# AOAC INTERNATIONAL

# FINAL REPORT AND EXECUTIVE SUMMARIES
## FROM THE
# AOAC INTERNATIONAL PRESIDENTIAL TASK FORCE ON BEST PRACTICES IN MICROBIOLOGICAL METHODOLOGY

## CONTRACT DELIVERABLE DUE TO THE U.S. FOOD AND DRUG ADMINISTRATION CONTRACT# 223-01-2464 MODIFICATION 12 AUGUST 10, 2006

*The Scientific Association Dedicated to Analytical Excellence®*


**AOAC INTERNATIONAL**

**481 North Frederick Avenue, Suite 500**
**Gaithersburg, MD 20877-2417 USA**

**Telephone: +1-301-924-7077**

**Fax: +1-301-924-7089**

**Internet e-mail: aoac@aoac.org**

**World Wide Web Site: http://www.aoac.org**

AOAC INTERNATIONAL
Presidential Task Force on
Best Practices for Microbiological Methodology

US FDA Contract #223-01-2464, Modification #12
Task Force Final Report
August 10, 2006

# TABLE OF CONTENTS

---

P. Task Force Membership

1. Steering Committee
2. Detection Limits Working Group
3. Matrix Extension Working Group
4. Sampling Working Group
5. Statistics Working Group
6. AOAC Staff

# AOAC INTERNATIONAL
# Presidential Task Force on
# Best Practices for Microbiological Methodology

US FDA Contract #223-01-2464, Modification #12
Task Force Report

## I. Background

During the past several years, issues have been raised about the limitations of the current AOAC guidelines for validation of microbiological methods. These issues have included the high rate of apparent false negative results when unpaired samples are used, the lack of a definitive acceptable range for "fractional positive" results for qualitative studies and the lack of appropriateness of the guidelines to bacterial toxins. A statistical task force was formed in 2003 to try to address the statistical issues, especially in the case of unpaired samples, and propose solutions. A set of recommendations was drafted, but as yet the recommendations have not been adopted by the Official Methods Board of AOAC. This task force did not address all the issues and concerns previously raised relative to validation of microbiological methods, but focused on the issues of importance to the US FDA as outlined in the objectives of the contract.

Modification #12 of USFDA Contract #223-01-2464 arose from discussions of the limitations of the current AOAC microbiology guidelines and a proposal to re-evaluate the AOAC guidelines was created. Modification #12 of the contract is focused on developing recommendations on the best practices for validation of microbiological methods by an international team of experts. The goal of the group was to consider the technical and statistical aspects of the current AOAC guidelines and ISO 16140 and to recommend new approaches where needed, without regard to harmonization, consensus within AOAC INTERNATIONAL or consensus among international validating organizations.

To lead the project, AOAC appointed Russ Flowers to Chair the Presidential Task Force on Best Practices for Microbiological Methods (BPMM, hereafter referred to as Task Force). A task force structure quickly took shape, comprising a Steering Committee (SC) of key individuals with varying expertise and four Working Groups – Detection Limits (DLWG), Matrix Extension (MEWG), Sampling (SAWG), and Statistics (STWG). The working groups were chaired by Steering Committee members and populated by international experts in that topic area. Great effort was expended to identify technical experts from government, industry, reference laboratories, and academia with varied backgrounds in food safety, quality assurance, clinical diagnostics, veterinary diagnostics and engineering. Not surprisingly, some of these experts also serve on committees for other standards organizations, such as ISO (International Standards Organization), ASTM (American Society for Testing and Materials), CLSI (Clinical and Laboratory Standards Institute, formerly NCCLS), and CEN, the European Committee for Normalization. Care was taken, however, to select scientists and technical experts, without introducing political agendas.

It is interesting to note that ISO TC 34 SC 9 is also considering a revision of the ISO 16140 guidelines and the recommendations from the BPMM task force will provide valuable input to that process. The BPMM project is an important step in the international harmonization of microbiological methods.

The objectives contained in Modification 12 of the US FDA contract were assigned to the working groups as appropriate, and the task of the working groups was to address the objectives by developing recommendations based on sound scientific and statistical principles. The Steering Committee provided guidance to the working groups and served as editors for the final reports. There were some topic areas that overlapped between Working Groups and, therefore, for future publication purposes, the ideas and recommendations of the task force would be best organized by topic area rather than by contract objectives as contained in this report.

This report summarizes the recommendations of the task force and is supported by appended working group reports, which provide the details behind the recommendations. The goal of the BPMM task force, in the short period of time allotted for the contract, was to determine the best practices for validation of microbiological methods and to make recommendations for consideration and further research by AOAC and US FDA. The Task Force did not attempt to create new microbiology validation guidelines as many of the recommendations represent new approaches that must be further evaluated from the perspective of practical application. There is no expectation of adoption of the recommendations by AOAC INTERNATIONAL. After discussing the merits and limitations of the BPMM recommendations, it is hoped that additional work will be funded to further refine and practically evaluate the recommendations presented herein. The Steering Committee recommends first that existing data be used to compare the statistical recommendations to current practice, and then laboratory feasibility studies be conducted to test proposed study designs and sample preparation techniques. These additional efforts would be expected to lead to development of new detailed guidelines for validation of microbiological methods that will be proposed for adoption by AOAC.

## II. Executive summary

The Presidential Task Force for Best Practices in Microbiological Methods (BPMM) makes the following recommendations relative to the objectives of Contract #223-01-2464, Modification #12. A more complete explanation and justification for the recommendations is given in the appended documents. A glossary of terms is found in Appendix O.

*Objective 1: Once a microbiological method has been validated for an array of specific foods and specific strains of a microorganism:*

a) *To what extent can these results be extrapolated to other foods and other strains?*
b) *Are there abbreviated but scientifically/statistically appropriate procedures/protocols by which a validation can be expanded to include additional foods and/or strains?*
c) *How can methods be applied to specific foods, where no validation has been performed?*

The BPMM Task Force recommends new food sub-categorization schemes based on proximate analysis, level and types of background microflora, presence of inhibitors and other characteristics of food matrices that may affect microbial growth, recovery or analytical procedures. Based on the new scheme, varying degrees of verification or validation (from no verification or validation to harmonized collaborative validation) would be required in order to apply a method to a new food matrix. The degree of validation or verification is dependent on how closely related the new matrix is to previously validated matrices and on the current validation status of the method (single lab validated, multiple lab validated or harmonized collaboratively validated).

A list of essential reference organisms and toxins was compiled to address the issue of variability of strains. The organisms and toxins represent antigenic and genomic variability and are recommended to be used as part of the inclusivity testing as appropriate for the method target. Other food-borne isolates should be added to the inclusivity list based on the claimed application of the method.

*Objective 2: What are the scientific/statistical bases for developing performance standards against which the validation of methods should be based*?

