UNIVERSITY OF CALIFORNIA, SAN DIEGO UCSD

BERKELEY • DAVIS • IRVINE • LOS ANGELES • MERCED • RIVERSIDE • SAN DIEGO • SAN FRANCISCO SANTA BARBARA • SANTA CRUZ

Department of Economics, 0508

9500 GILMAN DRIVE
LA JOLLA, CALIFORNIA 92093-0508

April 1, 2012

David A. Eisenberg, President
Micro-Tracers, Inc.
1370 Van Dyke Avenue
San Francisco, California 94124

Mr. Eisenberg,

This letter is in response to your request for an overview and assessment of general statistical methods for testing the level of homogeneity in a mixture of feed based on tracer particle distribution estimates. I have reviewed the literature you provided, particularly your 1976 paper "The Use of Microtracers in determining the uniformity of formula animal feeds" and the 2010 report produced by TNO Science and Industry. I have attached a copy of my full report which goes into greater detail than the summary presented here.

The procedure you outline in your paper, and the probability theory explained therein, is accurate and appropriate for testing mix homogeneity. The use of both Poisson and chi-squared tests is theoretically justified based on the assumptions and conditions of feed mixing. I think it is clear from your writing, and from our conversations, that you have a firm grasp on how every aspect of the mixing process may impact the validity and precision of the test results.

From my perspective, it seems that the following questions are of particular interest when it comes to the use of micro-tracer particles in the feed mixing process:

1. How many samples should be taken from the mix for analysis?
2. How many particles should be added to the mix initially in order to ensure accurate test results?
3. How big should each sample be (measured in grams)?
4. From where in the mix should the samples be taken?
5. How many sub-samples should be taken from each sample?

The theme that surfaced in my exploration of these issues is that there is no single answer to any of these questions. As with most statistical testing and inference, the number of samples is very important since it drives convergence results and indicates whether our tests are appropriate for interpreting the data. It follows that, since many samples dominate few samples, the answer to most of these questions should be "as many as possible". However, I think it is paramount that we consider the answer to a more fundamental question first: How complete does the mix need to be?

It is true that testing more samples or adding more particles can only improve our results, but this improvement may be irrelevant in many cases. I believe, based on the context in which this testing is done, that the statistical theory should not be used to formulate policies concerning marginal changes in sample size, *et cetera*. Instead, the probability theory would be better suited to establishing upper and lower bounds on the parameters of our procedure. As long as we are within these bounds, it may be best to let economic factors—such as the cost of mixing, the cost of ingredients or the costs of producing an incomplete mix—determine how these parameters are fine tuned. It is for this reason that I suggest a more intuitive approach to testing procedures over the analytically complex methods suggested by the TNO report.

With this in mind, I recommend testing no fewer than 10 samples of about 100g each, containing 50 to 100 particles that have been taken from several locations within the mixer at several different periods. More precision can be gained by increasing these numbers; more efficiency (lower testing cost) is gained by lowering them. Reconciling the tradeoff between these options is ultimately the biggest problem to be solved. The use of sub-samples will not greatly improve our measurements, and should only be used if the cost of doing so is close to zero.

I don’t think any of these recommendations are different from what you currently hold to be true.

It has been a pleasure working with you on this project. Your mastery of the industry, and the theories on which it is built, is very impressive and I have found this subject matter to be very intellectually stimulating. Please do not hesitate to contact me at any time if I can be of further assistance, or if anything here or in my report is unclear.

Sincerely,

David Bernotas, Ph.D.

On the Use of Particulate Distributions for Determining Degree of Homogeneity in a Feed Mixture

David Bernotas, Ph.D.
Department of Economics
University of California, San Diego
April 1, 2012

I have been asked by Microtracers Inc. to summarize appropriate statistical methods for determining how well a batch of feed is mixed based on the estimated distribution of tracer particles.[1] The seminal bit of literature on which this is based was produced by David Eisenberg of Microtracers Inc. in a document circa 1976 called "The Use of Microtracers in Determining Uniformity of Formula Animal Feeds" where it is proposed (correctly) that a Poisson distribution in conjunction with a Pearson's chi-squared test can be used to evaluate the completeness (homogeneity) of the mixture. Although the statistical theory underlying the problem is not in question, some issues do arise in the context of retrieving microtracers from a mixture of feed to determine the degree of homogeneity of the mix. For example:

1. How many samples should be taken from the mix for analysis?
2. How many particles should be added to the mix initially in order to ensure accurate results?
3. How big should each sample be (measured in grams)?
4. From where should the samples be taken?
5. How many sub-samples should be taken from each sample?

I will offer an answer to each of these. I say "offer an answer" because there are several viable ways to address these issues–each has its own costs and benefits–and the best alternative will depend on the situation and preferences of the individual at that time. I will try my best to highlight the trade-off between cost and benefit for each option.

1 Probability basics

This section and its sub-sections are provided for completeness and to develop consistency in notation.

[1] Terminology turns out to be delicate here. A mix is "complete" if it meets the standards and preferences of the person mixing based on the context. "Good" and "bad" are less precisely defined, but I will use "good" and "complete" interchangeably, likewise with "bad" and "incomplete". "Homogeneous" means that the mix is as random as it possibly can be–technically this is called "maximum entropy". Thus, from most mixed to least:
”*homogeneous*” > ”*complete*” = ”*good*” > ”*incomplete*” = ”*bad*”.

1.1 Poisson Distributions

The Poisson distribution is appropriate for modeling the distribution of rare, discrete events during a period of time, distance or volume. Classic examples are the number of telephone calls per hour (time), the number of mutations on a strand of DNA (distance) and the clustering of galaxies in the universe (volume). (See Hammersly (1972), Saslaw (1989) and Hope (1979) for examples and rigorous treatment.) The location of particles in a homogeneous feed mixture will similarly be distributed as a Poisson process. Recall the properties of a Poisson distribution:

$$prob(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{1}$$

where k is the number of observed particles and λ is both the expected number of particles (mean) and the variance. It follows that the coefficient of variation (CV), which relates the variability of observations to the mean of the distribution, is given by the following:

$$CV = \frac{\sqrt{\lambda}}{\lambda} = \frac{1}{\sqrt{\lambda}} \approx \frac{1}{\sqrt{\overline{X}}}. \tag{2}$$

Skewness and kurtosis (peakedness) of the Poisson distribution are also determined entirely by the parameter λ.

$$\begin{aligned} skewness &= \frac{1}{\sqrt{\lambda}} \\ kurtosis &= 3 + \frac{1}{\lambda} \end{aligned} \tag{3}$$

Since $\lambda \geq 0$, skewness values are always positive (the distribution leans left of center) and kurtosis values are always ≥ 3 (the distribution is taller than a normal distribution).

Since the mean, variance, skewness and peakedness of the Poisson are all characterized by the value λ, this parameter tells us everything we need to know about the shape and location of the Poisson distribution. If λ is sufficiently large ($\lambda \approx 1000$), then the shape of the Poisson distribution can be approximated very well using a normal distribution. For values of λ considerably smaller ($\lambda \approx 10$), a Poisson can still be approximated with a normal distribution if a certain adjustment is made when calculating the mean and variance. This adjustment is called a "Yates correction" and it is really nothing more than cheating, for example, to the right because you know the small λ results in a skew to the left. We will not need this correction here.

