite plans). AQL and LQL values of various sampling plans are CFU/g. LQL values > 1000 CFU/g are shown in red.

Variability (sdlog)		Lot (sample) size					
Within-lot	Between-lot	10 (5)		50 (11)		100 (15)	
		AQL	LQL	AQL	LQL	AQL	LQL
0.2	0	38	96	33	74	32	68
	0.2	28	146	25	117	23	109
	0.4	20	344	17	278	15	261
	0.8	18	3857	12	3154	12	2891
0.4	0	20	112	14	69	12	57
	0.2	18	153	12	96	10	80
	0.4	18	334	9	212	9	180
	0.8	15	3451	8	2197	8	1849
0.8	0	12	297	4	109	4	76
	0.2	8	374	4	140	4	97
	0.4	5	685	4	258	3	190
	0.8	6	5829	6	2465	3	1802

LQL tables can be used to estimate the performance of these plans against other desirable LQL targets.

As can be seen from the tables, efficiency of the sampling plan using the $\sqrt{N}+1$ equation for the sample size calculation varies across the lot sizes. The extent of this effect, however, depends on within- and between-lot variability. If there is high variability between lots (sdlog[between] = 0.8), none of discussed plans should be used, even if mixing is perfect and within-lot variability is low. The high between-lot variability (sdlog[between] = 0.8) is used in this project as a "limiting" case and is implemented to see the trends clearly. It requires tuning of the manufacturing process since acceptance sampling here would be very costly.

If within-lot variability is low, n-plan can be used with all lot sizes, regardless of mixing quality, as long as between-lot variability is moderate at most. When within-lot variability is moderate, n-plan can still be used unless the mixing quality is poor and the lot size is small. If there is high variability within a lot and the quality of mixing is poor, the n-plan cannot be

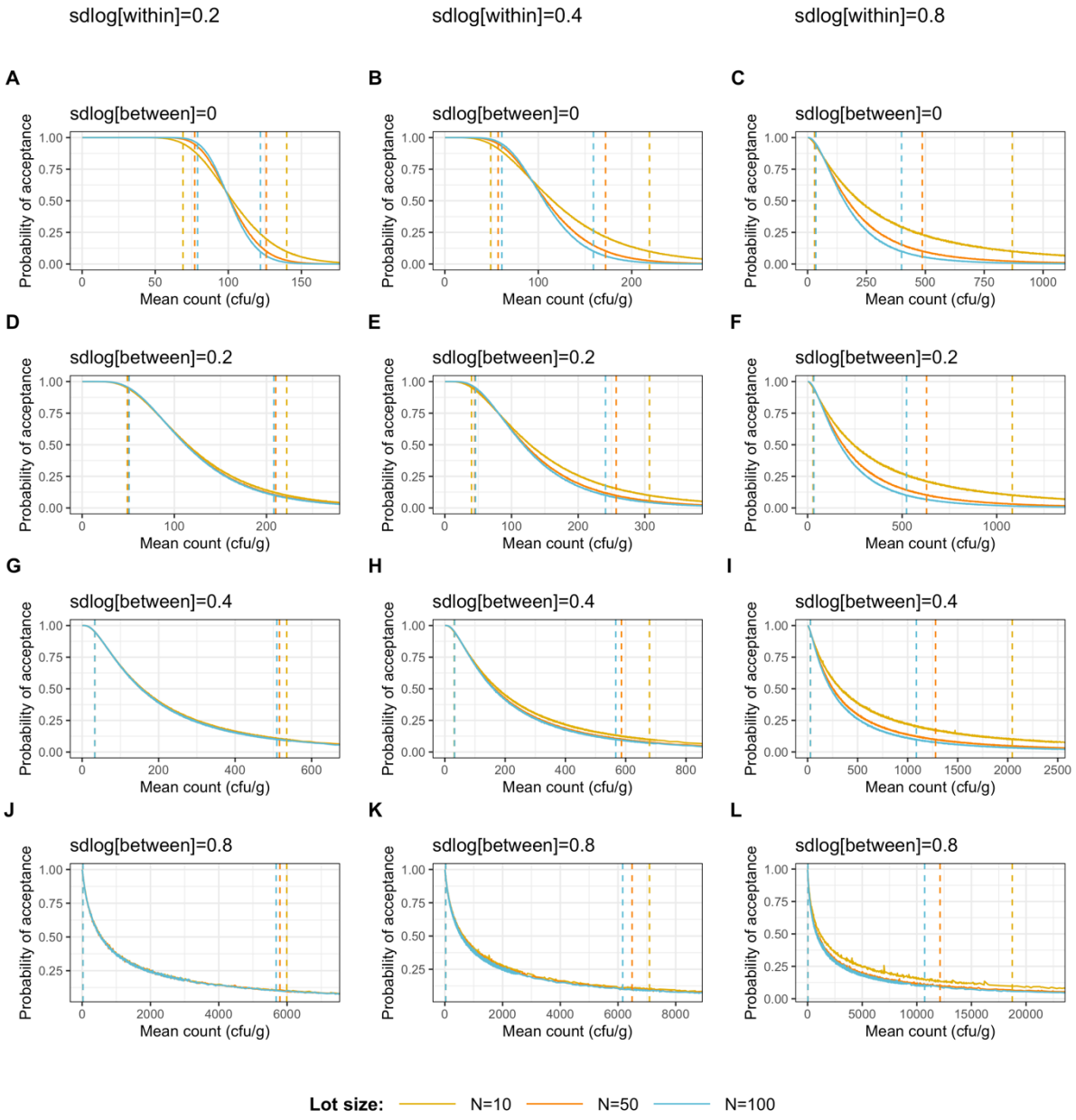

Figure 3. OC curves of n-plans with perfect mixing. The dotted lines correspond to the AQL and LQL values. Within-lot variability is shown at the top; therefore, three columns correspond to three levels of within-lot variability. The between-lot variability increases along vertical lines.

implemented. When the mixing quality is moderate at most and within-lot variability is low or absent, the n-plan can be used with medium and large lots. In these cases, n-plan can also be applied to small lots only when mixing is perfect. Instead of n-plan, r-plan was sufficient to maintain LQL below 1000 CFU/g in all conditions, except when lot-to-lot variability was high.

OC curves

OC curves provide a more complete picture of the sampling plan performance. Thus, Figures 3-6 show that as within- or between-lot variability increases, AQL and LQL tend to decrease and increase, respectively. The OC curves of plans with good mixing are not shown

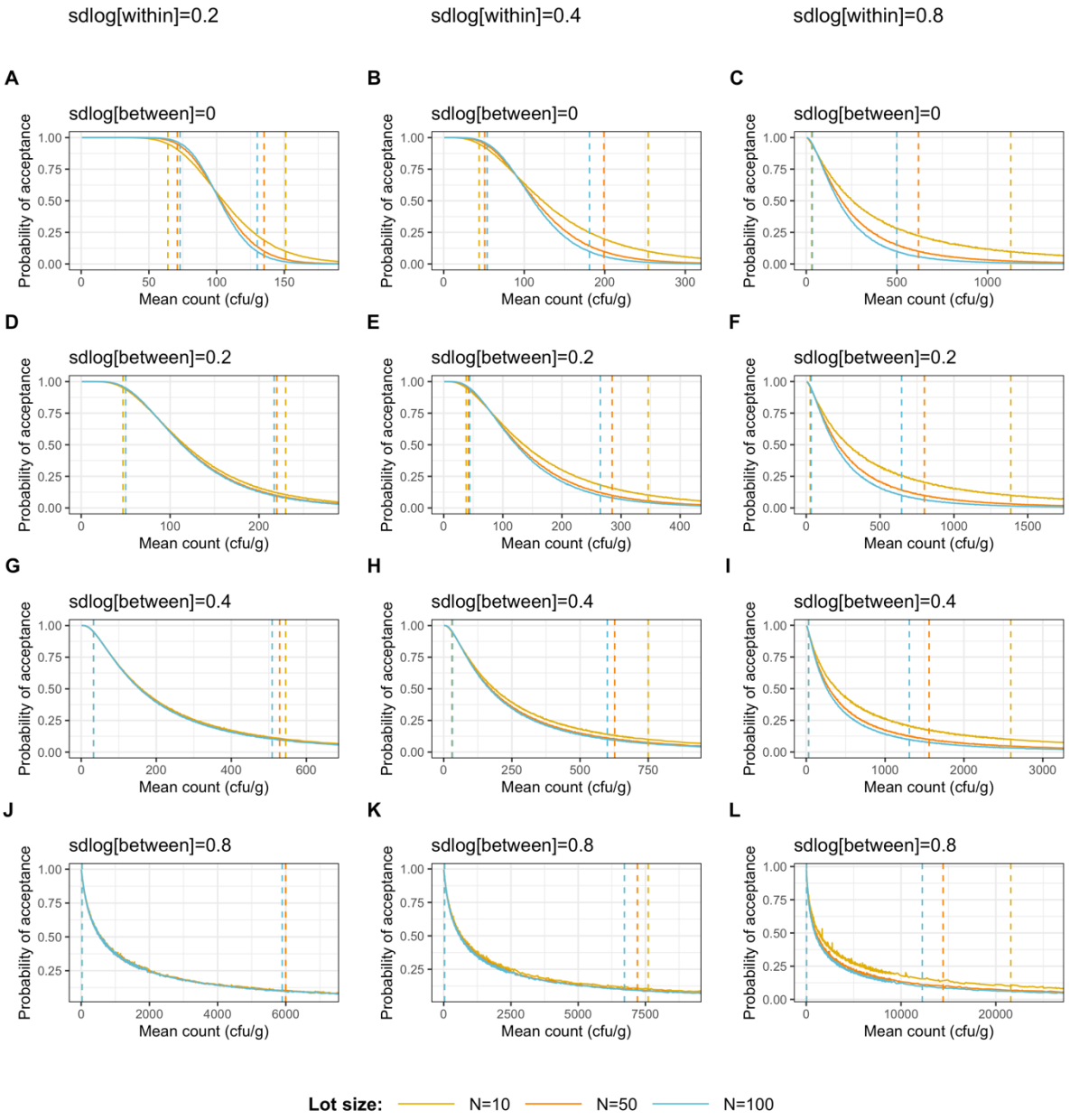

Figure 4. OC curves of n-plans with moderate mixing. The dotted lines correspond to the AQL and LQL values. Within-lot variability is shown at the top; therefore, three columns correspond to three levels of within-lot variability. The between-lot variability increases along vertical lines.

because they closely resemble the OC curves of plans with perfect mixing. Hence, the distance between these points on the x-axis increases; thus, the discriminating ability of the sampling plan decreases. Similarly, when lot, and therefore, sample size decreases, the discriminating ability of the sampling plan also decreases. Consequently, the sampling plan for large lots with low within-lot and no between-lot variability showed the steepest OC curve among all n-plans. Importantly, plans with good discriminating ability show different performances across the three lot sizes, while plans with poor discriminating ability are equally bad across the three lot sizes. The figures also show that between-lot variability seems to have a higher effect on the performance of the sampling plan than within-lot variability. Furthermore, when between-lot variability is high (sdlog=0.8), neither n- nor r-plans provide satisfactory performance. Unfortunately, even if the sampling plan is stringent enough to provide satisfactory

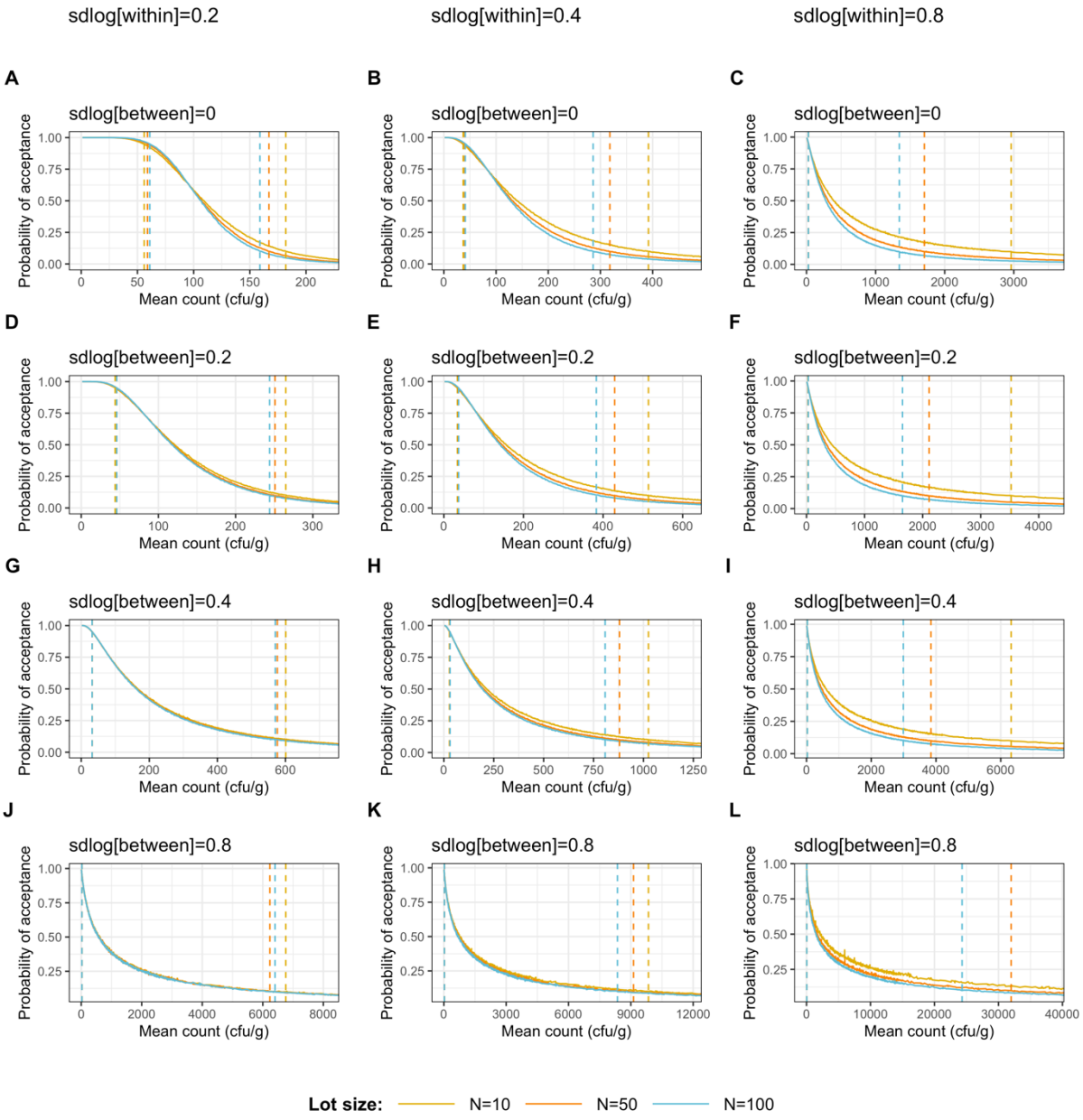

Figure 5. OC curves of n-plans with poor mixing. The dotted lines correspond to the AQL and LQL values. Within-lot variability is shown at the top; therefore, three columns correspond to three levels of within-lot variability. The between-lot variability increases along vertical lines.

performance, the sampling process will be very costly as a large number of lots will be returned to the producer and additional sampling will be required. Therefore, it is recommended to adjust the manufacturing process in such a scenario rather than improve sampling.