The Task Force recommends that performance standards be based on public health objectives (PHOs) and/or fitness for purpose criteria. In general, statistical methods should be used to assist in setting realistic performance standards. These procedures should be based on control of error related to a true negative testing positive (Type I) and error related to a true positive testing negative (Type II). Levels of poor performance that must be detected (with stated probabilities) should be determined. Appropriately determined sample sizes should be used to meet the stated goals. This approach would be a change from current practices where studies are accepted on the basis of standard designs for number of laboratories, materials, and replicates, and standard criteria for suitability of the summary statistics. The design specifications and resulting reliability

estimates should form the basis of applicability statements for test and measurement methods.

*Objective 3: What are reasonable performance standards [criteria] when microbiological methods are to be validated for use for: 1) Attribute (presence/absence) testing (for both 2-class and 3-class sampling plans), 2) Variables (quantitative) testing of batches, and 3) Process control testing of processes or cross-batch testing?*

Whether the method is to be used for attribute or variables testing, performance standards are similar. Ruggedness tests should be performed on the analytical procedure being used. The validation of a test should include estimates of sensitivity, specificity and accuracy. Reproducibility and repeatability should be determined through a detailed collaborative study and ranges of these measures should be published for quality assurance purposes. Results reporting should include a 95 % confidence interval rather than a point estimate of the mean. More detailed and specific recommendations are given in Appendix C.

Recommendations relative to Statistical Process Control (SPC) include Shewhart Charts of control samples with statistical control limits. Standard rules for setting control limits and evaluating control of these charts with respect to Type I and II errors should be followed. Specification limits should not be part of a SPC system. Further details of the SPC recommendations are described in more detail in Appendix F.

*Objective 4: What are the scientific/statistical bases for determining the lower limit of detection for microbiological methods? How is the lower limit of detection validated during the validation of a method? How is the relative performance of a method determined as the lower limit of detection is approached and what is the best way of characterizing this performance?*

The detection limit for qualitative tests is best described as the "$LOD_{50}$", or number of organisms per gram of sample at which 50% of the tests are positive. "$LOD_{50}$" not used in the analytical chemistry sense of LOD and LOQ. It is used in the microbiological sense of an endpoint where the methods are able to estimate around the level of a few particles (bacterium, virus, or genetic macromolecule) per analytical portion. This is possible because virtually unlimited amplifiability of such particles is possible due to their ability to multiply themselves in appropriate conditions. Fifty percent endpoint calculation methods allow for failures to inoculate resulting from imperfect homogeneity at low particle numbers per analytical portion. Such calculations do not assume or require paired samples. $LOD_{50}$ is determined with a nonparametric (distribution free) version of probit analysis, and an experimental study using at least 4 dilutions in which at least two of the dilutions have "fractional positives" in order to estimate better the $LOD_{50.}$ Estimates of other percentiles, such as the $LOD_{90}$ (number of organisms per gram of sample where 90% of results are positive) may be possible in the future development of this approach. The $LOD_{50}$ procedure requires that one dilution level has 0% positive results and one dilution level has nearly 100% positive results (allowing for measurement error in the test laboratories). The associated confidence limits refer to the uncertainty of

the estimated $LOD_{50}$. As proposed here, the $LOD_{50}$ would be calculated for the pool of results from a multi-laboratory study with or without removing outlying laboratories. However, $LOD_{50}$ values could be calculated for each participating laboratory for the purpose of removing outlying laboratories from a study.

The $LOD_{50}$ approach assures that the lower limit of detection for a method is described. However, the number of organisms/g at the limit of detection must be determined on the day of analysis of the test portions. This can be accomplished by analyzing the seed (inoculated or naturally contaminated product diluted into the test samples) on the day of analysis, or by employing a reference method, if one exists, with a known limit of detection in that matrix. In a single method validation, where no reference method exists, the number of organisms/g at the limit of detection must be measured or calculated from measuring the level of organisms in the seed on the day of analysis. Methods for measurement are described.

For quantitative methods, the committee recommends use of the ISO 16140 procedure, which presents limits of detection and quantification as functions of the variability of blank (or very low) samples. The committee recognizes, however, that alternative procedures exist that should be investigated, such as the ISO 11843 series on capability of detection, or the nonparametric analog of that procedure described in the CLSI document EP17-A on Limits of Detection and Quantitation. These procedures recognize the importance of Type I and Type II errors, and that variances of signals from truly negative and truly positive samples can be different. There are related strategies for designing experiments to use the ISO/CLSI approach.

*Objective 5: What are the scientific/statistical bases for developing validation protocols that adequately take into account the biological variation that exists within both the microorganisms and toxins produced by these microorganisms for which methods are developed and the foods which will be analyzed?*

See Objective 1 for discussion of biological variation of microorganisms and categorization of foods.

Validation of methods for toxins produced by microorganisms present a different set of challenges than validation of methods for microorganisms themselves. It is strongly recommended that methods targeting toxins be validated according to the AOAC chemistry guidelines and reviewed by the Chemistry and Microbiology Methods Committees. The methods for preparation of the microbial toxins, dilution into the food matrix and end-user sampling plans may need to be defined by consulting microbiologists, but the validation protocol appropriate for other chemical contaminants should apply.

*Objective 6: What are effective means for articulating the uncertainty associated with microbiological methods?*

Uncertainty in measurements using quantitative procedures is best estimated following an all-inclusive, or "top down" approach. This approach does not attempt to estimate all components of uncertainty separately and it does not require a detailed mathematical model of how those components are combined (ISO 19036). This approach is in contrast to a "bottom up" approach, which requires estimation and combination of variances at all stages of an analysis. This cannot be done routinely, however, so standard, or assumed, variances are used which align the combined estimate to the basic method rather than the analytical result. The "bottom up" approach is likely to underestimate uncertainty due to sources of uncertainty that are not considered. By contrast, the "top down" approach makes no attempt to set generic estimates of uncertainty for specific test methods and rightly aligns the estimate of uncertainty with a specific analysis (or set of analyses). The "top down" approach is consistent with the Guide to the Expression of Uncertainty in Measurement (GUM (2000), Quantifying uncertainty in measurement, BIPM/IEC/IFCC/ISO/IUPAC/OIML, published ISO) principles that allow combination of sources of uncertainty that are difficult to estimate individually.

For qualitative methods, measurement uncertainty for the result cannot be expressed directly – instead, the observed effect is on the probability of reporting an incorrect result. This can be estimated with false negative and false positive rates, for those methods with confirmation procedures. For some measurement procedures, uncertainty can be expressed as the standard error of the $LOD_{50}$, as estimated by the Spearman-Kärber method. This procedure estimates uncertainty where it is most important, which is at the border of the determination of "present" or "absent". The work of ISO Technical Committee 34, Subcommittee 9 is not yet completed, so the STWG recommends active participation in the efforts of this subcommittee.