A note on convergence: throughout the paper, I will make statements like "if λ is sufficiently large" or "this holds if $\lambda > 10$", *et cetera*. These statements are based on accepted results concerning convergence of distributions. For example, the Law of Large Numbers states that the sample mean will converge to the true mean if enough samples are taken. There is no clear cutoff for "enough samples". Convergence only describes the direction of movement–we never actually arrive at the destination. More samples or bigger λ may improve the results, but it is entirely up to personal preference and the context to determine when the point of "close enough" is reached. This is difficult in this setting, because making a decision about things like sample size and number of tracer particles is part of the production process. I cannot answer these questions or say what is right,–I can only comment on what is reasonable. Essentially we know only what's horrible. Everything else is simply "not horrible".

1.2 Testing

Ultimately, we will be using the theoretical model above for comparison to determine if our observed samples in fact came out of a Poisson distribution. Visually, we can picture plotting point by point our measurements on top of the theoretical drawing of the Poisson distribution and seeing how things line up. (In reality, statistics tables tells us how things "line up", as David Eisenberg's original paper–and the next section–show.). If they do line up, then the particles are where they should be–they are Poisson. If they don't line up, then more mixing is required. The key point is to trust that the particles inevitably will be Poisson in the mixture if mixing is thorough enough. This is a property of nature.

Think of it this way: In the feed mixture example, there are two distributions I must consider. The first is the true Poisson distribution that the tracer particles will eventually become after mixing. The shape of this true distribution is determined by the true λ, which is known as a function of the number of tracer particles added and the amount of feed in the mixer (equation (4)). The second distribution is the actual distribution of the particles in the mixer. At certain times during the mixing process (particularly early on), I may have no idea what this actual distribution looks like. To estimate this, I take samples. The tests that follow will use the samples to determine when the actual distribution is sufficiently close to the true distribution. In other words, these methods test if the particles are finally where they should be. A diagram in the following section illustrates this.

The feed mix context here is very unique for statistical modeling because we know what the actual value of λ is. Typically, this is unknown. (How many galaxies are there in the universe?) This knowledge is what allows me to use the word "true" to describe the theoretical distribution of particles (usually this is called the "population" distribution). Calculation of the true λ is simple:

$$\lambda = \frac{\text{\# of particles added}}{\text{total volume of feed}}. \tag{4}$$

Of course there is some uncertainty in both of these values. For the number of particles, we know the volume added (or the mass added) and we can use this to calculate the number of particles, but there is obviously some measurement error. Likewise, calculation of the total volume (or mass) of feed involves some error. The precision and accuracy of our λ calculation will be revisited later in this paper.

I want to reiterate a point made in a footnote earlier in the paper. A "homogenous" mix means the particles have finally reached their Poisson distribution–further mixing will not increase randomness. A "complete" or "good" mix means that the particles aren't quite yet Poisson, but they are random enough for the purposes faced by the person mixing. "Homogenous" is as random as possible, "complete" is as random as needed.

1.2.1 Poisson test to determine the shape of our observed distribution

Recall that the shape of the Poisson distribution is completely described by the parameter λ. The Poisson test presented here is designed to estimate the value of this parameter for the observed distribution, and in doing so it will determine the "shape" of the observed distribution. Since λ is also the mean, this test also estimates the position of the observed distribution, but I want to focus on shape to exaggerate the difference between this test and the chi-squared test below which emphasizes position. So, assume the Poisson test is for shape–it answers this question: "Is the plot of data points shaped like a Poisson?"

Each sample drawn from the mixer should be consistent with a Poisson distribution with known parameter λ if the mix is homogeneous. To test this for a single sample, calculate the following using the tracer particle count X and the known value for λ :

$$\frac{\lambda^X e^{-\lambda}}{X!}. \tag{5}$$

This will give me the probability that I see X given the fact that λ is true. For example, let $\lambda = 100$ and suppose I count 90 particles from a sample. Equation (5) gives a probability of 0.025, meaning that the odds of me seeing 90 particles given a Poisson with $\lambda = 100$ is only 2.5%. This may seem small, but keep in mind, there are millions of possible outcomes for each sample–I could get a particle anywhere between zero and the total number of particles added to the mixer. In fact, the largest probability given only one sample and $\lambda = 100$ is about 4% (occurs when $X = 100$).

To make our test more robust, we construct a range of values and determine the probability that X falls within this range. For example, given $\lambda = 100$, what are the odds that $90 \leq X \leq 110$? It turns out that the probability of this is about 70.6% (found by taking the sum of the individual probabilities found by equation (5) for $X = 90$, $X = 91$... and so on). Widening this range will increase the probability. For example, if I want 90% odds that X will be in a certain range (given $\lambda = 100$), the range around X should be $84 \leq X \leq 116$. In this fashion, the probability tables, like those used in the original Eisenberg paper, are constructed. A further improvement would be to use multiple samples, take their average, and use this average value in place of the single observation. This is very similar to (although not technically identical to) the procedure outlined in the Eisenberg paper. There he calculates the average of several samples, and uses this average to define a range for λ. (The former determines a range around X, the latter a range around λ.) Either method is valid, but there is a difference in flavor between the two–do you want to test the odds of a certain λ given sample data or do you want to test the odds of a certain sample given knowledge of λ?

In the context of feed mixtures, equation (4) states I should have a good idea what λ is if I know the number of particles added and I know the amount of feed being mixed. Obviously, uncertainty in either of these will increase uncertainty in my knowledge of λ. Strong knowledge of λ may lead me to a test that constructs a confidence interval around the sample value X or $\overline{X}$. On the other hand, if I have great confidence in my counting accuracy for samples, but I may be uncertain about how many particles I added (or what happened to them after I added them), then I may instead prefer to test a confidence interval around λ using my known value of X or $\overline{X}$.

Here is a danger: if I use a sample value X (or sample average $\overline{X}$) to construct a confidence interval, I can say something like this: "given this observed $\overline{X}$, the true mean λ will fall between A and B 90% of the time". I cannot say this: "given $\overline{X}$, 90% of samples will fall between A and B". The former involves making an inference about the population based on a sample–a standard statistical practice. The latter involves making an inference about future samples, based on the current sample. This is invalid. I can either predict the truth (λ) based on a sample ($\overline{X}$), or I can predict a sample knowing the truth. I cannot predict a sample based directly on another sample. In statistics, we usually are forced to predict the truth based on a sample because the truth is usually unknown. In the feed mixture problem, we are in the unique position of knowing what λ actually is since we add a known amount of tracer particles, so we can make either prediction. This is a luxury of the context.

The probability that an observed X would fall within an interval from $\underline{X}$ to $\overline{X}$ given known λ is this:

$$\text{Probability that } X \in [\underline{X}, \overline{X}] = \sum_{X=\underline{X}}^{\overline{X}} \frac{\lambda^X e^{-\lambda}}{X!}. \tag{6}$$

The general formula for the confidence interval around λ is this:

$$\frac{\chi^2(\alpha/2; 2k)}{2} \leq \lambda \leq \frac{\chi^2(1-\alpha/2; 2k+2)}{2} \tag{7}$$

where $\chi^2(p; n)$ is the χ^2 value for significance level p (for example 5% or 1%) and degrees of freedom n.