Accuracy

The more primary samples included in a composite sample, the closer the mean count of the composite sample will be to the mean count of the lot. However, when the number of primary samples is too high, the risk of dilution arises. The risk of dilution is a known constraint

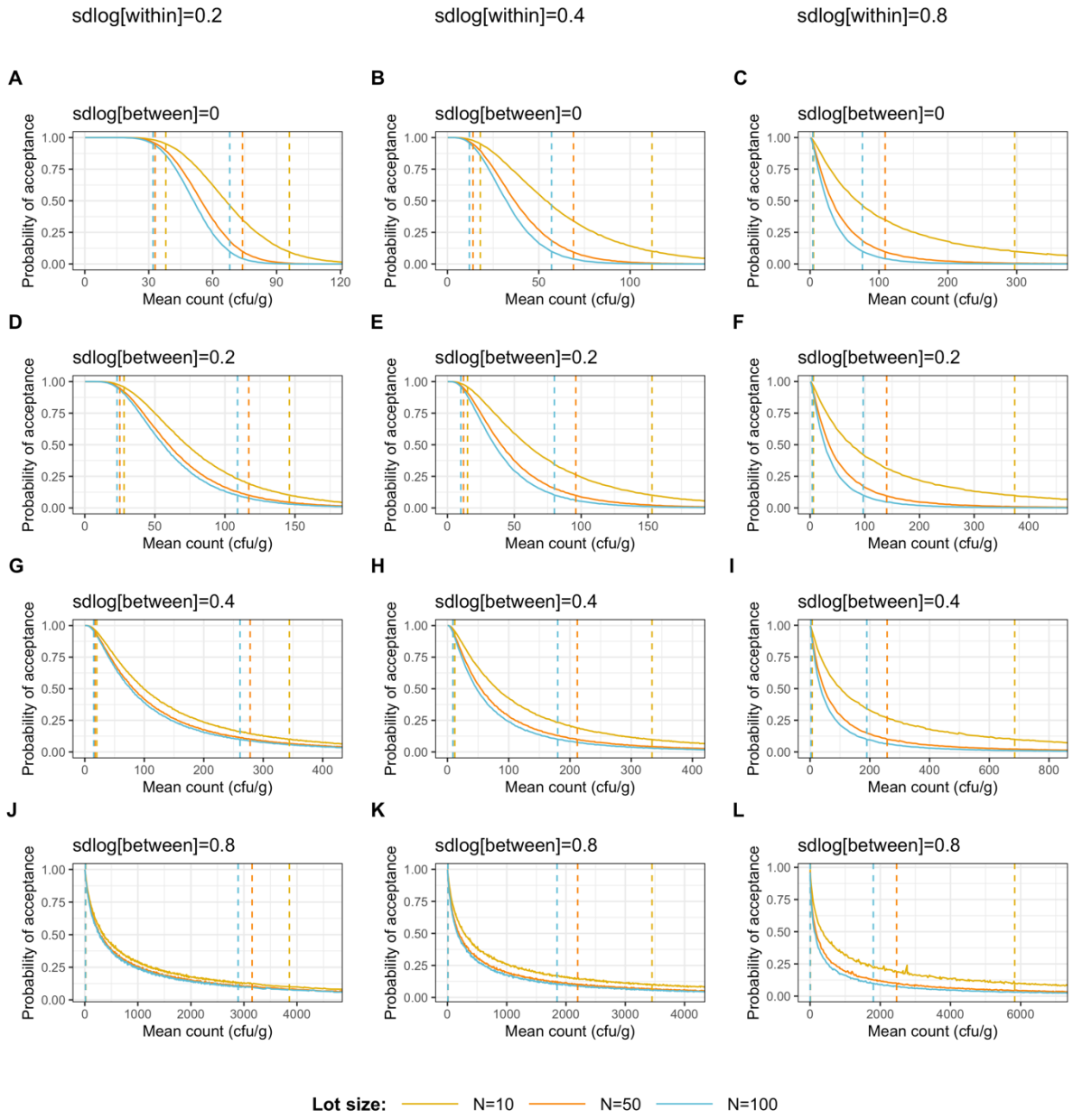

Figure 6. OC curves of r-plans. The dotted lines correspond to the AQL and LQL values. Within-lot variability is shown at the top; therefore, three columns correspond to three levels of within-lot variability. The between-lot variability increases along vertical lines.

of composite sampling, which arises when a single sample with a high bacterial concentration level is mixed with low contamination samples, leading to a composite sample that inaccurately tests negative. If the primary samples were tested individually, the lot would be rejected. To decrease the probability of such a scenario, the maximum limit on the primary sample number is often imposed. Alternatively, the specification limit may be reduced to decrease the probability of FN results. For instance, one recommendation is to divide the specification limit by the number of primary samples in one composite sample. For example, if one aims to reject lots with mean bacterial counts below 1000 CFU/g and use 100 CFU/g as a specification limit for acceptance sampling, the maximum number of primary increments would be 10 (EPA, 1995).

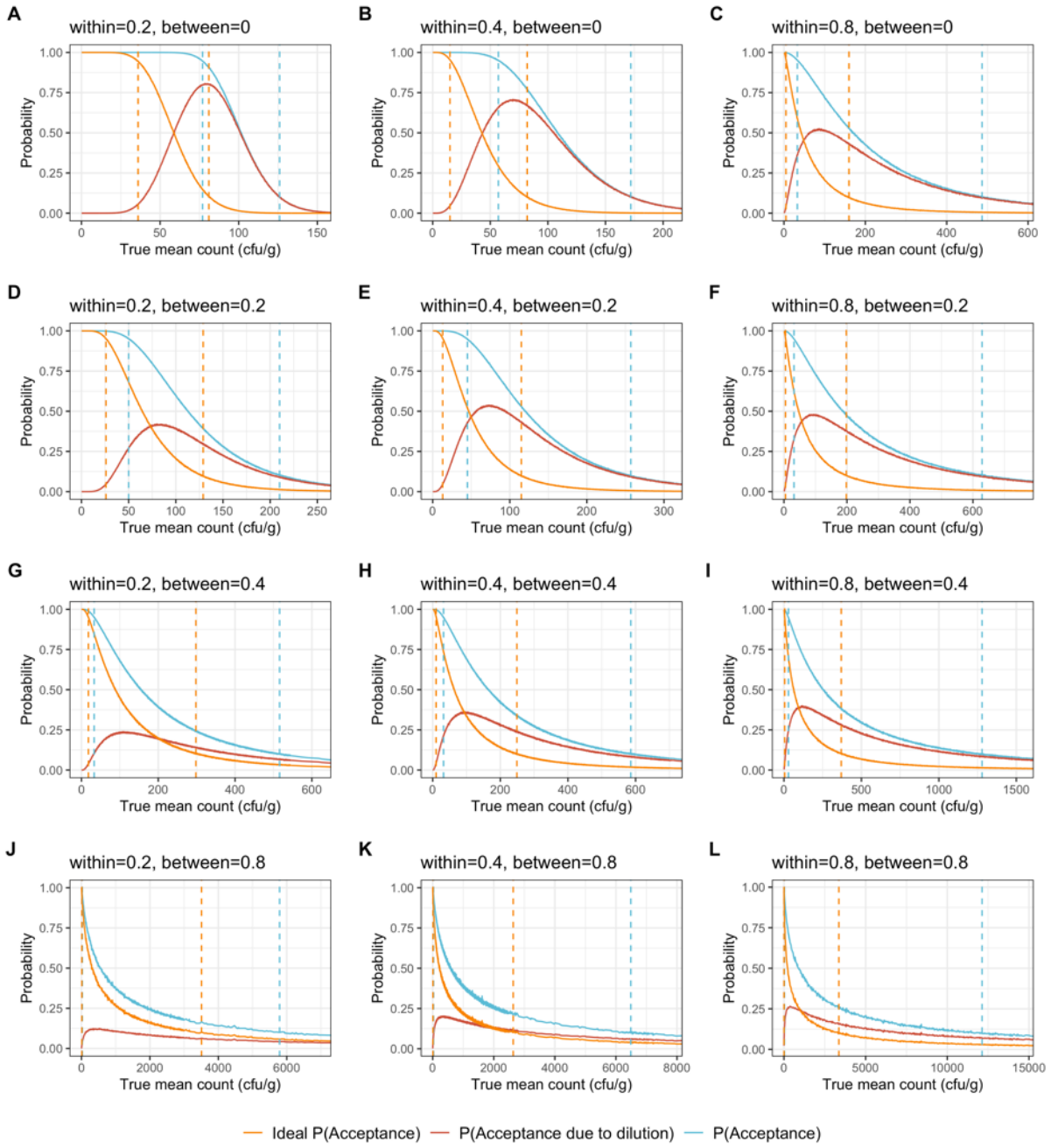

Figure 7. OC curves of n-plan with perfect mixing, Lot size = 50. The dotted lines correspond to the AQL and LQL values. Within-lot variability is shown at the top; therefore, three columns correspond to three levels of within-lot variability. The between-lot variability increases along vertical lines. Teal lines are real OC curves, orange lines show as OC curves would look if samples with outliers were rejected. Red line is the probability of accepting a sample containing an outlier.

Therefore, the accuracy of the n-sampling plan was studied. First, the probability of accepting a sample containing an "outlier" was studied. Note that here, the term "outlier" is used to describe the primary sample above the specification limit, which would be rejected if tested individually. Figures 7 show several examples of real OC curves of sampling plans with perfect mixing and lot size=50 (in teal), the probability of accepting the sample containing at least one outlier (in red), and the ideal OC curve (as the OC curve would look if all samples with outliers were rejected (in orange). The probability of acceptance due to dilutions for the rest of the sampling plans can be seen in Figures 8-19. Plans with a high risk of accepting samples due to

Table 4. Relative % FN AUC. The percentage of area under the OC curve resulting from false negative decision. Green: relative % FN AUC < 10%; yellow: relative % FN AUC < 30%; orange: relative % FN AUC is 30% and above.

Within-lot	Between-lot	Perfect mixing			Good		
		10 (4)	50 (8)	100 (11)	10 (4)	50 (8)	100 (11)
0.2	0	8 %	6 %	6 %	9 %	7 %	6 %
	0.2	7 %	6 %	5 %	8 %	6 %	6 %
	0.4	6 %	5 %	5 %	7 %	6 %	5 %
	0.8	5 %	4 %	4 %	6 %	5 %	4 %
0.4	0	19 %	17 %	15 %	20 %	18 %	16 %
	0.2	17 %	15 %	13 %	19 %	17 %	14 %
	0.4	16 %	14 %	12 %	17 %	15 %	13 %
	0.8	12 %	11 %	9 %	13 %	12 %	10 %
0.8	0	44 %	47 %	45 %	46 %	50 %	48 %
	0.2	42 %	45 %	43 %	44 %	48 %	45 %
	0.4	39 %	41 %	38 %	41 %	44 %	41 %
	0.8	32 %	33 %	31 %	34 %	34 %	32 %

Within-lot	Between-lot	Moderate mixing			Poor mixing		
		10 (4)	50 (8)	100 (11)	10 (4)	50 (8)	100 (11)
0.2	0	11 %	9 %	8 %	19 %	17 %	16 %
	0.2	10 %	8 %	7 %	17 %	16 %	14 %
	0.4	9 %	7 %	6 %	15 %	14 %	13 %
	0.8	7 %	6 %	5 %	12 %	11 %	10 %
0.4	0	26 %	23 %	20 %	43 %	41 %	38 %
	0.2	24 %	21 %	18 %	41 %	39 %	36 %
	0.4	22 %	18 %	16 %	37 %	34 %	31 %
	0.8	17 %	15 %	13 %	29 %	27 %	25 %
0.8	0	54 %	56 %	53 %	79 %	80 %	78 %
	0.2	52 %	54 %	51 %	77 %	79 %	77 %
	0.4	48 %	49 %	47 %	74 %	76 %	73 %
	0.8	39 %	40 %	37 %	63 %	64 %	61 %

Within-lot	Between-lot	r-plan		
		10 (5)	50 (11)	100 (15)
0.2	0	0 %	0 %	0 %
	0.2	0 %	0 %	0 %
	0.4	0 %	0 %	0 %
	0.8	0 %	0 %	0 %
0.4	0	1 %	0 %	0 %
	0.2	2 %	0 %	0 %
	0.4	2 %	0 %	0 %
	0.8	1 %	0 %	0 %
0.8	0	12 %	3 %	1 %
	0.2	12 %	4 %	1 %
	0.4	11 %	4 %	2 %
	0.8	9 %	3 %	1 %

dilution would have a large distance between real and ideal OC curves as well as a high probability of accepting a sample with outliers below LQL.

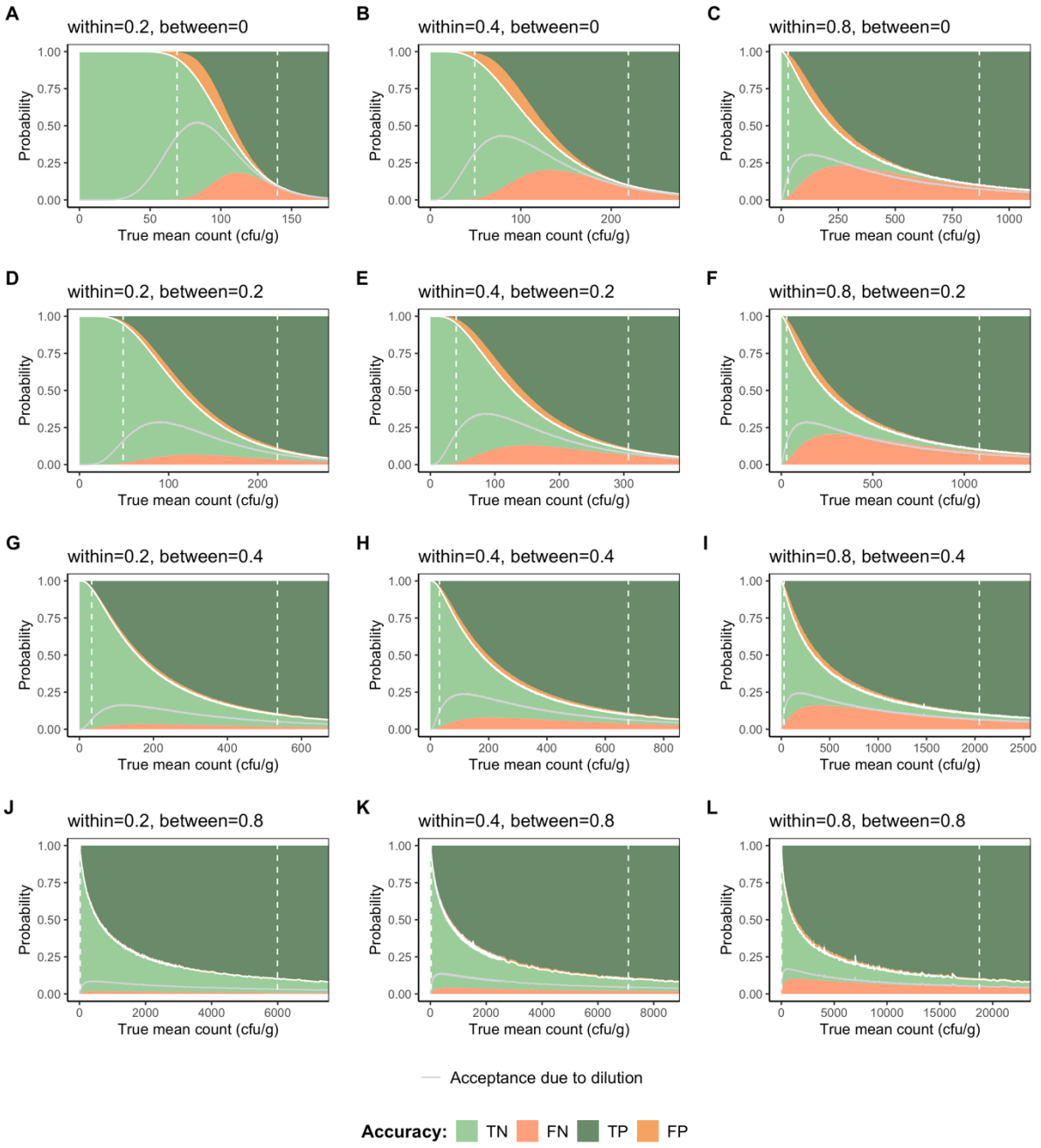

Figure 8. Classification of the sampling decision in the framework of hypothesis testing, where H_0 is that the lot bacteria mean count is < 100 CFU/g. n-plan: perfect mixing, lot size = 10 items.