*Objective 7: How is the statistical basis of a method influenced if the homogeneity of the sample cannot be assumed, particularly at the very low CFU level? How does this influence the performance parameter of a method? How can samples be prepared to minimize this effect? Define optimum procedures for sampling.*

With regard to the statistical validity of low level contamination, a supplemental statistical treatment is presented. This technique, $LOD_{50}$, is not suggested as a replacement for existing tests for significant differences, e.g. Chi Squared, but rather offered as a data treatment that could provide some measurement of the potential variability associated with low level contamination. This may be particularly relevant given that the LOD Working Group believes that low level contamination of matrices, whereby fractional recovery of positives samples occurs within an inoculation level, is the preferred method for defining assay performance.

Even though the homogeneity of the sample cannot be assumed, protocols are presented to minimize this impact. Furthermore, specific protocols are recommended for different categories of food matrices (high moisture food and low moisture food).

*Objective 8: Can a 2-dimensional classification matrix be developed using (1) rating of importance/urgency of intended use and (2) degree of validation, as the dimensional*

*factors? Examples of intended use include: (a) response to a recently emerged microorganism, (b) process control, (c) regulatory screening, (d) regulatory confirmation, and (e) forensic attribution. Examples of validation include: (a) published paper, (b) single laboratory validation, (c) multiple laboratory validation, and (d) full collaborative validation.*

Using the fit for purpose concept, the task force recommends varying degrees of validation for microbiological methods, depending on their intended use (see Appendix M). The study design (number of levels and number of replicates) and type of validation [single lab (SLV), multi-lab (MLV) and harmonized collaborative (HCV), or variants on these per specific design recommendations] will be dependent on whether the method is intended for widespread use, such as a screening method, or regulatory use in one or a few labs and what level of statistical confidence is required by the end user for that intended use. The degree of uncertainty that can be tolerated will depend on many factors; including urgency, cost, availability of confirmatory methods, laboratory (or field) analytical capabilities, etc. However, the degree of confidence required for regulatory, legal and forensic applications will certainly require the highest level of analytical confidence in the data, but may not be as demanding in terms of speed, ruggedness, reproducibility (inter-laboratory variance) and cost per test. Compiling a detailed set of recommendations requires further research and input from end users to define the limits of acceptable performance for each application.

*Objective 9: What are the minimum performance criteria (percentage correct answer on known controls with defined confidence limits) for each factor listed in Objective 8?*

Ideally, performance specifications or acceptance criteria should be based on risk analysis and historical analytical capabilities. In practice, however, developing statistically derived (comparative where available) performance characteristics through validation studies (SLV, MLV, HCV, or studies designed for the intended purposes and needs of the method as recommended by the BPMM study) is more reasonable, allowing potential users of the method to determine its application based on the fitness for purpose concept.

As the number of laboratories increases, the apparent dispersion in results will increase, but properly constructed confidence limits for performance measures will decrease due to having better estimates of the largest source of variability, the inter-laboratory variability. Clearly written package inserts, detailed validation protocols and method training are key factors to controlling this variability in the collaborative study and in the end-user application of the method.

The task force recommends that the level of confidence for different applications of methods be defined using the fit for purpose approach (intended use and end user requirements), and then appropriate validation study designs and verification criteria can be developed within the constraints of practicality. For example, regulatory agencies may determine that a method for detection of *Listeria monocytogenes* in food must have an $LOD_{50}$ of 1-3 CFU/25g at 95% confidence in the single laboratory study. The SLV study design and acceptance criteria would be based on this target value and variance.

Likewise, acceptable inter-laboratory variances can be used to design multi-lab and collaborative studies and set acceptance criteria for these studies. In reality, the needs of the end user and the practicality of the study design must be balanced.

A retrospective analysis of current AOAC, OMA and PTM methods using the $LOD_{50}$ approach would provide a starting point for determining target performance criteria for various intended uses. A corollary to this recommendation is that OMA precollaborative studies must be published as these studies generally provide SLV performance characteristics for a wider variety of matrices than collaborative studies.

*Objective 10: What are the appropriate statistical tools to be used for interpretation of validation studies?*

The Statistics Working Group recommends the use of robust statistical procedures that are not as severely affected by extremely large or small results that can be misleading with more conventional procedures. It also recommends against the removal of outliers from collaborative studies, except for assignable causes. The group recommends review of instances of laboratories in a collaborative study that give indications of having a different application of a method, to see if the method is clearly defined. The committee prefers strong cautions about the concept of "false negative" and "false positive" results due to the vagaries of microbial distribution, the difficulty of confirming all positives and negatives, and the likelihood of misinterpretation. Alternative confirmation procedures should be considered, such as nucleic acid testing. Any estimates of "sensitivity" for low level samples should be corrected statistically for the assumed number of true negatives, based on an assumed Poisson distribution of organisms in the samples.

Chi Squared analysis according to McNemar for paired samples is recommended where a reference method is available. An alternative formula for Chi Squared analysis of unpaired samples should be considered.

It is important that collaborative (interlaboratory) studies be analyzed carefully. The group recommends that current practices of deleting statistical outliers be replaced with a procedure to investigate laboratories that perform differently for an analyte, to see whether the cause can be explained, often because the laboratory had an incomplete understanding of the method. In these cases the method needs to be described better.

*Objective 11: What are the test variables (e.g., number of strains, foods, inoculum levels) that should be considered for each of the factors listed in Objective 8?*

The basic elements of validation studies include inclusivity and exclusivity, characterization of the method performance, and, where applicable, estimate inter-laboratory variation. Additional elements include ruggedness, stability, lot-to-lot variation, and instrument variation (if applicable). The test variables to consider in the design of a validation study include:

- Intended use
- Confidence required for intended use

- Number of inoculum levels
- Number of replicates per level
- Number of labs
- Food claim (from single matrix to multiple categories)
- Analyte claim (Genus, species, or strain)

*Objective 12: Can acceptance criteria be established for methods modification/substitution?*

It is logical to say that when a method is modified, its performance should be at least as thoroughly evaluated as was the original method. However, recognizing that the modification of a method may have benefits other than enhanced performance parameters, a modified method cannot be required to perform better than the original. Further, since there are many applications for methods (screening, regulatory action, process control, etc.) a modified method used for a different application may be acceptable even though some of its performance characteristics may be inferior to the original method. For example, increased sensitivity or broader inclusivity for a "screening method" may result in poorer specificity and/or exclusivity compared with the original method. Therefore, the acceptance criteria for method modification must be based on the claim being made (broader inclusivity, enhanced exclusivity, increased sensitivity, faster time to result) and the intended use of the method (screening, confirmation, process control, etc.).

*Objective 13: Define performance criteria for discrete vs. attribute testing methods.*

This objective was removed from the contract on July 21, 2005, following a request for clarification on April 18, 2005. The BPMM and contractor agreed that this objective is sufficiently covered under Objective 3.