1.2.2 Pearson's chi-squared (χ^2) test to determine position of our observed distribution

Since our observations are likely to be independent and random if our test procedure is reasonable, it is appropriate to use the Pearson's chi-squared test to determine whether the observed distribution of tracer particles is consistent with the theoretical Poisson distribution. This is called a "goodness of fit" test–how well do the observations "fit" our theory? Although the Poisson is not a continuous distribution, it is accepted that cases where λ is sufficiently large do not require adjustment from the standard chi-square test. Again, a Yates correction is an option (that we will not use). I will comment on what "sufficiently large" means in a following section. (Equation (8) is valid for $\lambda > 5$.)[2]

The Pearson's chi-squared test is straightforward. If the independence assumption holds, then

$$\sum_{i=1}^{N} \frac{(X_i - \lambda)^2}{\lambda} \tag{8}$$

follows a χ^2 distribution. The procedure is simply to count the particles in each sample, use the known value of λ to compute the sum above, and look at a chi-squared table to determine the likelihood of seeing the outcome we saw if the underlying Poisson characteristics are visible in our observed data. Low probabilities (low p-values) imply that the data points don't "line up" with the Poisson as we thought they would. This means that the mix isn't yet homogeneous.[3] Low p-values do not imply the theory about particles distributing via Poisson processes in the mixture is wrong–they eventually will, and we never doubt this. We are simply testing to see if enough time has passed.

An alternate form of the chi-squared test, one presented in the Eisenberg paper, uses the following statistic:

$$\sum_{i=1}^{N} \frac{\left(X_i - \overline{X}\right)^2}{\overline{X}} \tag{9}$$

which also follows a χ^2 distribution. The test procedure is the same, but there are a few details that deserve mentioning. First is the interpretation of the results. The test performed by equation (8) will test whether our observations are consistent with the true mean λ, while the test in equation (9) tests if our observations are consistent with each other. A second difference involves the degrees of freedom, which is addressed below (equation (9) will have one less degree of freedom since we use one to estimate $\overline{X}$). Both tests have merit, but care must be taken in determining which to use. It is true that both will test the odds that the set of observations were drawn from the same population, which may imply homogeneity of the mix (existence of different populations in the mixer clearly means it's not homogeneous). However, it is possible all of my samples appear to come from the same population, even though the mix is not homogeneous. Maybe I sample repeatedly from a quiet area in the mixer, or maybe there are places in the mixer that are identical to each other, but

[2] See Box, Hunter and Hunter. Statistics for experimenters. Wiley. p. 57

[3] The importance of differentiating between "homogeneous" and "complete" is clear at this point.

very different from every other location. Equation (8) will do a better job of catching this than equation (9) because (8) requires not only that the samples be taken from the identical populations, but that each of these population has the *correct* mean. Equation (9) only requires that the samples be taken from populations with the *same* mean. If we are confident that we know the true value of λ we should use equation (8).

1.2.3 Degrees of freedom

The rule for degrees of freedom in the chi-squared test of goodness of fit test is this: $N - r - 1$, where N is the number of observations and r is the number of parameters we need to estimate in order to know the shape of the theoretical distribution. For the Poisson, the only parameter needed to completely describe the shape is λ, so the degrees of freedom should be either $N - 1$ (if we use the known λ) or $N - 2$ (if we use $\overline{X}$ instead of λ) Note that decreasing the number of samples N will reduce my ability to correctly identify when the mix is still heterogeneous since N affects degrees of freedom, which in turn affects chi-squared critical values.

1.3 Type I and Type II errors, and the definition of a "good mix"

The previous paragraph made an assumption that should be stated explicitly: the null hypothesis we are testing is that the mix is either homogeneous or complete. This means that a Type I error (rejecting a true null) results in further mixing when we could have stopped. A Type II error (accepting a false null) means we conclude the mix is complete when it is not. The likelihood of a Type II error is inversely related to the "power" of the statistical test we are using. If we want a powerful test (so we can reject more nulls), then we will make Type II errors more often. The power of the test is also related to the optimal sample size needed to perform the statistical analysis. As mentioned above, smaller samples (smaller N) mean smaller power.

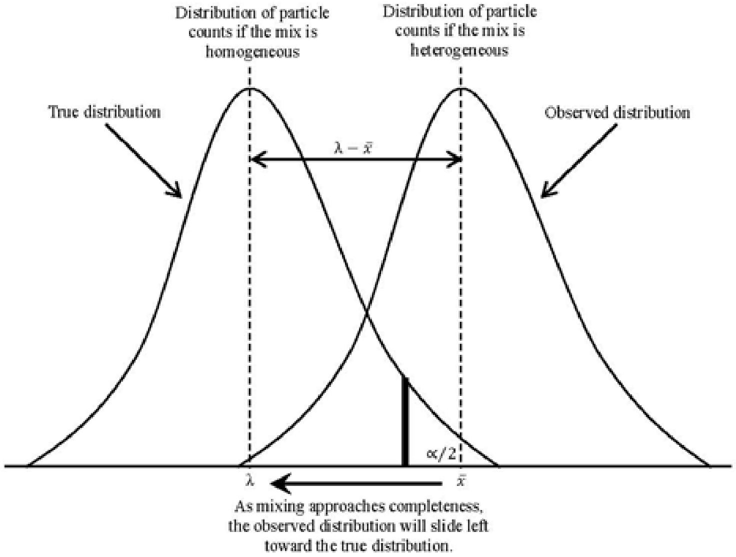
This highlights a fundamental question in determining the condition of a mix–how do we describe different degrees of heterogeneity? What does it mean to say something like "the batch is only 80% mixed"? Is this the same as saying "the odds of a Type II error is 20%"? No it is not. The former deals with the level of homogeneity within the mix, the latter deals with a critical value only. Essentially "mixed well" and "mixed very, very well" will be evaluated differently in the former but identically in the latter. If you want to say "I will correctly identify a good mix 80% of the time", this can be accomplished by adjusting the power of the test, but this says nothing about what those 80% look like (they may all have been mixed twice as long as necessary).

A similar point can be made about selection of a critical value for rejecting the null (we call this the "size" of the test). Both 5% and 1% test sizes have been proposed for this process, where the smaller size means I will reject the null less often. This is not the same as saying that I will catch only really bad mixtures. The size of the test determines how carefully I separate the mixtures into two groups: "good mix" and "bad mix". Within each group, there will be a variety of levels of mixing–bad mixes and absolutely horrible mixes will be in the same group–and the size of the test can do nothing to distinguish between these two. Both will be rejected in the same fashion.

So what is the thrust here? Determining a comfort level with Type I and Type II errors is key. We have to keep in mind that the statistics used here only give us "yes" or "no" answers–they do not tell us "how well is it mixed?". As a result, careful consideration about the costs of each type of error is absolutely paramount. It is possible to refine a statistical test until it nails your chosen critical value with immeasurable precision, but this test will be useless if you choose an inappropriate critical value.