It seems that the probability of accepting a sample with an outlier peak soon after the AQL, as before the AQL, only a few samples would contain outliers, especially in low variability conditions. The probability of accepting a sample with outliers is already quite low around LQL as outliers there are probably quite extreme and have higher prevalence to measurably affect the mean of a composite sample, resulting in rejection. Within-lot variability seems to increase the relative proportion of acceptance decisions owing to dilution. However, between-lot variability seems to have an opposite effect. For some plots, the majority of accepted samples seem to contain at least one outlier.

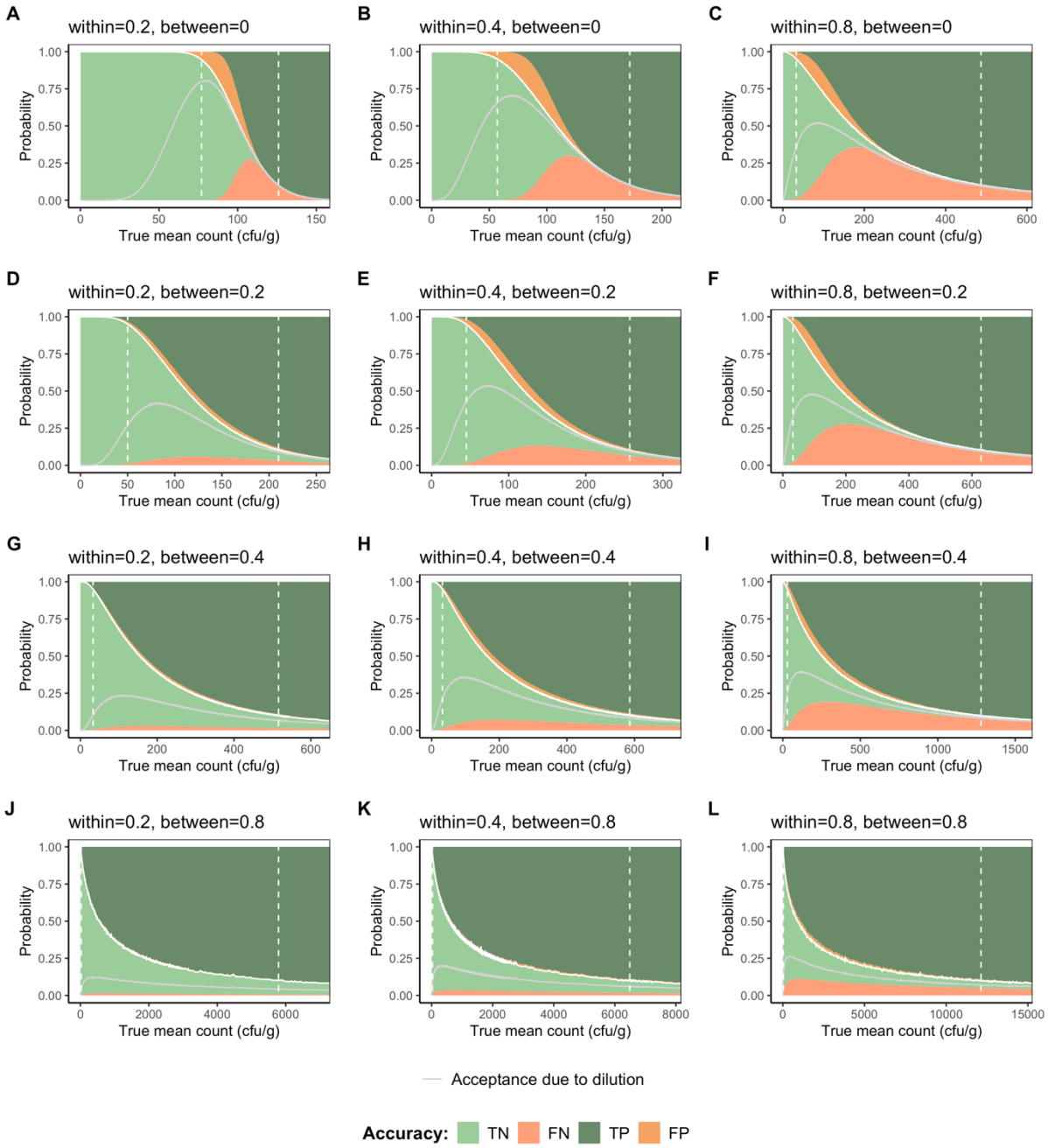

Figure 9. Classification of the sampling decision in the framework of hypothesis testing, where H_0 is that the lot bacteria mean count is < 100 CFU/g. n-plan: perfect mixing, lot size = 50 items.

Acceptance of the sample containing an outlier, however, does not mean that all these decisions are false negatives concerning the lot mean count. Therefore, all accepted decisions were classified in the context of hypothesis testing. Special emphasis was placed on the % of FN decisions (lot mean count > 100 CFU/g, but lot is accepted) as they are the costliest for the consumer. The next series of plots (Figures 8-19) show all decisions classified in the hypothesis testing paradigm. The plots for good mixing quality are not shown as they are almost identical to the perfect quality plots. The decisions are categorized as follows (see also Table 1): (1) True Negative (TN): correctly accepting a lot (lot mean count < 100 CFU/g); (2) False Negative (FN): incorrectly accepting a lot (lot mean count ≥ 100 CFU/g); (3) True Positive (TP): correctly

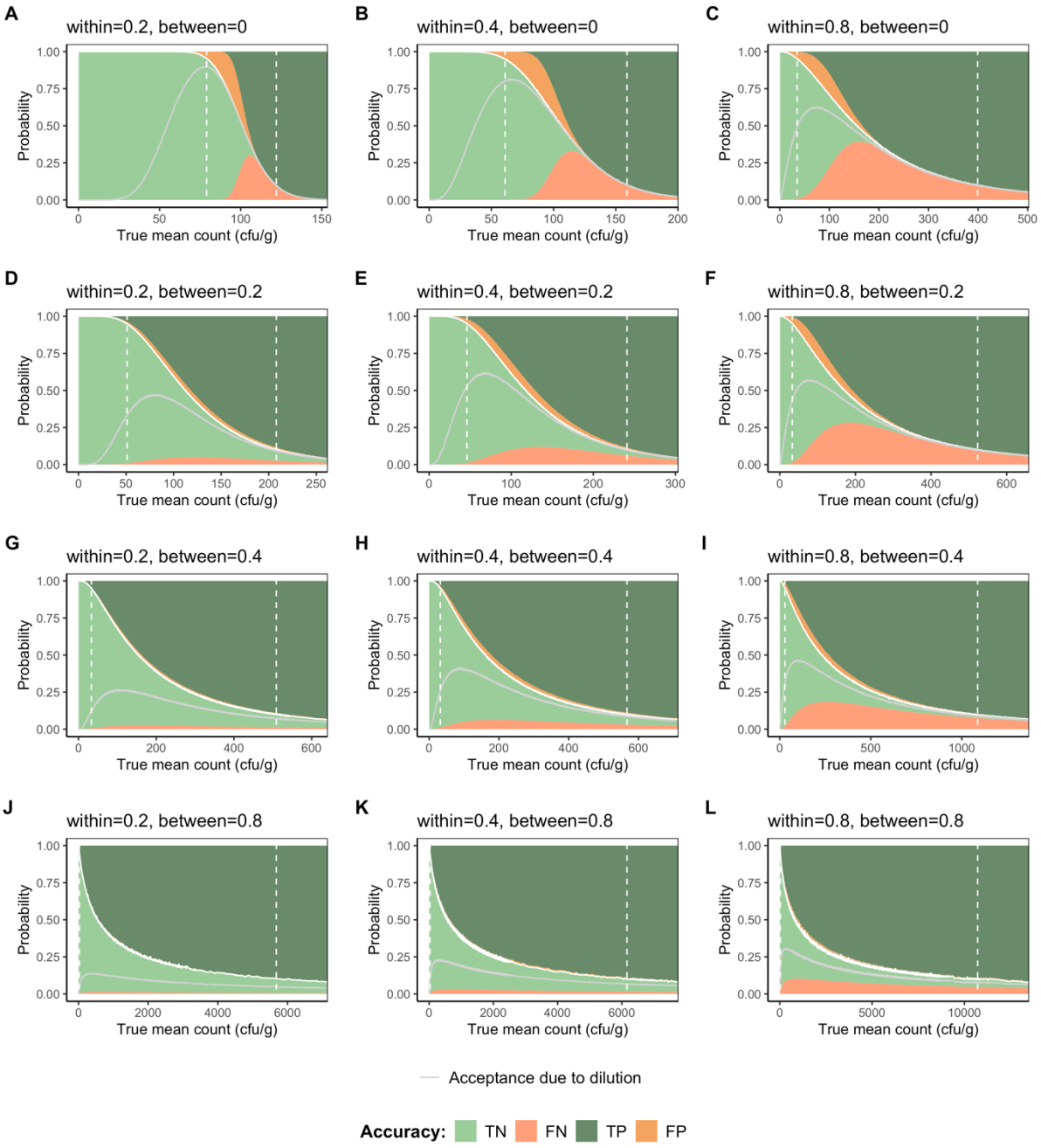

Figure 10. Classification of the sampling decision in the framework of hypothesis testing, where H_0 is that the lot bacteria mean count is < 100 CFU/g. Sampling plan: perfect mixing, lot size = 100 items.

rejecting a lot (lot mean count ≥ 100 CFU/g); (4) False Positive (FP): incorrectly rejecting a lot (lot mean count < 100 CFU/g).

A successful plan is expected to exhibit a low occurrence of both false negative and false positive determinations. However, given that consumer interest primarily lies in minimizing the false negative rate, this aspect is the primary focus of the current study. Hence, the greater presence of green hues and fewer instances of red hues in Figures 8-19 indicate a higher level of success in terms of sensitivity and specificity. When the red color occupies a larger proportion of the area under the acceptance curve, the false negative rate is increased. The

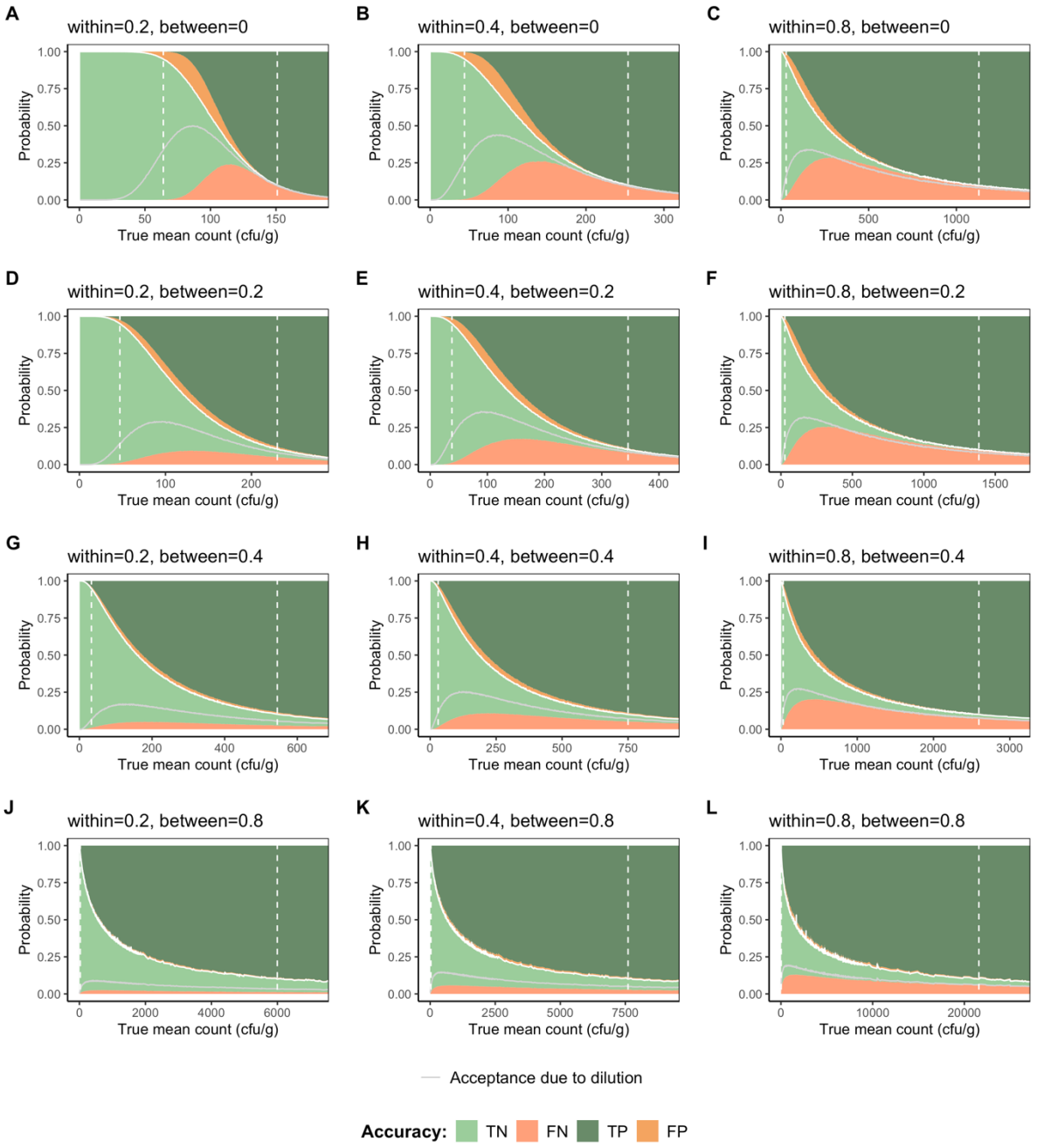

Figure 11. Classification of the sampling decision in the framework of hypothesis testing, where H_0 is that the lot bacteria mean count is < 100 CFU/g. Sampling plan: moderate mixing, lot size = 10 items.

difference in rates is especially clear when one compares n- and r-plans. For example, Figures 8A and 17A depict n- and r-plans designed for scenarios involving small lot sizes, low within-lot variability, and no between-lot variability. Figure 8A displays a significant red area under the acceptance curve, while the acceptance curve at Figure 17A is predominantly green. Consequently, for those seeking to minimize the false negative rate, the r-plan demonstrates notable superiority. Nonetheless, the r-plan exhibits a considerable red area above the curve (FP decisions), contrasting with the smaller area in Figure 8A. Although this study focuses on consumer interests, a very high FP rate is detrimental to consumers as it would lead to the false rejection of multiple lots of acceptable quality, necessitating resampling of replacement