*Additional Recommendation*

As briefly mentioned above in response to objective 3, the task force recommends that ruggedness testing be included as part of every microbiological method validation, similar to what is currently done in the AOAC Research Institute *Performance Tested Methods*$^{SM}$ program. Ruggedness testing involves the deliberate introduction of minor variations in a method procedure. These minor variations should be of a magnitude that might be expected to occur in the hands of the end user. Parameters to be tested might include reagent volumes, reaction temperature and time, enrichment temperature and time, and the like. The specific parameters to be varied would depend on the test method technology and type (quantitative or qualitative method; bacterial, viral or toxin method) and would be determined on a case-by-case basis. Ruggedness testing would target those parameters deemed most critical to method performance in order to provide guidance to the end user regarding the control of those parameters. Ruggedness testing is included in the plans for future research.

*Future Research*

Many of the recommendations and ideas of the task force require further review and development. For a description of the suggested areas for future research, see Appendix N.

III. Appendices

A. Detection Limits WG Report
B. Matrix Extension WG Report
    1. Matrix Extension Essential Organisms List
C. Sampling WG Executive Summary
D. Sampling WG Introduction
E. Sampling WG Enclosure A – Measurement Error
F. Sampling WG Enclosure B – Statistical Process Control
    1. Appendices for Statistical Process Control
G. Statistics WG Executive Summary
H. Statistics WG Report Part 1 – Developing Standards and Validating Performance
I. Statistics WG Report Part 2 – Study Variables
J. Statistics WG Report Part 3 – Uncertainty
K. Statistics WG Report Part 4a – $LOD_{50}$
L. Statistics WG Report Part 4b – $LOD_{50}$ Spearman-Kärber Worksheet
M. Classification Matrix
N. Recommendations for Future Research
O. Glossary
P. Task Force Membership
    1. Steering Committee
    2. Detection Limits Working Group
    3. Matrix Extension Working Group
    4. Sampling Working Group
    5. Statistics Working Group
    6. AOAC Staff

# AOAC INTERNATIONAL
# Presidential Task Force on
# Best Practices for Microbiological Methodology

US FDA Contract #223-01-2464, Modification #12

Executive Summary
Detection Limits Working Group (DLWG)

## EXECUTIVE SUMMARY

For the Limits of Detection (LOD) Working Group, five novel recommendations are offered for consideration:

1. In addition to more precisely defining how preparation and stabilization of inoculated samples should occur, the LOD Working Group has proposed an alternative challenge procedure, Dilution to Extinction (DTE). This methodology has an advantage in that it does not necessarily require the simultaneous analysis of the matrix by a cultural reference method. In certain instances, for example, where the alternative method may be more sensitive than the reference method DTE may have advantages in that the consideration of false positives and false negatives is eliminated. The calculation of the inocula levels of the target analyte organism on the day of initiation of analyses is done only in the organizing laboratory and not in the collaborating laboratories. The performance calculations may be applied to both method comparison and collaborative studies.

2. Alternatives are presented for novel approaches to be taken when the proposed method is suspected of being more sensitive than a reference method.

3. In addition to the considerations in (2) (above), specific consideration is given to molecular based methods validation. While the LOD Working Group is not specifically endorsing a position that molecular based methods are superior or preferred to assays based on other detection technologies, there was a consensus in the Group that confirmation of molecular based assays using traditional cultural procedures may be problematic. Specifically, when molecular assays have improved sensitivity and/or a better limit of detection compared to cultural methods, this may result in the incorrect perception of higher levels of false positive results. On the other hand, molecular assays may be more prone to matrix associated inhibition, leading to reduced assay specificity and a higher incidence of false negative results. Certainly, for routine analysis of samples, alternative methods based on detection technology other than traditional culture or molecular are suitable for many applications. The impact of these assay designs on assay sensitivity and specificity must be considered when establishing the limits of detection for these alternative molecular assays.
4. With regard to the statistical validity of low level contamination, a supplemental statistical treatment is presented. This technique, the $LOD_{50}$, is not suggested as a replacement for existing tests for significant differences, e.g. Chi Square, but

rather offered as a data treatment that could provide some measurement of the potential variability associated with low level contamination. This may be particularly relevant given that the LOD Working Group believes that low level contamination of matrices, including levels providing high and low fractional recovery and levels at or near the endpoint of recovery, is the preferred method for defining assay performance.

5. With regard to quantitative methods validation, the LOD Working Group supports methodologies presented in ISO 16140. It should be noted, however, that when the only available reference method is based on 3 tube MPN analysis, validation may best be performed using the DTE approach. This alternative may be preferred because of the lack of precision associated with MPN measurements.

## Objective 4:

*What are the scientific/statistical bases for determining the lower limit of detection for microbiological methods? How is the lower limit of detection validated during the validation of a method? How is the relative performance of a method determined as the lower limit of detection is approached and what is the best way of characterizing this performance?*

## Summary of Recommendations

It is the opinion of this Working Group that achieving an endpoint of microorganism recovery for the alternative method is the most reliable means for defining method performance and equivalence to a reference procedure. The reference procedure chosen may be a traditional culture procedure or a well-defined rapid method. Endpoint analysis may be applied to detection of bacteria, fungi, viruses and toxigenic compounds, assuming that a detection procedure and a reference procedure are available. In the case of viruses and toxins, enrichment procedures do not apply as these materials do not replicate in culture media.

Even though the homogeneity of the sample cannot be assumed, protocols can be developed to minimize this impact. Furthermore, specific protocols can be designed for different categories of food matrices (high moisture food and low moisture food). The Limits of Detection Work Group has divided the consideration of this topic into:

i. Inoculum preparation and uniform contamination of the matrix of interest
ii. Confirmation of results, particularly when the alternative method may be more sensitive than the reference method
iii. Analyzing and presenting summarized data

Discussion for sections i) and ii) pertain to contract question C1-4 and follow. Discussion for section iii) pertains to contract questions C1-7 and is presented in the next section.

When validating quantitative methods, it is preferable to inoculate the food matrix at three contamination levels. These levels should occur at approximately one logarithm increments within the range expected in the food matrix.

**Preparation of Inoculum and Artificial Contamination**

To determine the detection limit of a qualitative method, the method should be tested on appropriate food samples naturally contaminated or inoculated with microorganisms above and below the anticipated detection limit. For quantitative methods, a minimum of three inoculation levels should be prepared within the expected range of application for the method and the food matrix.

Inoculated food samples for validation of methods should be prepared according to the standard protocols used for AOAC precollaborative and collaborative studies (see article by Andrews, W. A.: J Assoc. Off. Anal. Chem. 1987 Nov-Dec: 70(6):931-6).

Recently, reference materials and certified reference materials have become available that may be more precise than inoculation doses prepared by traditional dilution methods, and may also be utilized when the appropriate levels of target organisms can be obtained (see below).