1.4 Probability revisited

As I described earlier, in sampling from the mixer, we are essentially plotting data points and seeing if this plot lines up with the true distribution. The probability theory and testing described above can be clarified using the simple diagram below. As mixing progresses, the average number of observed particles per sample should approach the expected (true) mean. Essentially, the observed distribution moves toward the true distribution, as illustrated in the figure below where the true mean is λ and the observed mean is $\overline{x}$.


Distribution of particle counts if the mix is homogeneous
Distribution of particle counts if the mix is heterogeneous
True distribution
Observed distribution
$\lambda - \bar{x}$
$\propto/2$
λ
$\bar{x}$
As mixing approaches completeness, the observed distribution will slide left toward the true distribution.

Although the figure above implies normally distributed variables, please keep in mind that the true distribution is Poisson (although for large λ the two look very similar). Also, although I have drawn the observed distribution to be smooth, it should really be a collection of points. Finally, the variance of each distribution above appears to be identical–a detail that may not hold in reality (the observed distribution may be wider than the true distribution, for example). However, these assumptions are only made in an effort to reduce clutter in the figure. Relaxing these assumptions does not undermine the results or intuition of what follows.

Based on the situation pictured above, we would reject the null hypothesis that the mix is complete (or homogeneous) because the observed mean $\overline{x}$ lies to the right of the critical value (the dark line separating λ from the region labeled $\alpha/2$). If $\overline{x}$ is further from λ than the critical value, then we have confidence that $\overline{x}$ would be so rare under a true null that we reject the null.

How do Type I and Type II errors factor into this intuition? Consider the following four possible outcomes of a statistical test based on the figure above:

1. The distributions are different and I get a value for $\overline{x}$ that is to the right of the critical value, in which case I reject the null (correctly).

2. The distributions are different and I get a value for $\overline{x}$ that is to the left of the critical value, in which case I accept the null (incorrectly). Type II error.

3. The distributions are the same and I get a value for $\overline{x}$ that is to the right of the critical value, in which case I reject the null (incorrectly). Type I error.

4. The distributions are the same and I get a value for $\overline{x}$ that is to the left of the critical value, in which case I accept the null (correctly).

Clearly, adjusting the critical value will affect the odds of the errors in #2 and #3, as will the actual position of each distribution. If I chose a very small α such that I must be very certain before I reject the null, then the critical value (dark line) moves to the right–which means the group that I consider to be good mixes will start to include more and more bad mixes. You can picture nudging the dark line delineating the critical value to the right in this diagram until it is just right of $\overline{x}$, in which case I would accept the above observed distribution as being statistically the same as the true distribution and would thus conclude that the mix is good, even though it is not.

This figure, and the understanding of the relationship between the shape and position of the distributions relative to each other, will help us answer questions concerning things like optimal sample size and particle count. The moral of everything that follows is how important it is to have an intimate understanding of what a Type I and Type II really mean in terms of the feed mix process.

2 How many samples should be taken from the mix?

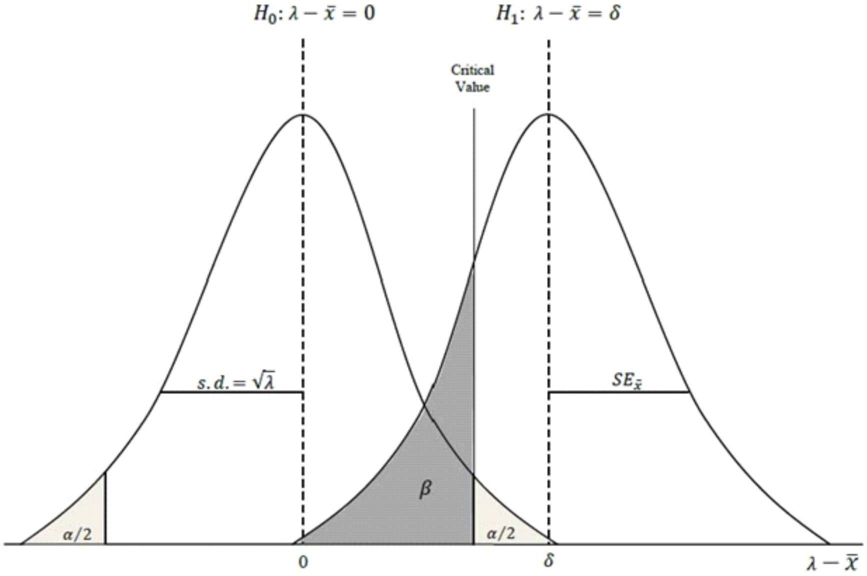
2.1 The importance of sample size N

This question addresses the issue of "sample size" in the statistical sense–this is about the number of observations, which I call N. Sample size N is not the same as the "size of each sample", which should refer to the amount of feed (in grams) taken out of the mixer for analysis. As a general rule, thanks to things like the Law of Large Numbers (the sample mean approaches the true population mean as the number of observations grows) and the Central Limit Theorem (the sample means will be normally distributed for big sample sizes), large sample sizes are always desirable. Unfortunately, there is not one single rule or accepted definition for "sufficiently large". This matter is complicated further because the Poisson distribution measures successes and failures–so the sample size and the number of observations may be difficult to keep straight. For example, if I toss a coin 10 times and see 3 tails, what is the sample size N? It is 10, not 3. In this sense, "sample size" should be interpreted as "number of trials", not "number of successes". In the context of testing feed, sample size is determined by the number of samples taken out of a batch for testing. The number of particles counted is not sample size. Consider another example, where I am measuring the height of people in a population, and I want to know if heights are normally distributed. I grab a person, get my ruler and record their height. I now have one observation, so $N = 1$. Using this height measurement, I put a dot on my normal distribution graph and I should realize that I won't be able to tell anything about the shape of observed distribution–I only have one dot and I need to know if it looks like a bell shaped curve. Grabbing a second person will help–grabbing a more accurate ruler will not. Increasing the number of samples is the former, increasing the number of tracers in a batch is the latter.

The statistical tests we use are analogous to comparing a plot of the data points to a plot of the correct Poisson distribution and comparing their shapes. As a result, if I have only a few data points, I won't be able to infer much about the shape of the observed distribution, and this will reduce my confidence in the test. However, if I have 50 observations, then I should be able to tell very easily what the graph of the distribution looks like. How many is enough? Again, there is not one correct way to answer this question, but I would be quite uncomfortable having fewer than 10 observations. 20 would be good, 30 would be luxurious–more than this is unnecessary.

3 Lehr's equation and sample size N

A formalized method for determining an appropriate sample size for statistical testing was offered by Lehr (1992). Some terminology and notation is required, which I will develop here. The following figure is designed to help motivate the intuition for this section. Please note: although the figure below looks very similar to the figure in section 1.4, they are VERY different. The previous figure plots the mean on the x-axis. The figure below plots the *difference* in means on the x-axis.