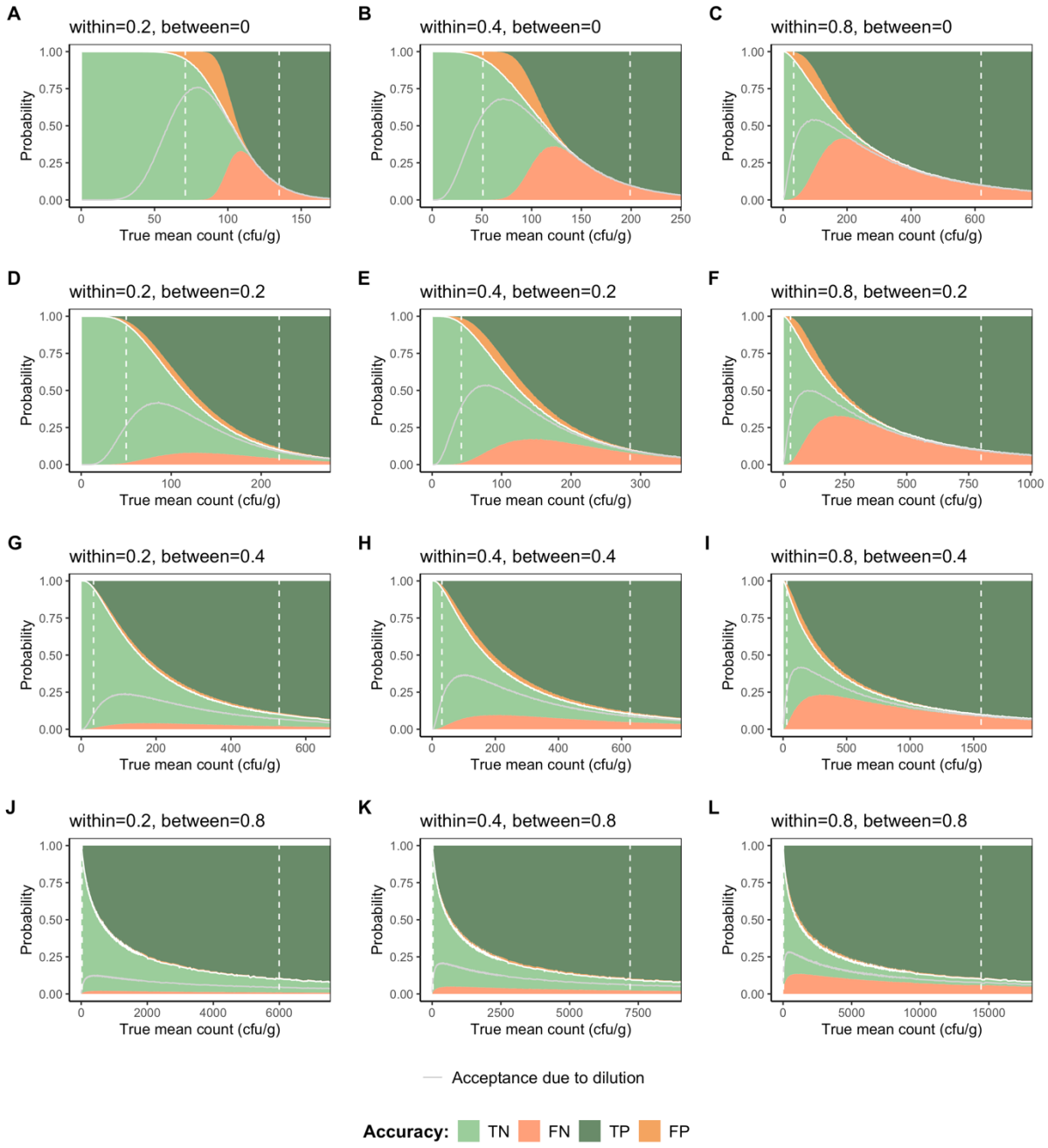

Figure 12. Classification of the sampling decision in the framework of hypothesis testing, where H_0 is that the lot bacteria mean count is < 100 CFU/g. Sampling plan: moderate mixing, lot size = 50 items.

lots, thereby increasing sampling costs and disrupting logistics. It is worth noting that false positive decisions tend to occur at lower mean count values, while false positives are more frequent prior to the LQL. Therefore, prior knowledge of the mean count distribution may also influence the choice of a sampling plan.

The grey line at Figures 8-19 shows the probability of false negative decision as in the Figure 7. Therefore, Figures 8-19 show that, although in low variability conditions and small lot sizes, many accepted samples contained outliers, only a small portion of accepted lots had concentrations below 100 CFU/g (FN decisions). Within-lot variability and poor mixing quality

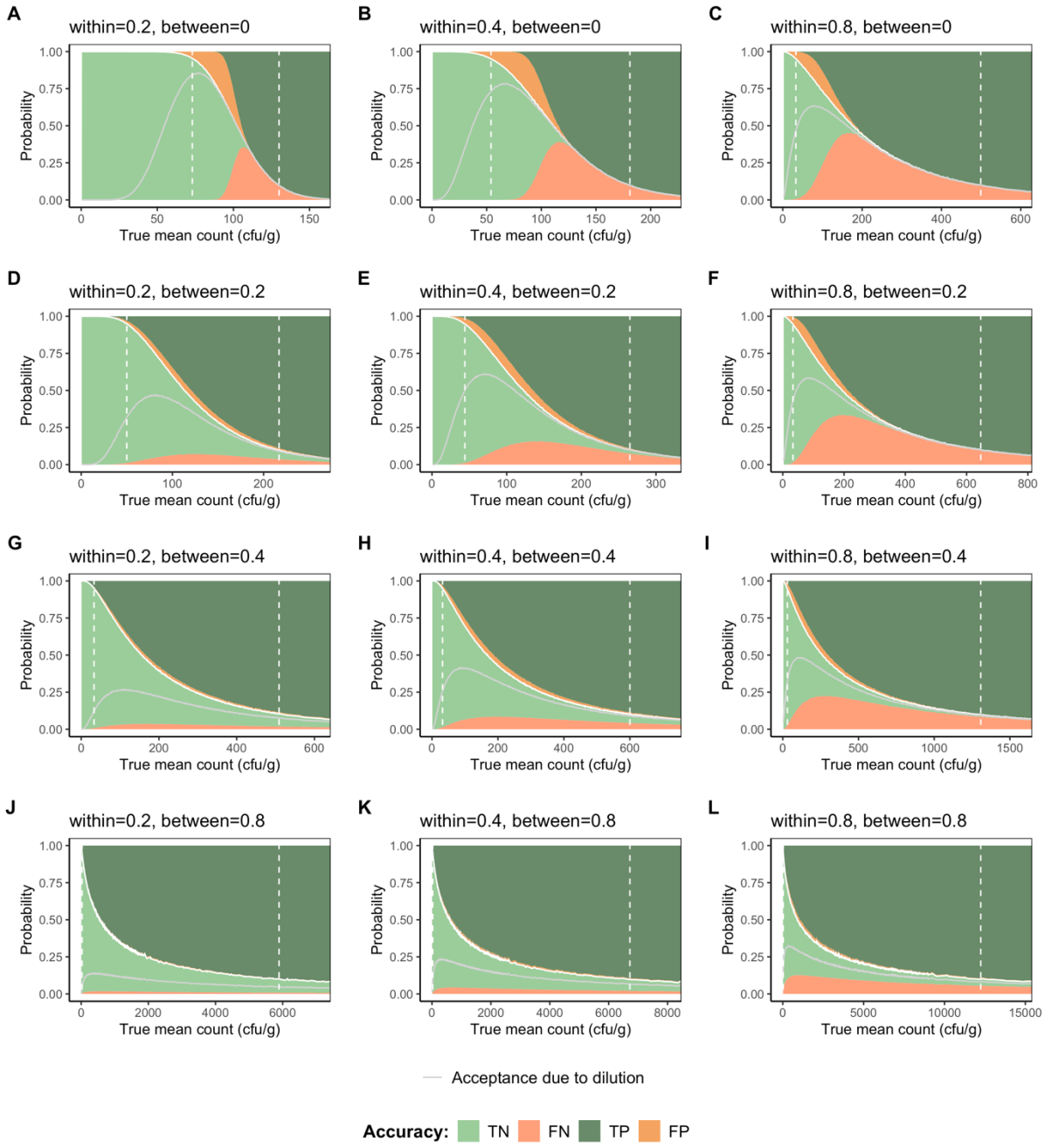

Figure 13. Classification of the sampling decision in the framework of hypothesis testing, where H_0 is that the lot bacteria mean count is < 100 CFU/g. Sampling plan: moderate mixing, lot size = 100 items.

increase the proportion of accepted lots with a mean count below 100 CFU/g (FN decisions). High between-lot variability seems to decrease the proportion of accepted lots with a mean count below 100 CFU/g (FN decisions), and a larger portion of these decisions seems to result from accepting a sample with outliers. These results can be explained by the fact that high lot-to-lot variability results in a high difference between the outgoing and accepted quality.

The Figures 18-19 show that although r-plan considerably decreases the proportion of false negative decisions, it dramatically increases the proportion of false positive decisions. This difference is especially high in low-variability conditions, where the n-plan seems to perform well enough. Although false positive decisions are of less concern for consumers,

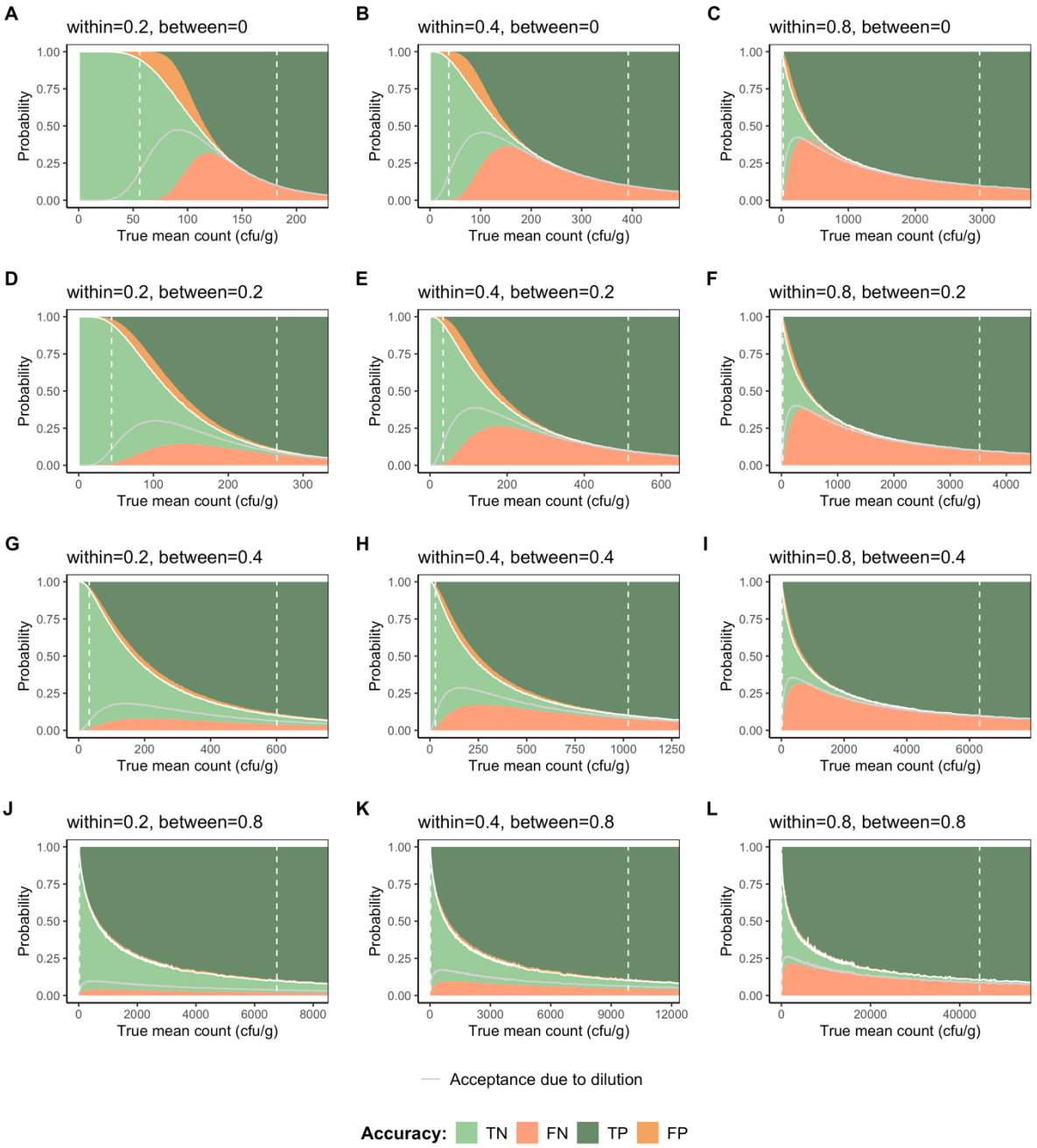

Figure 14. Classification of the sampling decision in the framework of hypothesis testing, where H_0 is that the lot bacteria mean count is < 100 CFU/g. Sampling plan: poor mixing, lot size = 10 items.

their amount still has to be kept below a certain limit; otherwise, too many lots of sufficient quality would be returned to the producers, and additional sampling of replacement lots will increase the sampling costs.

Table 4 shows the percentage of area under the OC curve consisting of false negative decisions (FN AUC). Plans were color-coded depending on the FN AUC: low FN AUC (< 10%, green), medium FN AUC (< 30%, yellow), and high FN AUC (≥ 30 %, orange). As all plans with high (sdlog-0.8) within-lot variability had a relative % FN AUC of more than 30%, these plans are not recommended. One may choose to use r-plans or alternative to composite plans instead, which will be further elaborated upon in subsequent discussions.