For viruses and toxins, a high level concentration titered by an accepted reference procedure is prepared in the food matrix, stabilized and then diluted in the food matrix as described by Andrews (1987) for bacteria and fungi.

**Precise Reference Materials and Certified Reference Materials**

Recent developments have made it possible to produce samples that contain precise numbers of microorganisms for use as quantified standards in microbiological analyses. Flow cytometry has been adapted as the platform to analyze and sort cells, and dispense precise numbers of the cells in liquid or freeze dried forms. These precise samples can be used as quantitative Reference Materials.

The International Organization for Standardization has an accreditation system for reference materials known as ISO 34 that enables the production of Certified Reference Materials (CRMs). These CRMs are supplied with a certificate that specifies the amount of bacteria and the variability.

**General Protocols for Limit of Detection Studies**

Determine level of viable target organism(s) in the “seed.” Normally this has been accomplished by MPN procedure with the reference method. However, if the target level of organism present in the “seed” is higher than the background flora, this may be accomplished by non-selective plating or MPN procedures, followed by confirmation of colonies or growth that is typical of the target organism. If a certified reference material

is used, the data provided on the QC certificate should be used for determination of inoculum level.

Once the level of target organisms present in the seed is determined, the seed can be used to prepare method validation samples by either dry or wet dilution methods. This can be done in the organizing laboratory only, or can be done in multiple laboratories by splitting the seed matrix and sending portions to additional laboratories.

1. Wet Dilution Method

    a. Prepare an enrichment of the seed matrix.
    b. Prepare 90 ml enrichments of uninoculated product matrix (same product used to prepare seed) one part matrix to 9 parts enrichment broth.
    c. Set up dilutions of the seed by adding 10 ml of seed homogenate into 90 ml of uninoculated enrichment, thus producing 10-fold serial dilutions of the seed with the same ratio of matrix to broth.
    d. Continue to serial dilute into uninoculated enrichments, until the expected lower limit of detection is exceeded.
    e. Set up multiple enrichments for each dilution for each method being evaluated.

2. Dry Dilution Method

    a. Prepare serial dilutions of seed culture by blending/mixing into the uninoculated portions.
    b. Analyze multiple samples (minimum of 5 per dilution) of each dilution by the reference method and method being evaluated.
    c. Analyze results and determine MPN based on the number of positive and negative tubes at the highest usable dilution according to the MPN rules.

**Limit of Detection Methodology to Determine Low Level Sensitivity: DTE as an Alternative to Use of Reference Culture Method Comparisons**

Certain situations may arise wherein a direct comparison to the reference culture method may not be the best microbiological practice. One such situation is where preliminary data indicate that the reference culture method may not be as sensitive as the alternative method. A second potentially problematic situation occurs when the alternative method and the reference method employ different primary enrichment media. In such a situation in which there is also a need to reach a fractional endpoint to determine the limit of detection, the incidence of positives and negatives would be expected to be random, assuming that proper homogenization of the matrix was accomplished. In this situation the performance data, as expressed as false positive and false negative results, will be very high for both methods, thereby rendering methods performance statistics of questionable utility to the analyst. Presently, the results are reviewed subjectively for "reasonableness." It is possible to employ an alternative study design to eliminate this anomaly in the data. This approach may be termed Dilution to Extinction. In this

approach, a concentrated inoculum is stabilized in a small amount of the matrix of interest. Subsequent dilutions are made in the matrix itself until the recovery by the alternative method becomes, at first, fractional and then progresses to all negative results. This approach also lends itself to validation of virus and toxin measurement assays where comparison to reference methods at low contamination levels may be problematic.

Using this technique, the sample is assayed by the alternative method but there is no unpaired companion sample run with the reference culture method. Instead the enriched sample is confirmed using the appropriate isolation and confirmation technique defined in the reference method. By proceeding in this manner there can be no disagreement in the results that is attributed to only the variability in uniformity of inoculum dispersion at the limit of detection. In such a scheme the sample size per level may be reduced from 5 replicates, but at least 3 replicates and as many as 5 levels may be run to reach fractional and finally completely negative determinations. The number of levels included in the methods comparison study may be increased from the current 2 levels to 5 and the number of replicates per level reduced to 4. For the collaborative study, 2 or 3 levels plus uninoculated controls should be run using 4 samples per level. Data generated by this protocol design are suitable for analysis by the $LOD_{50}$ method, which is described elsewhere in this document.

For methods chosen to be validated using the Single Laboratory Validation (SLV) or a Multi-Laboratory Validation (MLV) involving 2 or more laboratories but not a full Harmonized Collaborative Validation (HCV), the use of the $LOD_{50}$ analysis is an appropriate statistical methodology. This technique may be used in concert with the DTE methodology previously described, but only when an appropriate reference method is not available or different primary enrichment broths are specified. For HCV methods, the DTE approach may be employed in the methods comparison study. It may also be used in the full collaborative study, but only when an appropriate reference method is not available. For the collaborative study the $LOD_{50}$ method may be used to evaluate the data.

**Validation against a Less Sensitive Reference Method: General Approaches**

The need for highly specific and sensitive diagnostic methods for pathogens and toxins has always been critical to food safety, public health, and national security. For the past 50 years, the gold standard methods used to detect bacterial or viral pathogens has been culture-based analyses. Culture-based methods allow for non-directed analyses (i.e., can isolate/detect multiple pathogens from a single plate), are cost effective, have a true limit of detection (LOD) of one viable/culturable organism per sample size, and have been developed so as to be performed by a variety of well-trained professionals, including medical technologists, sanitarians, bacteriologists, and virologists who perform the detection and identification of the infectious agents. With the emergence of alternative, technically more sophisticated diagnostic methods like the polymerase chain reaction (PCR), a system is needed to compare the results from the two distinct platforms to validate or confirm results.

Where applicable, the use of an established reference method is recognized as a preferred means to confirm the results of an alternative method. More recently, however, it has become increasingly apparent that, in some circumstances, the alternative method may be more sensitive that the traditional reference method(s) which are available for confirmation. In such instances it is the opinion of the Working Group that it is appropriate to employ alternative methodology to resolve discrepant results. Such additional efforts would only be required when there was a difference manifest between the alternative and reference methods for an individual sample. Possible approaches could include re-assay of discrepant samples by both methods to confirm the validity of the preliminary determinations, use of a third assay that is based on a different detection technology and for which the performance characteristics of that third method are known, or use of molecular markers if they exist that could confirm the presence of the microorganism in the growth medium. Another attractive alternative is the limit of detection validation presented above, as it eliminates the mandatory use of a reference method.