As with the first figure in this paper, these appear to be normal distributions, but they would be Poisson. The distribution on the right is the observed data, and the distribution on the left is the true distribution. The notation used here is standard. $SE_{\overline{x}}$ is the standard error for the observed distribution and, since it is Poisson, $\sqrt{\lambda}$ is the standard deviation for the true distribution (these appear to be equal in the figure, but they need not be), $\overline{x}$ indicates the sample mean, λ is the true mean. Subscript 0 indicates the null, 1 indicates the alternative, and H_0 is the null hypothesis (homogeneous mix), which requires that the sample mean be the same as the true mean, $\lambda - \overline{x} = 0$. (I have drawn this such that the observed distribution is to the right of the null ($\overline{x} > \lambda$), but the converse could be true instead.) The probability of a Type I error is α (size of the test) and the probability of a Type II error is β (dark shading). Power of the test is thus $1 - \beta$. Notice that here I assume a 2-sided test, so each tail is labeled $\alpha/2$.

Here we would be testing the null hypothesis that the difference in means is zero. This is consistent with the test we use when mixing feed–we want to know if the actual average number of particles is the same as the predicted theoretical value. Notice that the null distribution identifies the size of the test α, and the alternative distribution identifies the power of the test $1 - \beta$. If I chose α such that I must be very certain before I reject the null (α would be small–1% or

0.1% for example), then the critical value (labeled in the figure) moves to the right and β grows–which means my ability to correctly identify an incomplete mix falls. I will be accepting more incomplete mixes as α approaches zero. Thus, lower Type I error (small α) implies higher Type II error (big β), *ceterus peribus.*

Formally, the distance between the critical value and the null H_0 is $0 - z_{1-\alpha/2}\sqrt{\lambda}$ and the distance between the alternative H_1 and the critical value is $\delta - z_{1-\beta}SE_{\overline{x}}$, where z comes from a standard normal table. The letter z is used because we are dealing with normal distributions. I will proceed under the assumption that $\sqrt{\lambda} = SE_{\overline{x}}$ only for simplicity. This assumption can be relaxed without damage to the results.

Using the information above, Lehr derived the following formula for sample size:

$$N = \frac{2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\left(\frac{\mu_0 - \mu_1}{\sigma}\right)^2} \tag{10}$$

The numerator for $\alpha = 5\%$ and $\beta = 20\%$ is equal to 15.68, which I round up to 16 in what follows.[4] This number is calculated using the normal table attached for the chosen values of α and β.

Now we can transition to Poisson. It is known that if a random variable X is distributed Poisson with mean λ, then $\sqrt{X}$ is distributed Normal with mean $\sqrt{\lambda}$ and variance $\sigma^2 = \frac{1}{4}$. This is the result of something called the "variance stabilizing transformation". This fact, and equation (10) imply the sample size rule for a Poisson distribution using $\alpha = 5\%$ and $\beta = 20\%$ is this:

$$N = \frac{4}{(\sqrt{\lambda_0} - \sqrt{\lambda_1})^2}. \tag{11}$$

All the math aside, how do we use this formula? Initially we me have the urge to claim that λ_0 and λ_1 are equal, so the denominator of this fraction is zero. However, λ_0 and λ_1 are equal only when the mix is perfectly homogeneous. If I added the tracer particles only moments ago, I have no reason to expect the sample mean to be equal to the true mean. Thus, this formula, in conjunction with my understanding of how a mix proceeds, could be used to determine how many samples should be taken as a function of time. As λ_0 and λ_1 get close (i.e. as the mixing nears completion), the denominator gets smaller, and so my optimal N must get larger. I thus need to take more samples as mixing progresses. How many depends on how I expect the observed mean to evolve over time.

For example, consider a case where $\lambda_0 = 50$, and I have been mixing for a certain period such that I believe that the mix is 80% complete. If this were the case, $\lambda_1 = 40$ may be the hypothesis I would test to confirm my beliefs about the status of the mix.[5] Given the formula above, I would test at least $N = 7$.

[4] Note that these are the same values chosen in the TNO report that will be discussed in a later section of this paper.

[5] Note that this isn't necessarily the only interpretation of the statement "the mix is 80% complete".

If instead, I believed that the mix was 90% compete, I would test $\lambda_1 = 45$ which would require $N = 30$.

4 How many particles should be added to the mix?

Due to things like measurement error, the exact count of Microtracer particles is not known with perfect precision. If we say 25,000 particles per gram, we actually mean something close to 25,000–but like any other estimate, the best we can do is a confidence interval. This uncertainty means that we don't really know precisely what the true value of λ is, but, as we have seen, we have methods to construct confidence intervals around λ based on sample data. The uncertainty about the particle count also complicates the discussion that follows, particularly in the context of the TNO report. For example, consider a situation where determine that increasing to $\lambda = 400$ will improve our results by 10% (a thought that we will address momentarily). In order to make this determination (or in order to execute this policy) we would require considerable precision concerning our initial tracer particle count to be added to the mixer. Knowing that $\lambda = 400$ is 10% better than $\lambda = 300$ is only valuable to me if I can actually distinguish $\lambda = 400$ from $\lambda = 300$. If measurement error prevents this level of accuracy in my measurement of initial particle count, then the statistical superiority of $\lambda = 400$ is irrelevant. The point is this: not knowing perfectly the number of particles per gram of mix means that marginal adjustments to λ are going to be problematic.

4.1 The (un)importance of expected particle count

Suppose we are mixing 3000kg of feed to which we add microtracer particles (25,000/gram) at a rate of 50g per 1000kg of feed. The number of particles we expect to find in a sample of 80g of mix is known at the moment the microtracers are added. At the ratio used here, we would expect to see 1.25 particles per gram of feed, or 100 particles in our 80g sample if the mix was perfectly homogeneous (i.e. perfectly Poisson)[6]. This is λ, the true population mean–which is where the actual Poisson distribution for this batch is centered. If I want a larger λ, there are two ways I can proceed: I can either take bigger samples from the mixer (160g instead of 80g), or I can add more particles to the mix. The latter option (adding more particles) will affect how I interpret my statistical testing results. The reason has to do with the fact that every aspect of the shape of the Poisson distribution depends on the value of λ. If I add more particles, I increase λ, which in turn increases the variance, which makes the distribution fatter. A larger variance will increase the size of the tails of my distribution, which will increase the odds of Type I errors unless I adjust my critical value (typically labeled α) to compensate.

[6] I also need to assume counting error is zero and no particles are lost–assumptions that may be brave in some situations.

4.1.1 Being misled by the CV

As formula (2) shows, the coefficient of variation is a function of λ. Specifically, since the CV is inversely related to λ, adding more particles to the mix will reduce the value of the CV. As with our previous discussion, interpreting what this means is key. The CV tells me how my measurements vary with respect to the mean (think of it as spread as a percent of the mean). If I can reuse the height example from above, the CV would tell me how my height measurements vary for short people when compared to measurement variations for tall people. I would hope that if I'm off by 2 inches on average for a 6 foot tall person, that I would be off by no more than 2 inches when I'm measuring a 5 foot tall person. If the opposite were true, I would be worse at measuring short people than tall people in percent terms, which would require some theoretical justification as to why. We can interpret a smaller CV in this example essentially as the result of using a better ruler. When it comes to microtracers in feed, the intuition is similar. If I expect only 4 particles in a sample, and I miss one due to measurement error, then I have missed 25% of the particles. Increasing the sample size will allow me to miss the same number of particles while decreasing the percent variation.