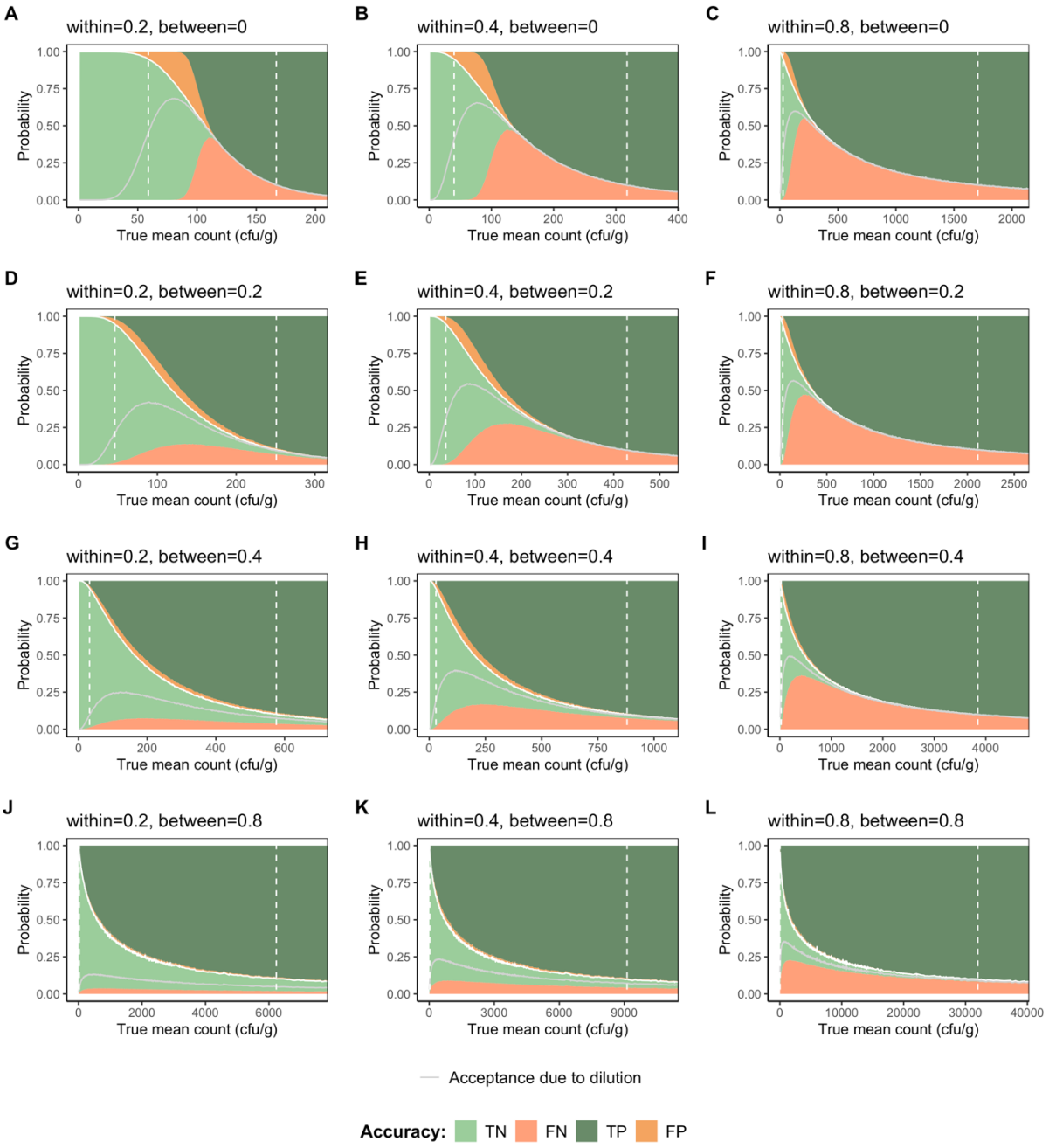

Figure 15. Classification of the sampling decision in the framework of hypothesis testing, where H_0 is that the lot bacteria mean count is < 100 CFU/g. Sampling plan: poor mixing, lot size = 50 items.

Outgoing quality

In addition to ensuring the rejection of unsafe or poor-quality batches, it is also important to monitor the concentration levels in accepted (outgoing) lots. Figures 20-21 show the examples of difference between incoming lots and accepted lots for n- and r-plans for low and high within-lot variability conditions. As expected, the mean count was lower in the accepted lots than in the lots prior to inspection in all sampling plans. This effect, however, is less pronounced when within-lot variability is low (Figure 20), as incoming lots are more consistent, and therefore the expected reduction in the variability between incoming and outgoing lots is low. On the other hand, when the variability of the incoming lots is high (Figure

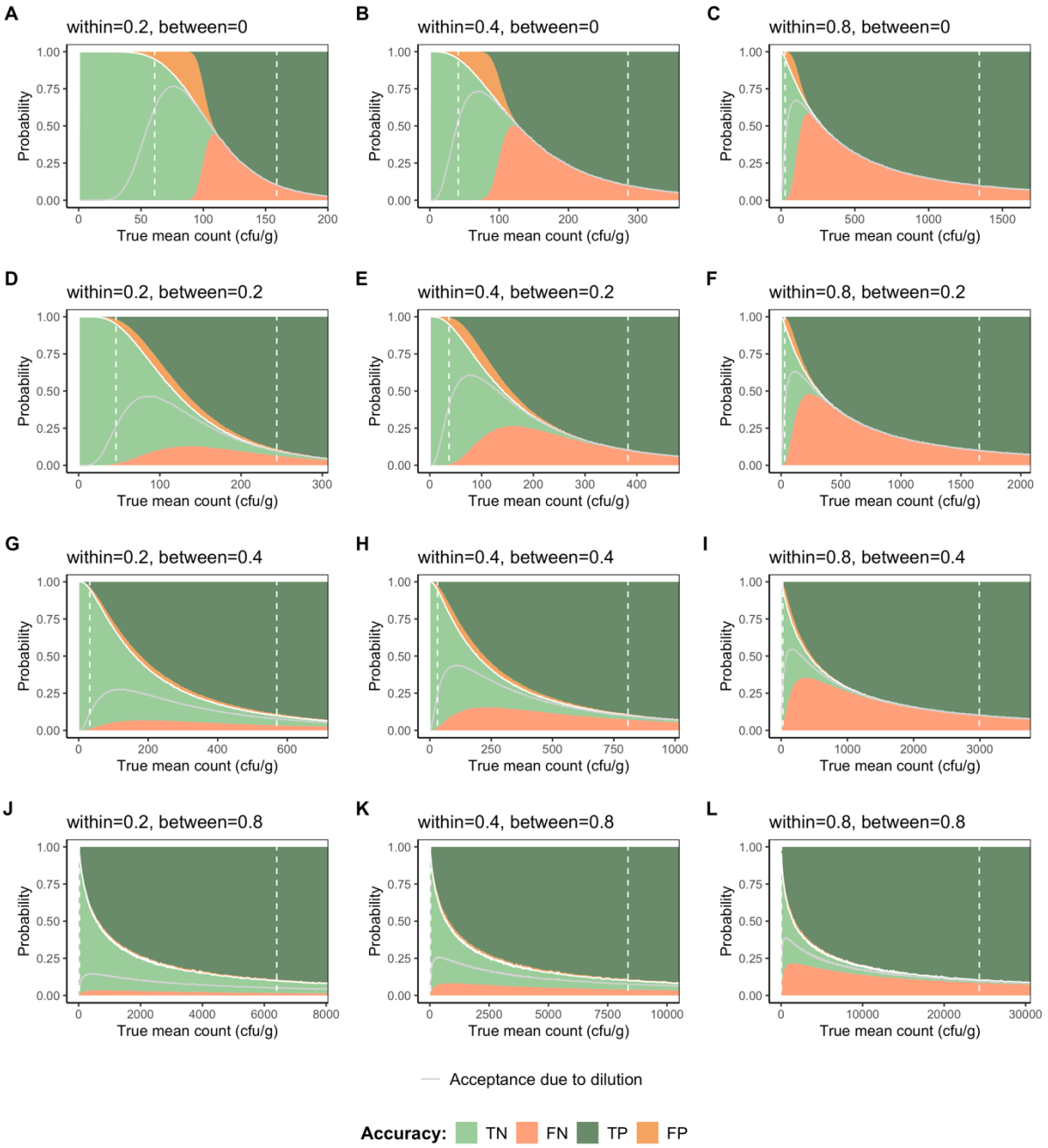

Figure 16. Classification of the sampling decision in the framework of hypothesis testing, where H_0 is that the lot bacteria mean count is < 100 CFU/g. Sampling plan: poor mixing, lot size = 100 items.

21), the reduction in variability and increase in quality are also high because the defective lots are filtered out.

A higher lot-to-lot variability leads to more significant differences in the quality levels of the accepted lots. Therefore, when the lot-to-lot variability is low, the accepted lots are closer in quality to the mean count of the incoming lots (the first column in Figures 20 and 21). In contrast, when the quality of the outgoing lots is inconsistent, the difference in the mean count between incoming and outgoing lots is higher (the third column in Figures 20 and 21). The latter case highlights underlying issues in the production process, resulting in high lot-to-lot variation. Ideally, one should first attempt to improve the production process rather than

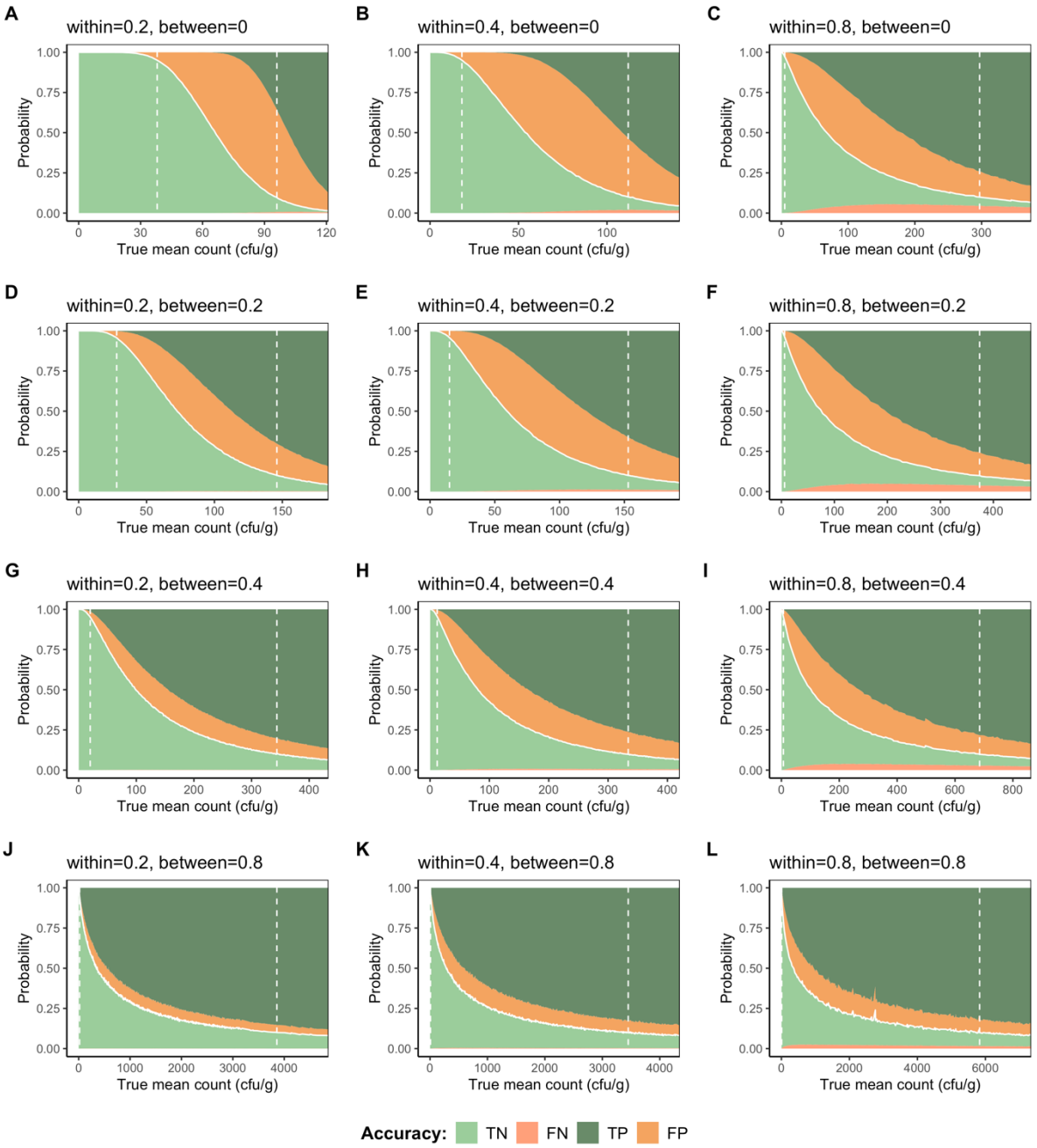

Figure 17. Classification of the sampling decision in the framework of hypothesis testing, where H_0 is that the lot bacteria mean count is < 100 CFU/g. Sampling plan: r-plan, lot size = 10 items.

make the sampling plan more stringent. Otherwise, even if outgoing lots would have sufficient quality, the rejection rate would be too high, making the sampling costs very high due to frequent re-inspections. However, if the production process cannot be improved, a very stringent sampling plan can serve as a temporary solution. In this case, the high difference between incoming and outgoing lots shows an effective filtering process, giving higher protection to consumers. Furthermore, the plots suggest that for both n- and r-sampling plans, not only high lot-to-lot variability but also medium variability will result in an expensive sampling process, as a large portion of lots has to be returned to the supplier.

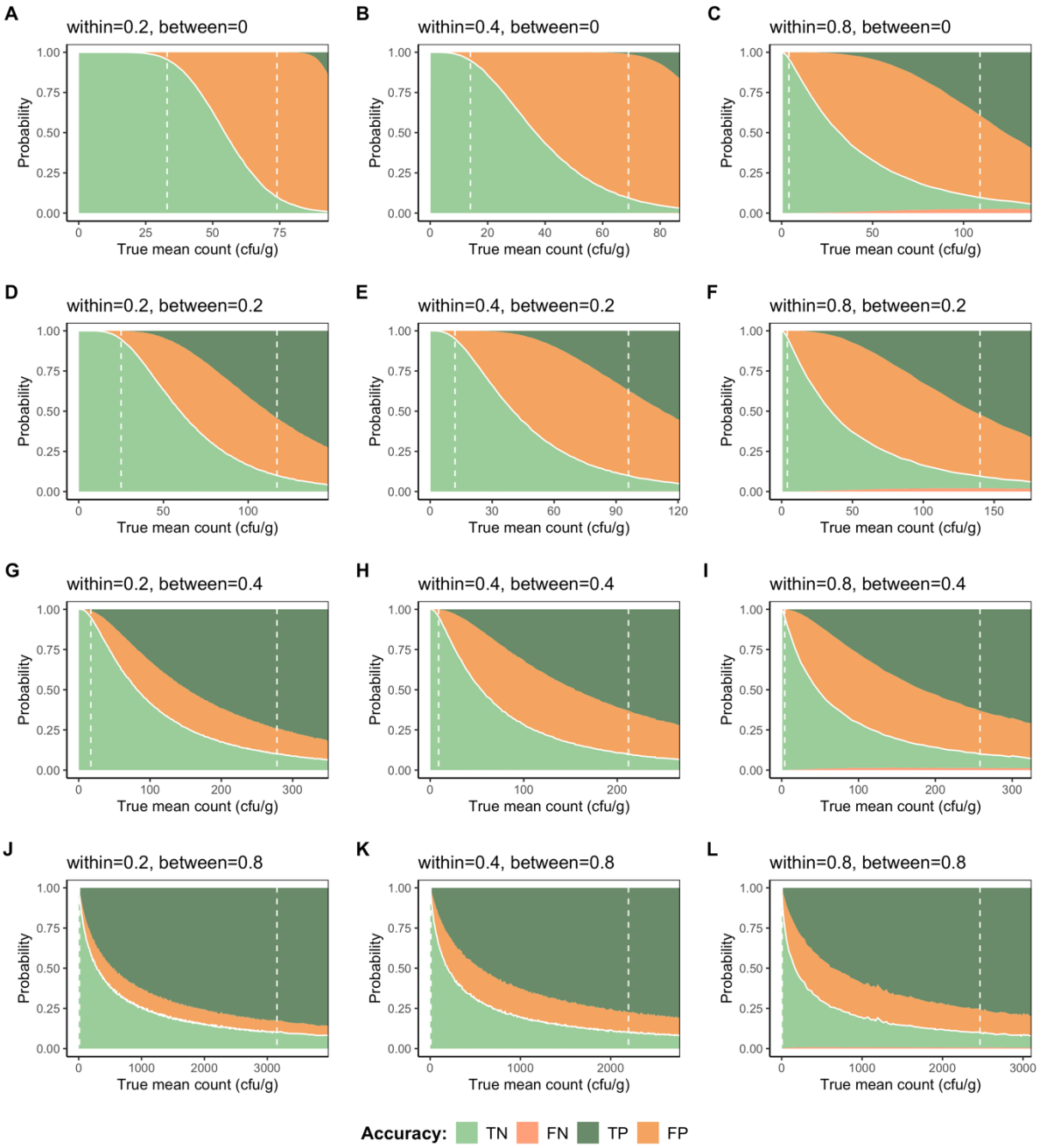

Figure 18. Classification of the sampling decision in the framework of hypothesis testing, where H_0 is that the lot bacteria mean count is < 100 CFU/g. Sampling plan: r-plan, lot size = 50 items.