**Matrix Inhibition of Molecular Methods**

When developing molecular based methods, specific consideration must be given to the potential for matrix related inhibition which may lead to both invalid and/or false negative determinations. A series of matrix addition experiments is in order to define at what, if any level, the matrix from which the isolation is attempted may be capable of inhibiting the amplification reaction itself. Furthermore, it is strongly suggested that, when using molecular based methodology in a routine testing environment, the method must contain an appropriate internal control which will fail to amplify in the presence of a matrix interference event. When validating molecular based methods, specific attention should be paid to results obtained using the proper internal controls to validate that no matrix inhibition has occurred.

## Objective 7:

*How is the statistical basis of a validated method influenced if the homogeneity of the sample cannot be assumed, particularly at a very low CFU level? How does this influence the performance parameter of a method? How can samples be prepared to minimize this effect? Define the optimum procedures for sampling.*

## Summary of Recommendations

**Procedures for Analysis of Data**

The Working Group supports long established methodologies for statistical analysis of test results contained in the guidelines developed by AOAC and contained in ISO16140, with one clarification about the appropriate use of Chi Square statistics. Chi Square calculations should include confirmed positives from presumptive positive samples. In other words, a negative Test Method result, even if confirmed positive, remains classified

as a negative result, and is not considered as a Confirmed positive result when calculating the Chi Square.

It may also be useful to consider an alternate method of compiling data that would allow for the presentation of confidence intervals for the alternative method as an additional statistic. One such technique is proposed below.

**Alternate technique: $LOD_{50}$**

An alternate approach is the $LOD_{50}$ Analysis. To augment and summarize the results of methods comparison and collaborative studies of qualitative microbiology methods, 50% detection endpoint values can be added to the result presentations. These can be calculated from the usual data obtained in such studies by the generalized Spearman-Karber method and would not require additional laboratory work. Some examples are presented. This procedure is also adaptable to experimental designs that differ from that of the traditional AOAC validation study. When analyzing data using $LOD_{50}$ Analysis, the study designs for the methods comparison and the collaborative studies may differ. For example, combining the limit of detection approach with $LOD_{50}$ Analysis, it is preferable to employ 4 or 5 levels of inoculation and 4 replicates per level for the methods comparison study. For the collaborative study, the existing study design of 2 levels plus uninoculated control with 6 replicates per level or, as an option, 3 levels of contamination plus controls with 4 replicates per level, may be preferred. The actual number of levels and number of replicates per level would be determined based on the fit-for-purpose concept and the level of confidence required for the intended use.

## Introduction

When an analyte is at the level of 1 particle per sample, heterogeneous distribution of analyte, as described by the Poisson distribution equation, becomes significant. In microbiology, the particle is either a single autoreplicative organism or a sub-cellular particle (virion or naked nucleic acid) capable of being replicated *in vivo* or *in vitro*. In chemistry, the analyte particle is an atom or molecule, generally in homogeneous solution, and routine qualitative chemical analysis is not performed at the Poisson level.

The limit of detection for qualitative microbiological methods is theoretically 1 organism per analytical portion of a sample (or 0.04 cfu per g in the typical 25-g portion). The endpoint is not a sharp cut-off of the % positive samples versus concentration (MPN/g or cfu/g) curve due to the Poisson distribution effect. As a consequence, the limit of detection curve has a sigmoid-like shape so that at the concentration level of 1 organism per sample only about 63 % of tested samples will test positive. That is, at least about 36% of samples will be true negatives because of failure of incurred or artificial inoculation (spiking) due to the Poisson distribution effect. This effect kicks in significantly below about 3 cfu per g. The region of the curve at about 100% is of particular interest because a significant deficit in positive samples by a test method, relative to an ideal percent positive value of 100%, represents mainly false-negatives, that is a failure to detect positive samples. As the curve approaches 100% asymptotically this

region is difficult to define experimentally. Therefore, it is easier to work at the 50% region of the curve where the curve is steepest and close to linear. Thus the performance of a test method can be defined as the concentration (MPN or cfu per g ± confidence limits) at which 50 % positive samples are observed. This will, allowing for the confidence limits of the estimate, not be less than about 0.028 cfu per g with a 25-g sample analytical portion. If it is significantly larger, it means the method is performing less than ideally because as well as true negatives there are false negatives.

**Fifty Percent Endpoints**

Expressing the limit of detection as the concentration corresponding to 50% positive samples ± confidence limits (usually 95%) is simply a shorthand way of summarizing the performance results for a method. A single number with its limits is used to express the result for a given food matrix. The current tabulations of results more or less nicely compare test and control results statistically but do not clearly tell us how well and with what degree of confidence the methods approach the theoretically maximum possible performance parameter of 1 organism detected per 25 g sample. Neither do they clearly distinguish between true negatives and false negatives. Nor does the current way allow us to easily compare the detection limits for different food matrices and/or analyte strains

Fifty percent endpoint values make efficient use of all the data from control, low and high inoculation levels as well as providing confidence limits to make significance comparisons. In the current method of presenting results, attention is generally focused on the results from only the one of the inoculation levels used that gives the lowest apparent false positive rate.

Some typical and hypothetical examples are given in Table 3. They were generated with an Excel spreadsheet program <Anthony.Hitchins@cfsan.fda.gov>. Approximate endpoint estimates are generally still possible with unusual positive response data patterns that sometimes happen with the 3-inoculation level study design (Table 1) or when a sample with a single level of naturally incurred contamination is used.

As with the traditional treatment of the validation study results, the endpoint calculation is also dependent on the accuracy of the enumeration of the sample contamination. The confidence limits of the 3-tube MPN enumeration typically used are quite broad. This aspect will be addressed in a future document.

The method requires just an extra calculation with the conventional validation data so its adoption would not be a dramatic change. It is interesting to note that the current AOAC experimental design already allows for a gradual detection limit cutoff. Thus, two levels of inoculation are specified so as to try to ensure at least one set of detection results that are not all positive or negative.

Confidence limits will usually be broader, but still tolerable, for the endpoints from pre-collaborative studies than for those from collaborative studies, which have a greater

number of replicates. For example, compare the 5 replicate and 75 replicate confidence limit ranges calculated for proportionate positive responses in Table 3.

**Table 3. Examples of the Application of the $LOD_{50}$ Calculation to Qualitative Microbiology Detection Method Data**

| No. replicates / level (labs x reps / level) | No. positives at control, low and high inoculation levels (MPN/25 g)[a] <1[b] MPN | 2.5 MPN | 10 MPN | 50% Endpoint (cfu/25g) and 95% confidence limit range[c] |
|---|---|---|---|---|
| A. Collaborative study type data | | | | |
| 40(10 x 4) | 0 | 16 | 40 | 1.33 (1.05-1.65) |
| 60 (15 x 4) | 0 | 50 | 60 | 1.93 (1.75-2.10) |
| 60 (15 x 4) | 1[d] | 25 | 58[e] | 3.3 (2.90-3.75) |
| B. Pre-collaborative study type data | | | | |
| 5 (1 x 5) | 0 | 2 | 5 | 3.15 (2.03-5.75) |
| 4(1 x 4) | 0 | 1 | 4 | 2.80 (0.78-10.32) |
| 3 (1 x 3) | 0 | 2 | 3 | 2.33 (1.28-4.23) |

[a]Levels of MPN/25g determined on the day analysis initiated.