However, in the context of testing a mixture of feed, we are not entirely concerned with improving the CV. True, a lower CV implies less variation percentage in my tracer particle counts for each sample, consistent with the "better ruler" analogy. Further, using a better ruler will allow me to more precisely place the measurement I have just made as a data point on my Poisson graph. However, as we have seen before, precise placement of each point is not as important as how several points line up with the theoretical Poisson distribution. Improving the CV will not help us with this. Even worse, falling in love with a low CV may give us a false sense of security in our results.

4.2 Sample size and the Poisson test

As we know, I can use the sample observations to estimate if the shape of the distribution is consistent with my true Poisson as discussed above. To do this, I would use formula (1) and statistical tables that give critical values of Poisson random variables. Keeping with the idea that increasing the number of tracer particles in a batch helps me accurately position data points on my graph, it follows that more particles means better Poisson test results. This is technically true, but as before, we need to be careful to interpret the improvement carefully. From what we know about the Poisson distribution, values of λ that are larger than 10 are sufficient for statistical testing.[7] Intuitively, the concern about small λ with respect to the Poisson test for shape is this: if λ is very close to zero, then I may find that the left tail of my Poisson distribution is chopped off, which may make it difficult for me to correctly plot the shape of the distribution. This intuition motivates our selection of λ. Larger λ values will not hurt, but

[7] Existing literature would suggest using the Yates correction for $\lambda = 10$. This should be thought of as a lower bound.

the benefit is asymptotically approaching zero–which means the cost of more particles must be zero if the benefit is to exceed the cost. Since there is again no single rule to follow, I suggest $20 \leq \lambda \leq 100$ based on what we know about asymptotics and the Poisson distribution.

4.3 Sample size and the χ^2 test

The chi-squared test outlined above essentially tests the location of the mean of my observed distribution and compares it to the location of the mean for the true distribution. If I add to the feed a number of microtracers such that I expect 100 of them in each 80g sample, then the location of the mean of the true distribution equal to 100. If I count a sample and find only 62 particles, then the chi-squared test will tell me the odds of seeing only 62 tracers given that I expected 100. If I add additional particles in order to increase λ then I am shifting both distributions–the true and the observed–to the right (see figure above). The distance between the two means will not change. However, since the variance of the Poisson distribution increases when λ grows, I would expect to see more observations far away from the mean–which may make me conclude that my observed distribution is not in the right place–so I will continue to mix unnecessarily. Thus if I do increase the number of tracers used, I must also increase the standards by which I evaluate my results–more particles must be accompanied by stricter critical values. This may sound like an improvement in our testing–and it technically is since we will be more precise about our rejections, but it is an improvement in statistical testing that likely results in zero benefit in the quality of the feed. Increasing the number of particles in the feed will help us determine how precisely the mean of the true distribution differs from the mean of the observed distribution. However, this is a benefit that applies only to the cases where we are very close to the critical value. Increasing λ does nothing in the context of helping determining an appropriate critical value, which as we have seen is a fundamental issue. Knowing that I am within an Angstrom of a critical value is only important if the critical value has important meaning. We will revisit the relationship between sample size and critical values in a later section.

5 How big should a sample be?

To avoid confusion between "the size of the sample" and sample size N, I will say "the mass of the sample". I will assume that samples are measured in grams. This assumption, although seemingly benign, actually could be problematic if the mass of some ingredients in the feed are very different per unit volume than the mass of others. A alternate method would be to define the size of a sample as a volume, but this could be equally problematic for different reasons (this would probably work best for liquids). Note that the mass of the sample taken is intimately related to the number of tracers we expect to find–a bigger scoop should contain more tracers. Knowing this, we should set the mass of a sample such that we do not risk getting into trouble by flirting with small values for λ. The guidelines in the previous section will help us with this.

Something else to consider is that measuring the mass of a sample itself is not an exact science–we will have a mean and variance for each mass, although the distribution will be approximately normal (not Poisson) as long as we take a reasonable number of samples. What we know about things like the CV, more massive samples will be less variable in percent terms. Think of using a scale that is accurate to 0.1 grams. We learn in high school chemistry that, if this is the case, 10.1g and 10.2g must be considered to be the same mass because of significant figures. It follows that keeping the mass sufficiently large will result in more precise testing. However, excessively massive samples may undermine the accuracy of the results. Intuitively, we can think of the limiting case where the sample size is so large it is equal to the entire volume of the mixer. In this case, we may accurately count the particles present, but we would have no information about how those microtracers were distributed throughout the batch. Of course the limiting case in the opposite direction is no better–taking such tiny samples that no particles are present would clearly not produce reliable results. These things said, when it comes to determining the mass of the sample to be take, our only concern should be that λ is acceptably large. Beyond this, as long as the mass of the sample is much smaller than the mass of the entire batch being mixed, any convenient size should suffice.

6 From where should the samples be taken?

This question was touched on in the original Eisenberg paper, although an answer was not formalized. The issue is that the microtracers are initially added to a single location (or possible several locations) in the mixer. As a result, samples taken from regions near this point will have a different distribution of tracer particles than samples taken from further away in the mixer, at least if the mixing time has been short and a complete mix has not been achieved. This could lead me to conclude that the mix is better than it actually is if I happen to pick from a spot in the mixer where the particles are coincidentally similar to my expected values. There are several solutions to this problem. The first is to take samples simultaneously from several locations within the mix. Alternatively, I could take several samples from the same location but at different times during the mixing process. These two can actually produce different results. For example, if heavy particles sink and lighter particles rise within the mix, then sampling from the exact center of the mixer repeatedly over time may imply a homogeneous mix when I actually have considerable heterogeneity. It might be possible that a certain region in the mixer doesn't agitate as well as others, and if I am unfortunate enough to sample from that location, my results will not accurately reflect the degree of homogeneity. It follows that a combination of both–samples from several locations taken at several different times–would give the best results.

Consider an example, where tracer particles are added at point A which is far away from some other point B. I always take a 100g sample from the exact midpoint of point A and point B at different times. During the time between my samples, the tracer particles migrate slightly within the feed mix. As a result, the number of particles observed at my chosen location should evolve over time. Initially I should see fewer than expected (since I am far from point A where the particles were added), and this number should approach my expectation as mixing continues. If my observations finally coincide with my expectations, is the mix perfect? Since I never sample from point B, I technically don't know how the mix looks at that location. Increasing the number of sample locations addresses this issue. It should also be noted that the physical world occasionally doesn't agree with the world we describe on paper. For example, tracer particles may get coated in goo as mixing time increases making them hard to find. This means that convergence to expectations is not perfect (it may not even be monotonic), and sampling from a single location over time may not be ideal.

So how should sampling location be determined? It would be difficult to justify not having more than one location and more than one sampling time. Keep in mind, the real concern here is sample size N. If we want $N = 10$, then perhaps 2 locations at 5 different times may be appropriate.[8] How would these results differ from taking from 5 locations at 2 different times? As I said above, one may be more vulnerable to things like quiet spots in the mixer while the other may be more vulnerable to things like floating and sinking particles. Consistent with the general theme, consideration of the costs and benefits must guide this decision.