Alternative plans and future research

With the exception of situations characterized by significant lot-to-lot variability (sdlog=0.8), in which modification of the manufacturing process is recommended over the adoption of a more rigorous sampling plan, the r-plan may be applied across all conditions where the n-plan does not achieve the desired LQL target or false negative rate. However, as r-plan is not a composite plan, it is quite costly owing to multiple testing especially when lot size is medium or high. Therefore, using the same simulation approach as before, the alternative composite plans were developed for medium and large lots to maintain the target LQL value and/or false negative rate. The sample size was optimized by iteratively cycling through the following steps with a limited number of simulations.

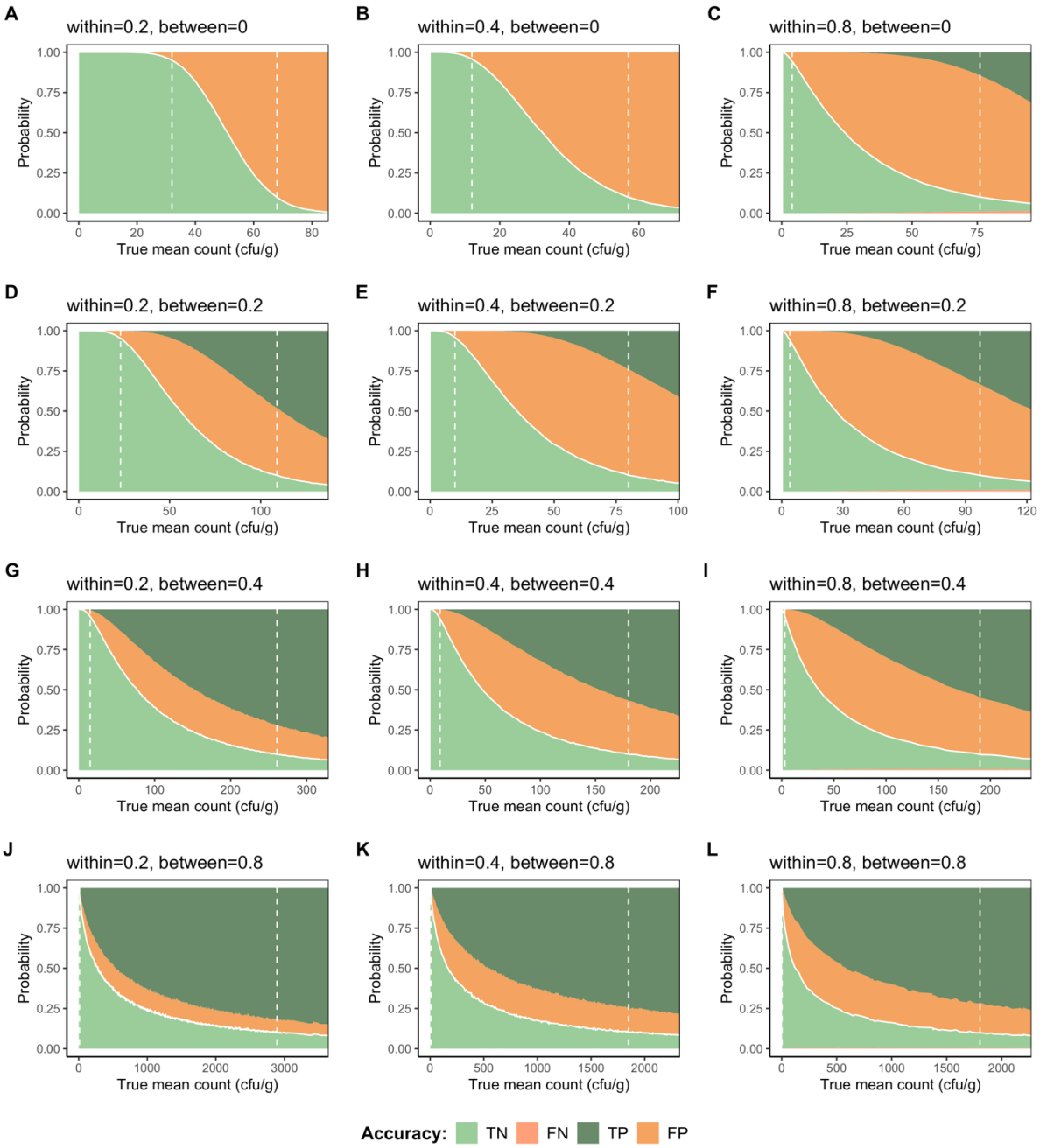

Figure 19. Classification of the sampling decision in the framework of hypothesis testing, where H_0 is that the lot bacteria mean count is < 100 CFU/g. Sampling plan: r-plan, lot size = 100 items.

(1) If desired LQL (LQL < 1000 CFU/g) and false negative rate (% FN AUC < 30%) are not reached, increase the number of primary samples by one;
(2) If desired LQL and false negative rate cannot be reached with 8 primary samples, increase the number of composite samples by one and start with 2 primary samples in each composite sample;

When the desired LQL and false negative rates were reached, the resulting plans were simulated with a higher number of simulations (due to time constrains the number of simulations was lower than 50 000 as used for the n- and r-plan simulations, therefore it is advisable to run the chosen alternative sampling plan with 50 000 simulations prior use). If the LQL was below the target of 1000 CFU/g and the false positive rate point estimate (% FN AUC) was below 30%, the plan was accepted, otherwise iterative cycling continued. For all

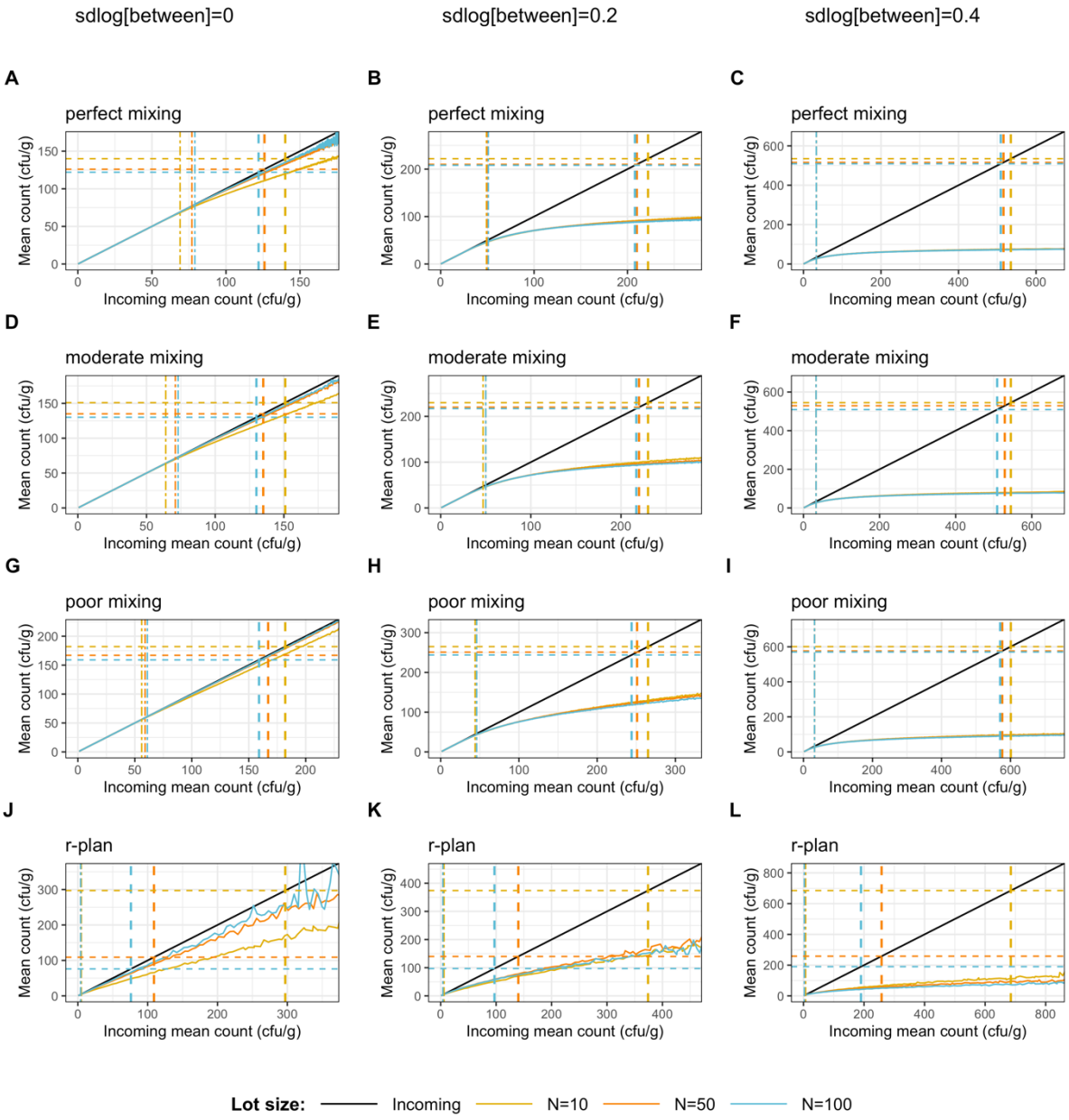

Figure 20. Incoming and outgoing quality for sampling plans with low within-lot variability.

scenarios requiring an alternative composite sampling plan, single composite sample was not sufficient to produce LQL below 1000 CFU/g, while two or at most three composite samples resulted in sufficient improvement in all of the cases. Such small number of composite samples, however, resulted in high false negative rate, which required further increase in the number of composite samples. See Table 5 for the number of composite and primary samples in the alternative composite plans.

The current project has identified several plans that can be used instead of r-plan for medium and large lots to have LQL < 1000 CFU/g and false negative rate below 30%. The same approach can be used to balance the costs of sampling and decrease the false negative rates for all other desired values of LQL, AQL, and false positive rate. Additionally, instead of having LQL target as a mean lot concentration, the target prevalence can be used. The simulation

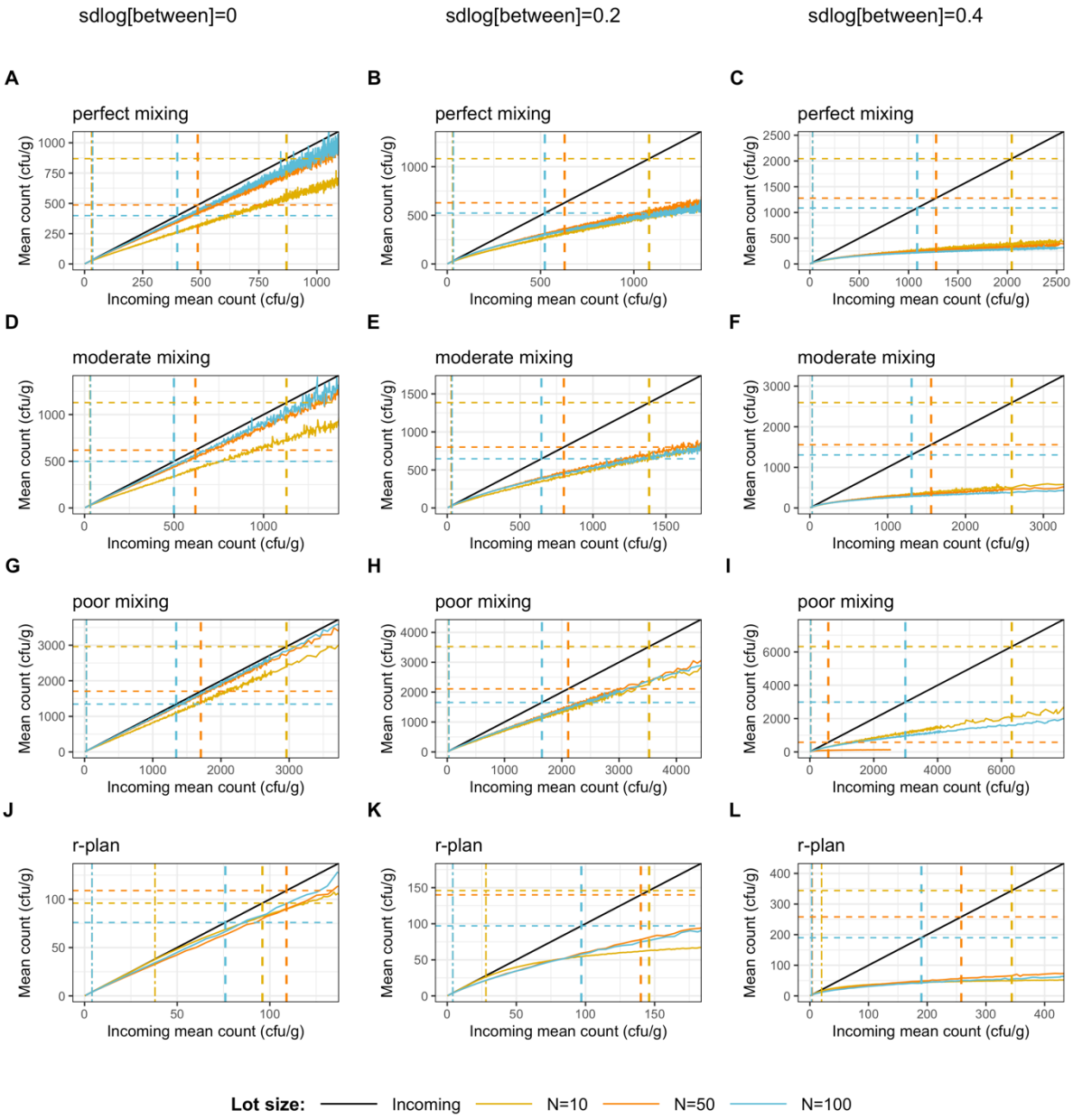

Figure 21. Incoming and outgoing quality for sampling plans with moderate within-lot variability.

approach allows to tailor the sampling plan to specific conditions and target requirements. As simulations are time-consuming, it was impossible to incorporate all these options in the current project. Additionally, future research should extend the modeling to cover the situation with heterogeneous contamination and conduct a sensitivity analysis for a range of distributions (discussed in the next section).

Possible drawbacks of the used methods

There are two main limitations to the current work. First, the current thesis is applicable only to situations where bacterial contamination is relatively homogeneous. A homogeneous sample is characterized by the uniform distribution of bacterial contamination over the whole item in the lot, such as drums. Conversely, a heterogeneous sample may have significantly

different counts of bacteria in different portions of an item. Heterogeneity is not a concern when the entire item is investigated, but this is not the case in this project. More complex modeling is required to model heterogeneous bacterial contamination.