[b] The method requires a definite concentration value for a zero positives response. Therefore 1 MPN/25g is chosen for the controls (uninoculated samples) rather than inoculating replicated sample sets with inoculum so dilute as to virtually ensure zero positive-responses. This is fair as the proportion of positive responses approach zero asymptotically according to the Poisson equation.

[c] Divide by 25 to obtain the 50 % detection endpoint expressed as cfu/g of the 25 g sample.

[d] This value exemplifies natural or accidental background contamination at a very low level. The calculated mean background concentration per 25 g, $m = -\text{Ln } P_0 = -\text{Ln } (60-1)/60 = 0.016$ MPN/25 g. Since this is <<< 1, a value of zero positive was used for the control level. For a significant level of naturally incurred contamination either do not calculate a 50% endpoint value or estimate it by assuming appropriate concentrations for 0 and 100 % positive responses as illustrated else where in these footnotes.

[e] The calculation requires a 100 % response value. In this case, a 100% response was very conservatively assumed to be at a 100 MPN/25 g inoculum level.

The $LOD_{50}$ method can be applied to quantitative methods when the Dilution to Extinction experimental protocol is used as described in the next section.

**Illustration of the application of dilution-to-extinction to quantitative microbiology using enumeration method study data**

The dilution to extinction method can be used to enumerate microbes in foods and other matrices. As a work in progress simulation, the natural micro-flora in five sub-samples taken from a homogenized food sample have been enumerated by dilution and plating of 0.1 ml amounts in duplicate on plate count agar. This simulation applies either to one laboratory enumerating 5 samples or to 5 laboratories enumerating one sample each. The colony counts obtained at the various dilutions are displayed in Table 4, Panel A. The calculated mean colony count and 95% confidence interval for the sample are presented in the second and third columns, respectively, of Panel D.

To apply the dilution to extinction method to this example, the actual colony counts for the 5 sub-samples presented in Panel A are transformed into presence and absence data in Panel B. To transform the data, the combined duplicated count for each subsample at a given dilution is designated a positive result if the combined colony count is > 0 and negative if it is 0. The number of positives per five sub-samples at each dilution is totaled (Panel B). The totals per dilution are processed by the Spearman-Kärber $LOD_{50}$ calculation. The $LOD_{50}$ result and its confidence interval represent the *limit of dilution* at which 50% of the diluted sub-samples are culture positive (Panel C). The reciprocal of the product of the $LOD_{50}$ value, or each of its uncertainty limits, and the volume of dilution tested (0.2 ml in this example) multiplied by a constant m (m=0.69) give the dilution to extinction estimate of the mean count of the sub-samples (Panel D, columns 4 and 5). The constant m represents the mean count per test volume corresponding to 50 % negative (or positive) results. It is obtained from the Poisson relationship, $P_o = e^{-m}$, for the proportion of negative cultures. The data in Panel B can also be used to calculate a not significantly different 5-tube MPN result of 6500/g with a confidence interval of 1800 – 20,000. Thus the dilution to extinction method performs as well as the MPN. However, it may be technically superior, at least in some applications, because it is a plate as opposed to a broth culture method.

It can be seen from Panel D of Table 4 that the dilution-to-extinction and plate-count results are not significantly different. The dilution to extinction method was subjected to the experimental design of the plate count method in order to directly compare the two calculations. That design is not optimal for the dilution to extinction method but it can be modified appropriately for optimization.

| **Table 4. Enumeration by Dilution to Extinction using the 50% Extinction Value** | | | | |
|---|---|---|---|---|
| | | | | |
| **A. Raw data** | | | | |
| | Colony count (CFU/ 0.1ml of stated dilution ) | | | |
| Sub-Sample | $10^{-1}$ dilution | $10^{-2}$ dilution | $10^{-3}$ dilution | $10^{-4}$ dilution |
| 1 | 49;51 | 4;5 | 1; 0 | 0; 0 |
| 2 | 62;72 | 6;7 | 1;1 | 0; 0 |
| 3 | 40;43 | 4;4 | 0; 0 | 0; 0 |
| 4 | 62;65 | 6;7 | 1; 0 | 0; 0 |
| 5 | 31;40 | 3;4 | 0;2 | 0; 0 |
| | | | | |
| | **B. Raw Data Processed for $LOD_{50}$ Calculation** | | | |
| | Positives (mean CFU > 0 per 0.2 ml) per 5 samples at stated dilution | | | |
| Sub-Samples | $10^{-1}$ dilution | $10^{-2}$ dilution | $10^{-3}$ dilution | $10^{-4}$ dilution |
| 1-5 | 5 | 5 | 4 | 0 |
| | | | | |
| | **C. $LOD_{50}$ Result** | | | |
| Sub-Samples | Dilution for 50% positive | | 95% confidence dilution interval | |
| 1-5 | $5.013 \times 10^{-4}$ | | $(1.838 - 13.677) \times 10^{-4}$ | |
| | | | | |
| | **D. Comparison of Counts by Plating and by $LOD_{50}$** | | | |
| | Av. CFU/plate x $[1 / (0.1 \times 10^{-1})]$ | | 0.69x1/[volume tested x $LOD_{50}$]# | |
| Sub-Samples | Av. CFU /g | ±1.96SD CFU | | |
| 1-5 | **5150** | 2430 - 7816 | **6900** | 2518 - 18800 |
| | | | | |

## Application of $LOD_{50}$ to existing methods

The $LOD_{50}$ may best be applied when the study design includes several levels of inoculation as described above. In situations where previous studies exist which were conducted under protocols which required two levels of inoculation plus negative controls, this analysis technique may be applied retrospectively. The confidence intervals of the data may suffer somewhat from the reduced number of levels; however, valid interpretations may be drawn. This would be the situation with the many AOAC collaboratively studied microbiological Official Methods of Analysis. In these studies, generally two levels plus an uninoculated control level were run. The data from such studies may be retrospectively analyzed to calculate confidence intervals.

# AOAC INTERNATIONAL
# Presidential Task Force on
# Best Practices for Microbiological Methodology

US FDA Contract #223-01-2464, Modification #12

Executive Summary
Method/Matrix Extension Working Group (MEWG)

## A. METHOD VALIDATION - BIOLOGICAL VARIATION OF MICROORGANISMS AND TOXINS

Contract Objective being addressed:

Objective 5

*What are the scientific/statistical bases for developing validation protocols that adequately take into account the biological variation that exist within both the microorganisms and toxins produced by these microorganisms for which methods are developed [and the foods which will be analyzed]?*

Summary of Recommendation

The MEWG has compiled a list of species and strains, taking into account immunological and genetic variation, that it recommends to be used in method validation studies. As new variants are discovered and made available, they can be added to the list.