I cannot overemphasize that sampling times must be selected carefully to avoid convoluting the statistical tests. If I intend to use samples from different sampling times, then I must be certain that all samples are taken to test a consistent alternative hypothesis. For example, I cannot sample 10 seconds after the mixing begins and then again after 20 minutes of mixing if I intend to use these samples in the same hypothesis test. The null at both periods is the same (compete mix), but the alternative hypothesis (λ_1) will be very different. (See the discussion about equation (11) above.) I can use these observations to estimate convergence of the mean, but I cannot use them to test position of the mean in the same test. This is consistent with the idea that a decision about whether testing position or convergence is desired. What would be sound is if I sampled at two different times–if at both times I had the same expectation about the value for λ_1. This is the basis for my recommendation about multiple sampling times.

[8] $N = 10$ is pushing the lower envelope in terms of convergence.

7 How many sub-samples should be taken from each sample?

Using multiple sub-samples from a single sample can be problematic in the sense that the observed particle count in sub-sample A will not be independent of the particle count in sub-sample B. For example, if the original sample was taken from a portion of the mix that was densely populated with microtracer particles, both sub-samples would be expected to have high particle count. If we violate the independence assumption, we undermine the validity of our statistical inference–for example, the chi-squared test may no longer be viable. That said, dividing a sample into sub-samples may still have value, but the sub-sample values should be combined somehow (e.g. averaged) instead of being used separately as two different observations. The intuition behind this risk is related to the discussion in the previous section about sampling time and location. Dividing a sample into two sub-samples is equivalent to taking two samples from the same location at the same time. Since tracer particles had no time to move between the sampling times, and the distance between the locations is zero, two sub-samples provides no more information about the homogeneity of the mixture than does a single sample taken from the same time and location.

8 TNO report

The TNO report begins with a couple of examples that motivate a complaint concerning the trade-off between Type I and Type II errors. Specifically, a 1% confidence interval was used in their simulations testing the null that the mix is good. The results in these models were acceptable for Type I errors, but unacceptable for Type II errors–they mistook too many bad mixes as being good. The standard solution is to move the critical values closer to the mean (larger α) which will result in more observations being captured in the tails. Since more observations are now further from the mean than the critical value, there will be more rejections of the null hypothesis (that the mix is good) Declaring a mix is good less often means fewer false "goods"–but it can also mean fewer true "goods". This is exactly what they found–increasing α to 5% reduced the Type II error to acceptable levels, but sacrificed too much in terms of Type I errors–they mistook too many good batches as being bad.

The problem the authors face is this: since this trade-off between Type I and Type II errors affects each of their simulation models differently based on the parameters of each (particle count, etc.), it becomes difficult to compare two different methods of evaluating the quality of a mix. One particle count or sample size might perform better for Type I errors, but the other is superior for Type II errors. How do we decide which dominates?[9] To answer this question, the TNO report uses computer simulations to compare hundreds of models

[9] At this point, I would ponder deeply the meaning and economic impact of what each type of error. Only an intimate understanding of the product being produced, the costs and benefits of each error, and the overall context of the problem can answer this question.

with different parameters.[10] They also varied things like measurement error and the level of mix heterogeneity. Based on the results of these simulations, a determination was made about the best number of particles, sample size and use of sub-samples. Before I proceed, I want to extend a warning about numerical methods: simulations can be very useful, but I think they have the tendency to hide the meaning of the results beneath masses of very attractive computer output. Essentially they shift the focus from intuition to interpretation of data, which can be risky. That said, the results are still very helpful in understanding the problem at hand, as we will see.

The comparison of different models in the TNO report is accomplished as follows. "Good" is defined as a distribution of particles that is sufficiently skinny (low variance) so that 68.2% of observations fall within ±4% of the mean.[11] The authors set an 80% acceptance rate standard for good mixes in every model tested by adjusting the size of the α until 80% of mixes that are actually good are labeled good. If a model with particular parameter values correctly accepts only 70% of good mixes, then the authors widen the confidence interval slightly (smaller α) to reduce the number of rejections and increase the acceptance rate to 80%. Once normalized this way every model will miss the same number of good mixes. Since the normalized models perform identically for good mixes, each model is evaluated based on how many bad mixes get through–obviously fewer is better.[12] A "bad" mix is defined as a distribution of particles that is sufficiently fat (high variance) such that 68.2% of observations fall within ±10% of the mean.

Since we have defined "good" as $SE \leq 4\%$ of the mean, we are essentially testing whether the standard deviation of the sample from which our data was drawn (with known λ) is less than or equal to 0.04λ. If it is, we need to identify this 80% of the time according to our normalization rule. This test thus concerns the confidence interval around the standard error for the distribution from which we drew our data. The formula for a confidence interval around the estimated standard error in this situation is straight forward:[13]

$$SE \cdot \sqrt{\frac{N-1}{\chi^2(\alpha/2; N-1)}} \leq \sigma \leq SE \cdot \sqrt{\frac{N-1}{\chi^2(1-\alpha/2; N-1)}} \tag{12}$$

where $\chi^2(p;n)$ is the chi-squared value for p level of significance with degrees of

[10] The parameters were λ (75 to 750), N (10 or 20) and the number of sub-samples (1 or 2).

[11] Here's where the 68.2% comes from: they define the variation in non-uniformity as $f_{nu} = \frac{SE_{nu}}{\lambda}$ and arbitrarily state that $f_{nu} = 4\%$ is the definition of "good". It follows that, for a good mix, $SE_{nu} = f_{nu} \cdot \lambda = 0.04 \cdot \lambda$. For a normal distribution (which they assume), 68.2% of observations fall within the interval $[\mu - SE, \mu + SE]$. This implies that 68.2% of observations here fall within $[\lambda - .04\lambda, \lambda + .04\lambda] = [.96\lambda, 1.04\lambda]$. I.e. 68.2 % fall within 4% of the mean.

[12] A more technical way of saying all of this would be as follows: A "good" mix is defined as a mix that is normally distributed with standard error equal to $0.04 \cdot \mu$. Fix the probability of Type I error at 20%, then minimize the odds of Type II error.

[13] See, for example, "Handbook of Parametric and Nonparametric Statistical Procedures: Fourth Edition" by David J. Sheskin, p.197.

freedom n. Recall that for the Poisson distribution, so $\sigma = \sqrt{\lambda}$. Also recall that our standard for a good mix is that $SE = 0.04 \cdot \lambda$. Making these substitutions into the equation above, we get the following:

$$0.04\lambda \cdot \sqrt{\frac{N-1}{\chi^2(\alpha/2; N-1)}} \leq \sqrt{\lambda} \leq 0.04\lambda \cdot \sqrt{\frac{N-1}{\chi^2(1-\alpha/2; N-1)}}. \tag{13}$$

Using this equation, we can calculate the value of λ that will result in both the lower bound and the upper bound for given N and α. For example, to find the lower bound:

$$0.04\lambda \cdot \sqrt{\frac{N-1}{\chi^2(\alpha/2; N-1)}} = \sqrt{\lambda} \tag{14}$$

which, for $N = 20$ and $\alpha = 20\%$, gives us the following:

$$\underline{\lambda}(N, \alpha) = \underline{\lambda}(20, 0.20) \approx 895.$$

Similarly, for the upper bound using the same values of N and α, we get

$$\overline{\lambda}(20, 0.20) \approx 383.$$

It seems logical that the larger value for λ would define the lower bound for the standard deviation since more particles can't hurt my estimation. The average of the two is 639, which is the value of λ that gives the highest likelihood of the standard deviation being in the center of the confidence interval. Thus, $\lambda = 639$ gives the highest probability of correctly identifying a mix with $SE = 0.04\lambda$ at least 80% of the time. These numbers for λ are also consistent with those proposed in the TNO report.