Second, the effectiveness of the sampling plans was verified for lognormal distribution only. Although this distribution is an appropriate model, the sensitivity analysis has to be conducted for at least the gamma distribution and Weibull distribution. Owing to time constraints, a sensitivity analysis was not conducted.

Ethical thinking, societal relevance, and stakeholder awareness

Evaluating the effectiveness of the WHO's n-sampling plan requires consideration of ethics, societal impact, and stakeholder awareness. Detecting microbial contamination in pregelatinized starch, which is used as a binder in the pharmaceutical industry, is vital for patient safety. High microbial counts may indicate pathogens, posing a risk to patients if they are not detected. Adhering to ethical principles should guarantee that the sampling plan applied would not only be the most cost-efficient but also have enough discriminating power to guarantee the acceptance of uncontaminated products. Although the sampling plans in this project aim to sample hygiene indicator microorganisms, their high counts suggest the presence of pathogens. Any failure in the detection mechanism could lead to the distribution of contaminated pharmaceuticals, potentially harming patients. Any failure in the detection mechanism could lead to the distribution of contaminated pharmaceuticals, potentially harming patients. Effective microbial sampling fosters trust in pharmaceutical products. When patients and healthcare providers are confident in the safety of medications, they support better health outcomes and adherence to treatment regimens. This study had three main stakeholders:

1. Patients benefit from effective microbial sampling of pharmaceutical products. This study aimed to ensure patient safety and well-being. Cost-efficient sampling processes can also reduce the pharmaceutical costs. If the sampling process is not sufficiently stringent and patients receive contaminated pharmaceuticals, this becomes a public concern.

2. Companies such as J&J and other pharmaceutical companies want inexpensive and effective sampling plans to identify safe products that meet international standards. Good sampling plans help prevent expensive recalls, damage to reputation, and legal issues. These consumers are more aware of projects like this compared to other stakeholders.

3. Regulatory bodies may use the results of this study and other similar studies to improve future guidelines and policies for pharmaceutical safety. While regulatory bodies monitor new studies for improvements, they often react slowly to ensure the validity of scientific results owing to potentially costly consequences.

This study suggests sampling improvements to potentially benefit all stakeholders by balancing sampling efficiency and costs. By addressing ethical, societal, and stakeholder considerations, this study seeks to enhance patient safety, support pharmaceutical companies, and inform regulatory practices for the betterment of public health.

Conclusion

This study evaluated the performance of n- and r-plans for microbial acceptance sampling plans with a specification limit of 100 CFU/g for small (N=10), medium (N=50), and large (N=100) lots in the presence of several variability sources: within-lot variability, between-lot variability, and variability due to imperfect mixing). The performance of the sampling plans was evaluated as follows:

1. Visual inspection of OC curves showed that steepness, and therefore discriminating power, of n-sampling plans varies across lot sizes and variability levels. The low variability conditions and large number of samples resulted in steeper OC curves and, therefore, more efficient sampling.
2. Assessment of AQL and LQL values for producer's risk α=0.05 and consumer's risk β=0.1, which showed a substantial difference for the sampling plans. The table with AQL and LQL can be used to separate the lots based on the desired LQL target. This project has used 1000 CFU/g as a cut-off, as regulatory guidelines suggest that the mean microbial count of non-sterile products used in pharmaceuticals should be below this value. Thus, projects with LQL below 1000 CFU/g would reject the lot of such quality with a probability of 90%. The code in the supplementary information can be used to study the AQL and LQL values for different consumer and producer risks.
3. Assessment of the risk of dilution and incorrect acceptance of an unsatisfactory lot. The probability of accepting a sample containing at least one item above the specification limit was studied graphically. In all cases, n-plans pose a considerable risk of accepting such a sample, and r-plan, being a noncomposite plan, does not pose such a risk by definition. However, accepting a sample containing an outlier does not automatically mean that the accepted lot is of unsatisfactory quality. In fact, this depends on how unsatisfactory the lot quality is defined. Therefore, all acceptance decisions were classified in a hypothesis testing paradigm as FN, FP, TN, and TP (H_0: the lot mean count is above 100 CFU/g). This is a stringent classification; therefore, depending on the purpose, H_0 can use a more relaxed cut-off. The proportion of FN acceptance decisions was evaluated, and sampling plans generating many FN decisions were identified.
4. The difference between the incoming and outgoing quality was evaluated graphically. The plots suggest that all sampling plans with medium between-lot variability, including r-plan, will be potentially costly due to frequent testing of replacement lots; therefore, optimizing the manufacturing process is important.

To summarize, when the variability is relatively low (low within-lot variability, low between-lot variability, perfect or good mixing), the performance of n-plan is often satisfactory in its ability to result in LQL below 1000 CFU/g (consumer's risk β=0.1) and to have a low level of false negative decisions (where false negative is described as acceptance of the lot with mean count above 100 CFU/g). When the n-plan is unsatisfactory, the r-plan can be used instead. However, as the r-plan requires multiple individual tests, it is costly for medium and large lots. Therefore, alternative composite sampling plans can be used instead.

Table 5 provides recommendations on how many primary and composite samples have to be taken if the main goal is LQL below 1000 CFU/g. The green cells indicate that n-plan is satisfactory in both resulting LQL < 1000 CFU/g and having a relatively low rate of false positive decisions. Yellow cells indicate that n-plan does not provide neither LQL < 1000 CFU/g, neither

Table 5. Recommendation for the choice of sampling plan. The green cells indicate that n-plan is satisfactory in both resulting LQL < 1000 CFU/g and having a relatively low rate of false positive decisions. Numbers in brackets show the required numbers to have LQL < 1000 CFU/g, but provided for the reference purpose only as they results into high false negative rate. Yellow cells indicate that n-plan does not provide neither LQL < 1000 CFU/g, neither low rate of false-positive decisions. The blue cells indicate that although LQL < 1000 CFU/g is achieved, the rate of false positive decisions is high. Recommendations:
Green cells - is WHO's n-plan Yellow or blue – WHO's r-plan
Yellow and blue cells with text – alternative sampling plan specified in these cells will allow to reach LQL < 1000 CFU/g and/or false negative rate < 30%.

Variability (sdlog)

Within n-lot	Between-lot	Perfect or good mixing			Moderate mixing			Poor mixing		
		N=10	N=50	N=100	N=10	N=50	N=100	N=10	N=50	N=100
0.2	0									
	0.2									
	0.4									
0.4	0									
	0.2								2 x n=2	2 x n=2
	0.4								2 x n=2	2 x n=2
0.8	0		2 x n=7	2 x n=8		3 x n=4	3 x n=4		4 x n=5 (2 x n=3)	4 x n=5 (2 x n=3)
	0.2		2 x n=7	2 x n=8		3 x n=4	3 x n=4		4 x n=5 (2 x n=5)	4 x n=5 (2 x n=5)
	0.4		2 x n=7 (2 x n=4)	2 x n=8 (2 x n=4)		3 x n=4 (2 x n=5)	3 x n=4 (2 x n=6)		4 x n=5 (3 x n=6)	4 x n=5 (3 x n=6)

low rate of false-positive decisions. The blue cells indicate that although LQL < 1000 CFU/g is achieved, the rate of false positive decisions is high. Therefore, the recommended sampling plan for green cells is WHO's n-plan; for empty yellow and blue cells, WHO's r-sampling plan is recommended; if yellow or blue cells are not empty, this indicates that either the r-plan has to be applied, or an alternative composite sampling plan with an indicated number of composite and primary samples (to reach both LQL < 100CFU/g and false negative rate < 30%).

The future research should extend the modeling to cover heterogeneous contamination and validate the sensitivity of the sampling plans against the broader range of distributions.

References

Brown, G. H., & Fisher, N. I. (1972). Subsampling a Mixture of Sampled Material. *Technometrics*, *14*(3), 663. https://doi.org/10.2307/1267294

Elder, R. S., Thompson, W. O., & Myers, R. H. (1980). Properties of Composite Sampling Procedures. *Technometrics*, *22*, 179–186. https://api.semanticscholar.org/CorpusID:122303752

EPA. (1995). EPA Observational Economy Series, Volume 1: Composite Sampling. In *EPA-230-R-95-005*.

European Directorate for the Quality of Medicines & HealthCare. (2023). *European Pharmacopoeia* (11th ed.). European Directorate for the Quality of Medicines & HealthCare.

Hund, L. (2014). New tools for evaluating LQAS survey designs. *Emerging Themes in Epidemiology*, *11*(1), 2. https://doi.org/10.1186/1742-7622-11-2

Izenman, A. J. (2001). Statistical and Legal Aspects of the Forensic Study of Illicit Drugs. *Statistical Science*, *16*(1), 35–57. https://doi.org/10.1214/ss/998929475

Montgomery, D. C. (2009). *Introduction to Statistical Quality Control* (6th ed.). John Wiley & Sons.

Patil, G. P., Gore, S. D., & Taillie, C. (2011). Composite Sampling. In *Composite Sampling*. Springer US. https://doi.org/10.1007/978-1-4419-7628-4

Rohde, C. A. (1976). Composite Sampling. *Biometrics*, *32*(2), 273. https://doi.org/10.2307/2529498

Samohyl, R. W. (2018). Acceptance sampling for attributes via hypothesis testing and the hypergeometric distribution. *Journal of Industrial Engineering International*, *14*(2), 395–414. https://doi.org/10.1007/s40092-017-0231-9

Santos-Fernández, E., Govindaraju, K., & Jones, G. (2015). Variables sampling plans using composite samples for food quality assurance. *Food Control*, *50*, 530–538. https://doi.org/10.1016/J.FOODCONT.2014.09.041

Saranadasa, H. (2003). The Square Root of N Plus One Sampling Rule How Much Confidence Do We Have. *Pharmaceutical Technology*, *27*, 50–62. https://api.semanticscholar.org/CorpusID:117858087

United States Pharmacopeial Convention. (2009). *United States Pharmacopeia and National Formulary (USP 32–NF 27)* (32nd ed.). United States Pharmacopeial Convention.

World Health Organization. (2005). TRS 929 - Annex 4: WHO guidelines for sampling of pharmaceutical products and related materials. *WHO Technical Report Series*, *925*.