What about rejection of bad mixes? According to the TNO paper, a bad mix means that $SE \geq 0.10\lambda$ for the underlying distribution from which we pulled our samples. Using the value for λ calculated above and $N = 20$, we could use a similar method to determine the odds of correctly identifying a bad mix. Keep in mind, this is not exactly how the TNO paper was structured, but the process illustrated here will drive the intuitive arguments that follow.

Are there any problems with this methodology? Sort of. One major issue is that the estimation of λ depends critically on the definition of a "good" mix. For example, if I change the standard for a good mix from ±4% to ±3%, the number of particles needed for the same level of confidence doubles. This is significant because the 4% standard is quite arbitrary. As a result, the calculated value of λ must also be considered arbitrary. This implies that the best value for λ should not be determined *ex ante* by way of simulation, but must be determined after consideration of the context in which the mix is being used. What level of heterogeneity is acceptable in the mixture? Using the same calculations and assumptions as above, if we are comfortable with 10% non-uniformity (as defined in the TNO paper), we require $\lambda = 100$. This is precisely the value proposed by the original Eisenberg paper 30 years ago. (Note that mass of the samples is not explicitly defined in the TNO report, so the expected particle count per sample requires some interpretation here.) There are other examples of similar elements throughout the paper, such as the assumption about measurement error and the arbitrary 80% standard for accepting good mixes. Weakening (or strengthening) any of these assumptions may change our result considerably, which implies to me that we shouldn't have tremendous faith in our numerical results unless we have tremendous faith in our assumptions. None of the assumptions mentioned here seem worthy of this faith.

It should also be noted that the TNO report did not include in their models many of the details mentioned in earlier sections of this paper. For example, the uncertainty about exactly how many particles we add, the time correlated goo covering issue, and diversity in the size and type of feed particles that may result in floating or sinking materials. I am not claiming that these warrant any more consideration than those factors considered in the TNO report, I am merely trying to highlight the delicacy of numerical methods in light of so many unmeasurable details in the feed mixing process. Can the TNO results be trusted? Yes they can be. Are the TNO results an absolute definition of optimal procedure? No they are not–what is optimal varies by situation.

8.1 Example

This is a simple example designed to illustrate the importance of understanding how Type I and Type II errors affect the testing process. Let's assume that we have a standard for 5% odds of Type I error and 20% odds of Type II error (consistent with above). Given these standards, I want to evaluate two different procedures:

1. A case where $N = 20$ and λ varies to achieve this goal;
2. A case where $\lambda = 100$ and N varies to achieve this goal.

Essentially I am asking how far from the true mean can the observed mean be if I am to detect this difference at the stated confidence levels. I can use equation (11) to answer this question. First, plug in $N = 20$, $\lambda = 100$ and my chosen critical values of 5% and 20%. Solving for the observed mean I get $\lambda_1 = 91.26$. Now, to understand which benefits me more–increasing λ or increasing N, I will compare this value to the case where $N = 21$, $\lambda = 100$ and to $N = 20$, $\lambda = 101$. The first gives $\lambda_1 = 91.46$ and the second gives $\lambda_1 = 92.21$. Thus, the additional observation (bigger N) allowed me to detect a difference in means that was smaller by 0.20, while the additional particle allowed me to detect a difference in means that was smaller by 0.95. It thus appears that the additional particle makes more of an improvement. However, in percent terms, the case with $N = 21$ allows me to identify a mean that is 91.46% of the true mean, while the case with $\lambda = 101$ only allows me to identify a mean that is 91.30% of the true mean. Now it appears that the additional observation allows me to identify more precisely difference in the mean that the additional particle. So which is better? The table below gives additional examples.[14]

	$\lambda = 10$	$\lambda = 50$	$\lambda = 100$	$\lambda = 500$
$N = 5$	5.1 (51%)	38.2 (76%)	82.9 (83%)	460.8 (92%)
$N = 10$	6.4 (64%)	41.5 (83%)	87.8 (88%)	472.1 (94%)
$N = 20$	7.4 (74%)	43.9 (88%)	91.3 (91%)	480.2 (96%)
$N = 50$	8.3 (83%)	46.1 (92%)	94.4 (94%)	487.4 (98%)

How should we interpret this table? It appears that increasing N improves our precision more than increasing λ. For example, doubling N from 10 to 20 improves our precision by 10% (from 64% to 74%) for the smallest value of λ. However, a similar doubling of λ from 50 to 100 increases our precision by only 7% (from 76% to 83%) for the smallest value of N. As N and λ grow, an increase in either becomes less valuable.

This example reiterates a point that has surfaced several times in this paper–numerical methods and technical rigor are not worth much if the problem is not understood. Is there actually a measurable difference between 91.46% and 91.30% in terms of feed mix output? If the answer is no, then solving these complicated mathematical problems is not helpful in terms of producing a better product. A summary follows.

9 Summary

Everything outlined above is consistent with the following:

1. Not fewer than 10 samples should be taken from a mixture. Results will improve if samples are taken from several locations at several different times. More samples is always better, but the cost of sampling should also be considered in determining the number of samples taken.

2. The number of tracer particles added must be determined in conjunction with the mass of each sample so that the expected number of particles is not smaller than 10. Assume approximately 100g samples, particle counts of 1 per gram or $\frac{1}{2}$ per gram of feed are appropriate lower bounds. This would result in an expected particle count of 50 to 100 per 100g sample.

3. The mass of a sample should simply be any mass that is small and convenient to collect, keeping in mind that sample mass affects expected particle count. Likely 50g to 250g seems appropriate. Again, this mass should be determined in conjunction with the number of particles added.

4. Samples should be taken from multiple locations within the mixer. Random locations are acceptable, but randomly selecting the sampling location may undermine predictive ability about particle movement.

[14] The cell entries are the value for λ_1 that we can identify with the stated confidence and the percent of λ that each represents. For example, $N = 10$, $\lambda = 10$ shows that we can identify a value of $\lambda_1 = 6.4$ which is 64% of the true mean.

5. Sub-samples are only useful in reducing expected counting error. This benefit is small given the sample size, so sub-samples can be eliminated.

Overall (arbitrary) recommendation: add particles at a rate of 1 tracer per every 2 grams of mix. Take 12 samples, each of 100g, from a unique location (or distinct times during the mixing process). Expected particle count will be 50, samples size is 12.