Supplementary r-code

```
# Required packages:
library(ggplot2)   library(ggpubr)   library(reshape2)   library(gtools)
library(dplyr)     library(tidyr)    library(pracma)

# Function creates a composite sampling plan and the metrics of its performance
attribute_plan.composite <- function(N, n, sdlog=0.8, sdlog.batch=0, limit=100,
        mixing = c("perfect", "good", "moderate", "poor"),
        n.sim=5000, seed=123) {
  set.seed(seed)
  df <- data.frame(means = c(seq(0.1, 1, by=0.1), seq(2, 500, by=1),
  seq(510, 1500, by=10)), p.accept = NA, concentration.incoming = NA,
  concentration.accepted = NA, outliers.undetected = NA,
  TP.conc = NA, FP.conc = NA, TN.conc = NA, FN.conc = NA)
  for (i in 1:length(df$means)) {
    p.nonconform.accepted <- conc.accepted <- outliers <- rep(NA, n.sim)
    accepted <- conc.incoming <- accurate.conc  <- rep(NA, n.sim)
    for (sim in 1:n.sim) {
      batch_mean <-  rnorm(1, mean=log(df$means[i], base=10), sd=sdlog.batch)
      batch <- 10^(rnorm(n=N, mean=batch_mean, sd=sdlog))
      sample <- sample(x=batch, size=n, replace = FALSE)
      if (mixing == "perfect") {sample_mean = mean(sample)}
      else {
        if (mixing == "good") {mixing = 5}
        if (mixing == "moderate") {mixing = 1}
        if (mixing == "poor") {mixing = 0.1}
        Dirichlet.weights <- rdirichlet(1, rep(mixing, n))
        sample_mean <- sum(sample*Dirichlet.weights)
      }
      conc.incoming[sim] <- mean(batch)

      if (sample_mean < limit) {
        conc.accepted[sim] <- mean(batch)
        accepted[sim] <- 1
        outliers[sim] <- any(sample >= limit)
        if (mean(batch) < limit) {accurate.conc[sim] <- "TN"}
        else {accurate.conc[sim] <- "FN"}
         }
      else {
        accepted[sim] <- 0
        if (mean(batch) < limit) {accurate.conc[sim] <- "FP"}
        else {accurate.conc[sim] <- "TP"}
      }
    }
    df$concentration.incoming[i] <- mean(conc.incoming)
    df$p.accept[i] <- mean(accepted)

    df$TP.conc[i] <- sum(accurate.conc=="TP")/n.sim
    df$FP.conc[i] <- sum(accurate.conc=="FP")/n.sim
    df$TN.conc[i] <- sum(accurate.conc=="TN")/n.sim
    df$FN.conc[i] <- sum(accurate.conc=="FN")/n.sim

    if (any(!is.na(conc.accepted))) {
      df$concentration.accepted[i] <- mean(conc.accepted, na.rm = TRUE)
      df$outliers.undetected[i] <- sum(outliers, na.rm = TRUE) / n.sim
    }
    else { df$outliers.undetected[i] <- 0 }
  }
  return(df)
```

```
}

# Function creates a noncomposite sampling plan and the metrics of its performance
attribute_plan.noncomposite <- function(N, n, sdlog=0.8, sdlog.batch=0,
         limit=100, n.sim=5000, seed=123) {
  set.seed(seed)
  df <- data.frame(means = c(seq(0.1, 1, by=0.1), seq(2, 500, by=1),
         seq(510, 1500, by=10)), p.accept = NA, concentration.incoming = NA,
         concentration.accepted = NA, TP.conc = NA, FP.conc = NA,
         TN.conc = NA, FN.conc = NA)
  for (i in 1:length(df$means)) {
    p.nonconform.accepted <- conc.accepted <- accepted <- rep(NA, n.sim)
    conc.incoming <- accurate.conc  <- rep(NA, n.sim)
    for (sim in 1:n.sim) {
      batch_mean <-  rnorm(1, mean=log(df$means[i], base=10), sd=sdlog.batch)
      batch <- 10^(rnorm(n=N, mean=batch_mean, sd=sdlog))
      sample <- sample(x=batch, size=n, replace = FALSE)

      conc.incoming[sim] <- mean(batch)

      if (all(sample < limit)) {
        conc.accepted[sim] <- mean(batch)
        accepted[sim] <- 1
        if (mean(batch) < limit) {accurate.conc[sim] <- "TN"}
        else {accurate.conc[sim] <- "FN"}
         }
      else {
        accepted[sim] <- 0
        if (mean(batch) < limit) {accurate.conc[sim] <- "FP"}
        else {accurate.conc[sim] <- "TP"}
      }
    }
    df$concentration.incoming[i] <- mean(conc.incoming)
    df$p.accept[i] <- mean(accepted)

    df$TP.conc[i] <- sum(accurate.conc=="TP")/n.sim
    df$FP.conc[i] <- sum(accurate.conc=="FP")/n.sim
    df$TN.conc[i] <- sum(accurate.conc=="TN")/n.sim
    df$FN.conc[i] <- sum(accurate.conc=="FN")/n.sim

    if (any(!is.na(conc.accepted))) {
      df$concentration.accepted[i] <- mean(conc.accepted, na.rm = TRUE)
    }
  }
  return(df)
}

# Function that creates a composite sampling plan with k composite
 and n primary samples
attribute_plan.k_composites <- function(N, n, k, sdlog=0.8, sdlog.batch=0,
         limit=100, mixing = c("perfect", "good", "moderate", "poor"),
         n.sim=5000, seed=123) {
  set.seed(seed)
  df <- data.frame(means = c(seq(0.1, 1, by=0.1), seq(2, 500, by=1),
          seq(510, 1500, by=10)), p.accept = NA, concentration.incoming = NA)
  for (i in 1:length(df$means)) {
    accepted <- conc.incoming <- rep(NA, n.sim)
    for (sim in 1:n.sim) {
      batch_mean <-  rnorm(1, mean=log(df$means[i], base=10), sd=sdlog.batch)
```

```
      batch <- 10^(rnorm(n=N, mean=batch_mean, sd=sdlog))
      samples <- sample(x = batch, size = n*k, replace = FALSE)
      comp.samples <- as.data.frame(split(samples, cut(seq_along(samples), k,
                                                       labels = FALSE)))

      if (mixing == "perfect") {
        samples_mean = colMeans(comp.samples)
        }
      else {
        if (mixing == "good") {mixing = 5}
        if (mixing == "moderate") {mixing = 1}
        if (mixing == "poor") {mixing = 0.1}
        Dirichlet_weights <- function(column) {
          Dirichlet.weights <- rdirichlet(1, rep(mixing, n))
          return(sum(column * Dirichlet.weights))
        }
        samples_mean <- apply(comp.samples, 2, Dirichlet_weights)
      }
      conc.incoming[sim] <- mean(batch)

      if (sample_mean < limit) {
        conc.accepted[sim] <- mean(batch)
        accepted[sim] <- 1
        outliers[sim] <- any(sample >= limit)
        if (mean(batch) < limit) {accurate.conc[sim] <- "TN"}
        else {accurate.conc[sim] <- "FN"}
         }
      else {
        accepted[sim] <- 0
        if (mean(batch) < limit) {accurate.conc[sim] <- "FP"}
        else {accurate.conc[sim] <- "TP"}
      }
    }

    df$concentration.incoming[i] <- mean(conc.incoming)
    df$p.accept[i] <- mean(accepted)

    df$TP.conc[i] <- sum(accurate.conc=="TP")/n.sim
    df$FP.conc[i] <- sum(accurate.conc=="FP")/n.sim
    df$TN.conc[i] <- sum(accurate.conc=="TN")/n.sim
    df$FN.conc[i] <- sum(accurate.conc=="FN")/n.sim
  }
  return(df)
}

# Function that find mean count corresponding to LQL and AQL
find.AQL_LQL <- function(plan, beta = 0.1, alpha = 0.05) {
  df <- data.frame(means = plan$concentration.incoming,
         accept_probs = plan$p.accept)
  AQL <- LQL <- NA
  if (min(df$accept_probs) <= beta) {
    df$differences <- abs(df$accept_probs - beta)
    LQL <- df$means[which(df$differences == min(df$differences))]
    if (length(LQL) != 1) {   LQL <- mean(LQL) }
    LQL <- round(LQL)
  }
  if (max(df$accept_probs) >= (1 - alpha)) {
    df$differences <- abs(df$accept_probs - (1 - alpha))
    AQL <- df$means[which(df$differences == min(df$differences))]
```

```
    if (length(AQL) != 1) {   AQL <- mean(AQL)   }
    AQL <- round(AQL)
  }
  return(c(AQL, LQL))
}

# Function that find mean count corresponding to LQL and AQL
on the geometric scale (helper function used in graphs)
find.AQL_LQL_geo <- function(plan, beta=0.1, alpha=0.05) {
  df <-data.frame(means = plan$means, accept_probs = plan$p.accept)
  AQL <- LQL <- NA
  if (min(df$accept_probs) <= beta) {
    df$differences <- abs(df$accept_probs - beta)
    LQL <- df$means[which(df$differences == min(df$differences))]
    if (length(LQL) != 1) {   LQL <- mean(LQL) }
    LQL <- round(LQL)
  }
  if (max(df$accept_probs) >= (1 - alpha)) {
    df$differences <- abs(df$accept_probs - (1 - alpha))
    AQL <- df$means[which(df$differences == min(df$differences))]
    if (length(AQL) != 1) {   AQL <- mean(AQL)   }
    AQL <- round(AQL)
  }
  return(c(AQL, LQL))
}

# Function that find mean count corresponding to LQL only
 (helper function used in graphs)
find.LQL <- function(plan, beta = 0.1) {
  df <- data.frame(means = plan$concentration.incoming,
         accept_probs = plan$p.accept)
  LQL <- NA
  if (min(df$accept_probs) <= beta) {
    df$differences <- abs(df$accept_probs - beta)
    LQL <- df$means[which(df$differences == min(df$differences))]
    if (length(LQL) != 1) {   LQL <- mean(LQL) }
    LQL <- round(LQL)
  }
  return(LQL)
}

# Function that plots 3 OC curves corresponding to 3 lot sizes on the same plot
 and draw dashed lines corresponding to AQL and LQL
OC_3sizes <- function(plan1, plan2, plan3, graph_label="",
                      beta=0.1, alpha=0.05, upper_limit = NULL) {
  LQL_AQL1 <- find.AQL_LQL(plan=plan1, beta=beta, alpha=alpha)
  LQL_AQL2 <- find.AQL_LQL(plan=plan2, beta=beta, alpha=alpha)
  LQL_AQL3 <- find.AQL_LQL(plan=plan3, beta=beta, alpha=alpha)
  if (is.null(upper_limit)) {
    upper_limit <- max(c(LQL_AQL1[2], LQL_AQL2[2], LQL_AQL3[2]),
           na.rm = TRUE)*1.2
  }
  plot <- ggplot() +
    geom_line(data=plan1, aes(x = concentration.incoming, y = p.accept,
         color = "N=10"))+
    geom_line(data=plan2, aes(x = concentration.incoming, y = p.accept,
         color = "N=50"))+
    geom_line(data=plan3, aes(x = concentration.incoming, y = p.accept,
         color = "N=100"))+
```

```
    geom_vline(xintercept = LQL_AQL1,  linetype = "dashed",
         color = "#E1AF00", size=0.5) +
    geom_vline(xintercept = LQL_AQL2,  linetype = "dashed",
         color = "#F98400", size=0.5) +
    geom_vline(xintercept = LQL_AQL3,  linetype = "dashed",
         color = "#5BBCD6", size=0.5) +
    theme_bw() +
    theme(legend.position="bottom", legend.box = "vertical",
          plot.margin = margin(15,15,15,15, "pt"),
          legend.text = element_text(size = 12),
          legend.title = element_text(size = 13, face = "bold"),
          legend.key.size = unit(50, 'pt'),
          panel.border = element_rect(colour = "black", fill=NA),) +
    scale_color_manual(name = "Lot size:",
                       breaks = c("N=10", "N=50", "N=100"),
                       values = c("N=10" = "#E1AF00", "N=50" = "#F98400",
                                  "N=100" = "#5BBCD6")) +
    labs(title = graph_label,
       x = "Mean count (cfu/g)",
       y = "Probability of acceptance")+
    coord_cartesian(xlim = c(0, upper_limit))
  return(plot)
}

# Function that plot real OC curve, how would OC curve look, if samples with
 outliers unaccepted, and absolute probability to accept a sample due to dilution
dilution.plot <- function(plan, graph_label="") {
  AQL_LQL <- find.AQL_LQL(plan)
  plan.ideal <- plan
  plan.ideal$p.accept <- plan$p.accept -plan$outliers.undetected
  AQL_LQL.ideal <- find.AQL_LQL(plan.ideal)
  upper_limit <- AQL_LQL[2]*1.2
  concentration_accuracy <- ggplot() +
    geom_line(data=plan, aes(x = concentration.incoming, y = p.accept,
                             color = "P(Acceptance)"))+
    geom_line(data=plan, aes(x = concentration.incoming, y = outliers.undetected,
                             color = "P(Acceptance due to dilution)"))+
    geom_line(data=plan, aes(x = concentration.incoming,
                              y = p.accept - outliers.undetected,
                             color = "Ideal P(Acceptance)"))+
    geom_vline(xintercept = AQL_LQL, color = "#5BBCD6", linetype = "dashed",
          size = 0.5) +
    geom_vline(xintercept = AQL_LQL.ideal,
               color = "#F98400", linetype = "dashed", size = 0.5) +
    theme_bw() +
    theme(legend.position="bottom", legend.box = "vertical",
          plot.margin = margin(15,15,15,15, "pt"),
          legend.text = element_text(size = 12),
          panel.border = element_rect(colour = "black", fill=NA)) +
    scale_color_manual(name = "",
                       values = c("P(Acceptance)" = "#5BBCD6",
                                  "P(Acceptance due to dilution)" = "tomato3",
                                  "Ideal P(Acceptance)" = "#F98400")) +
    labs(title = graph_label,
       x = "True mean count (cfu/g)",
       y = "Probability")+
    coord_cartesian(xlim = c(0, upper_limit))

  return(concentration_accuracy)
```

```
}

# Function that plots all acceptance/rejection decisions as FN, FP, TN, TP
accuracy_concentration.plot <- function(plan, graph_label="") {
  AQL_LQL <- find.AQL_LQL(plan)
  conc.df <- plan[, c("concentration.incoming", "TP.conc", "FP.conc",
         "TN.conc", "FN.conc")]
  colnames(conc.df) <- c("concentration.incoming", "TP", "FP", "TN", "FN")
  accuracy_conc <- melt(conc.df, id.vars = "concentration.incoming",
                        variable.name = "Accuracy", value.name = "Percentage")
  upper_limit <- AQL_LQL[2]*1.2
  concentration_accuracy <- ggplot() +
    geom_area(data=accuracy_conc, aes(x = concentration.incoming, y = Percentage,
                                      fill = Accuracy),
              position = "stack") +
    geom_line(data=plan, aes(x = concentration.incoming, y = p.accept),
              color = "white")+
    geom_line(data=plan, aes(x = concentration.incoming, y = outliers.undetected,
              color = "Acceptance due to dilution"))+
    geom_vline(xintercept = AQL_LQL,  linetype = "dashed", color = "white") +
    theme_classic() +
    theme(legend.position="bottom", legend.box = "vertical",
          plot.margin = margin(15,15,15,15, "pt"),
          legend.text = element_text(size = 12),
          legend.title = element_text(size = 13, face = "bold"),
          panel.border = element_rect(colour = "black", fill=NA),) +
    scale_color_manual(name = " ",
                       values = c("Acceptance due to dilution" = "lightgrey")) +
    scale_fill_manual(name = "Accuracy:", breaks = c("TN", "FN", "TP", "FP"),
           values = c("darkseagreen3", "lightsalmon",
           "darkseagreen4", "sandybrown")) +
    guides(color = guide_legend(order=1),
         fill = guide_legend(order=2)) +
    labs(title = graph_label,
       x = "True mean count (cfu/g)",
       y = "Probability")+
    coord_cartesian(xlim = c(0, upper_limit))

  return(concentration_accuracy)
}

# Function that plots quality of incoming and accepted lot vs quality of incoming
lots
 AOQ_3sizes <- function(plan1, plan2, plan3, graph_label="", beta=0.1, alpha=0.05)
{
  AQL_LQL1 <- find.AQL_LQL(plan=plan1, beta=beta, alpha=alpha)
  AQL_LQL2 <- find.AQL_LQL(plan=plan2, beta=beta, alpha=alpha)
  AQL_LQL3 <- find.AQL_LQL(plan=plan3, beta=beta, alpha=alpha)
  AQL_LQL_geo1 <- find.AQL_LQL_geo(plan=plan1, beta=beta, alpha=alpha)
  AQL_LQL_geo2 <- find.AQL_LQL_geo(plan=plan2, beta=beta, alpha=alpha)
  AQL_LQL_geo3 <- find.AQL_LQL_geo(plan=plan3, beta=beta, alpha=alpha)
  upper_limit <- max(c(AQL_LQL1[2], AQL_LQL2[2], AQL_LQL3[2]), na.rm = TRUE)*1.2
  plot <- ggplot() +
  geom_line(data=plan1, aes(x = concentration.incoming,
           y = concentration.incoming,
                             color = "Incoming")) +
  geom_line(data=plan2, aes(x = concentration.incoming,
            y = concentration.incoming,
                             color = "Incoming")) +
```

```
  geom_line(data=plan3, aes(x = concentration.incoming,
            y = concentration.incoming,
                              color = "Incoming")) +
  geom_line(data=plan1, aes(x = concentration.incoming,
            y = concentration.accepted,
                              color = "N=10")) +
  geom_line(data=plan2, aes(x = concentration.incoming,
           y = concentration.accepted,
                              color = "N=50")) +
  geom_line(data=plan3, aes(x = concentration.incoming,
           y = concentration.accepted,
                              color = "N=100")) +
  geom_vline(xintercept = AQL_LQL1[1],  linetype = "twodash",
             color = "#E1AF00", size=0.5) +
  geom_vline(xintercept = AQL_LQL2[1],  linetype = "twodash",
             color = "#F98400", size=0.5) +
  geom_vline(xintercept = AQL_LQL3[1],  linetype = "twodash",
             color = "#5BBCD6", size=0.5) +
  geom_vline(xintercept = AQL_LQL1[2],  linetype = "dashed",
             color = "#E1AF00", size=0.75) +
  geom_vline(xintercept = AQL_LQL2[2],  linetype = "dashed",
             color = "#F98400", size=0.75) +
  geom_vline(xintercept = AQL_LQL3[2],  linetype = "dashed",
             color = "#5BBCD6", size=0.75) +
  geom_hline(yintercept = AQL_LQL1[2],  linetype = "dashed",
             color = "#E1AF00", size=0.5) +
  geom_hline(yintercept = AQL_LQL2[2],  linetype = "dashed",
             color = "#F98400", size=0.5) +
  geom_hline(yintercept = AQL_LQL3[2],  linetype = "dashed",
             color = "#5BBCD6", size=0.5) +
  theme_bw() +
  theme(legend.position = "bottom",
        legend.box = "vertical",
        plot.margin = margin(15, 15, 15, 15, "pt"),
        legend.text = element_text(size = 12),
        legend.title = element_text(size = 13, face = "bold"),
        legend.key.size = unit(50, 'pt'),
        panel.border = element_rect(colour = "black", fill = NA)) +
  scale_color_manual(name = "Lot size:",
                       breaks = c("Incoming", "N=10", "N=50", "N=100"),
                     values = c("Incoming" = "black", "N=10" = "#E1AF00",
                                "N=50" = "#F98400", "N=100" = "#5BBCD6")) +
  labs(title = graph_label,
       x = "Incoming mean count (cfu/g)",
       y = "Mean count (cfu/g)") +
  coord_cartesian(xlim = c(0, upper_limit), ylim = c(0, upper_limit))
  return(plot)
}

 # Function that finds relative % FN AUC
 find.AUC <- function(plan) {
  AQL_LQL <- find.AQL_LQL_geo(plan)
  filtered_plan <- subset(plan, means <= AQL_LQL[2])
  df <- filtered_plan[order(filtered_plan$concentration.incoming), ]
  AUC.FN <- trapz(df$concentration.incoming, df$FN.conc)
  AUC.OC <- trapz(df$concentration.incoming, df$p.accept)
  AUC.relative <- AUC.FN/AUC.OC
  return(AUC.relative)
}
```