Details of Recommendation

In order to assess the ability of a method to detect or quantify its target specifically, the method must be challenged with a variety of target and non-target organisms or molecules. The details of the challenge will depend on two main factors:

1. The method type – For example, immunoassay, molecular method or metabolic-based method (e.g., chromogenic agar) and
2. The target – For example, genus (e.g., *Salmonella* spp.), species (e.g., *Listeria monocytogenes*), a group of organisms (Enterohemorrhagic *Escherichia coli*), a single molecule or a group of molecules (e.g., staphylococcal enterotoxins).

Challenge studies will be designed to test variations of the target as appropriate to the method type, as well as test the most common food and environmental strains. For example, the presence of *E. coli* O157:H7 can be presumptively targeted by detecting the O157 antigen, the verotoxin genes or the verotoxins. Challenge studies could include non-motile strains, genetic variants, verotoxin expression variants (various levels of expression), strains that produce VT1, VT2 and VT1+2 and strains that produce VT1 variants and/or VT2 variants, depending on the method type.

In order to compare the performance of one method to another, a set of common strains must be used for inclusivity. Appendix 9 is a table of microorganism strains (including bacterial, viral and parasitic strains) and toxin types that can be used as reference strains for challenge studies during the initial validation of a method. This table takes into account known genetic and immunological variations of microorganisms and toxins. It is

recognized that shipping microorganisms can be problematic and, therefore, the table is comprised of reference strains found in the American Type Culture Collection (ATCC), the National Center for Type Cultures (NCTC) as well as other collections. An increasing number of strains are being labeled as bioterrorism agents, and in this case it may be beneficial to use a contract laboratory with the proper facilities and licenses to obtain, maintain and use these strains.

Inclusivity studies, therefore, would comprise a small number of appropriate reference strains chosen from Appendix 9 and a larger number of food isolates for validation of the target analyte(s) claim. The reference strains will allow comparison of one validation study to another, but it is recognized that food isolates are the most relevant strains for validating the claims of a method. The food isolates must, however, be well characterized relative to the method technology (genotype and/or phenotype characterization). By testing a sufficient number of variants within the target group, using reference strains and food isolates, one can be more confident in the comparative test results between methods and in being able to extend the method to additional strains.

The number of target organisms or toxins required for inclusivity testing is dependent on the target scope and the known variants available and, therefore, cannot be generalized. For inclusivity studies, AOAC currently recommends 100 strains for validation of methods for the detection of *Salmonella*, and 50 strains for methods for detection of pathogens (and other organisms) other than *Salmonella*. For some pathogens, for which number and/or availability of strains may be limited (for example, hepatitis A virus), or which have been highly characterized on the genetic level, it may be appropriate to use less than 50 strains for inclusivity testing. It is recognized, however, that inclusivity testing for a method targeting a genus should logically require more strains than a method targeting a species.

## B. METHOD VALIDATION - VARIATION OF FOODS

Contract Objective being addressed:

Objective 5

*What are the scientific/statistical bases for developing validation protocols that adequately take into account the [biological] variation that exist within [both the microorganisms and toxins produced by these microorganisms for which methods are developed and] the foods which will be analyzed?*

Summary of Recommendation

Food commodity groups proposed by ISO and AOAC were extensively re-categorized based on physical structure, chemical parameters, bacterial load or other factors that would likely impact microbiological recovery and hence require different analytical approaches (See Appendices 1-8). The new food classification schemes are recommended to be used as the basis for method validation.

Details of Recommendation

The food categories found in ISO 16140 and the AOAC Guidelines are sub-categorized on the basis of broad food categories and microbial load and recovery. To make a broad

food claim, the AOAC Guidelines require 20 foods covering at least 6 of the 9 food categories in the precollaborative or single lab validation. ISO 16140 requires testing of three food types from each of five categories for an "all foods" claim. It has been observed, however, that methods validated with such broad claims do not necessarily perform well with new matrices that were not included in the validation study. Often times, matrix extension is not predictable within a food category. Therefore, the categories and sub-categories were redefined in an effort to make matrix extension more predictable.

The factors taken into consideration for sub-categorization include those that can affect microbial recovery or detection. Immunological and molecular methods can be affected by different factors, so both were considered. The factors used to sub-categorize foods included lipid or fat, protein, fiber, water activity or moisture level, presence of PCR inhibitors, microbial load, type of processing if any, presence of preservatives, surface structure, pH, salt and sugar. See Appendices 1-8 for the breakdown of each food category.

The concept of matrix extension is complicated and there are no hard and fast rules about how food products can or should be categorized for this purpose. Furthermore, this is a new way of thinking for most traditional food microbiologists. As we move away from qualitative assays towards quantitative, molecular-based methods, there will certainly be developments on this front. This document was constructed using the input of food microbiologists with expertise in a wide variety of matrices, as applied to many different detection methods, and is meant to be a guideline for future deliberations. Irrespective of how closely related a non-validated matrix may be to a validated matrix, the Matrix Extensions working group recommends that there needs to be some type of in-house verification conducted before using the alternative method on any previously un-validated matrix. This is particularly important when results are to be used for regulatory purposes.

To validate a category of foods, it is proposed that one matrix from each sub-category must be tested. This will no doubt increase the amount of work required to claim certain food categories, but will also increase the likelihood that the method is applicable to all types of foods in that category.

## C. METHOD/MATRIX EXTENSION

Contract Objective being addressed:

Objective 1

*Once a microbiological method has been validated for an array of specific foods and specific strains of a microorganism:*

a) *To what extent can these results be extrapolated to other foods and other strains?*

b) *Are there abbreviated but scientifically/statistically appropriate procedures/protocols by which a validation can be expanded to include additional foods and/or strains?*

c) *How can methods be applied to specific foods, where no validation has been performed?*

Summary of Recommendations

By using the food sub-categorization schemes shown in Appendices 1-8, matrix extension is simplified. The degree of validation required to extend a method to a new matrix is dependent on (1) how closely related the new matrix is to those that have been included in the initial validation, and (2) the level of validation initially performed (single lab, multi-lab, harmonized collaborative).

Details of Recommendations

With the exception of a few key methods (e.g., culture-based detection of *Salmonella* in "all foods," and and culture-based methods for the detection of *Listeria* spp. and *Escherichia coli* O157:H7 in broad categories of foods), when a method is validated by AOAC INTERNATIONAL or by the AOAC Research Institute, the claim is limited to those foods actually tested in the single lab validation (SLV), multi-lab validation (MLV) and/or in the harmonized collaborative validation (HCV). With sufficient representation within a food category, a claim can be made for that food category, although the actual foods tested must be clearly stated. There is a clear need to provide additional guidelines for matrix extension after appropriate laboratory validation has been completed.

When extending a validated method to a new matrix, then, it is logical to propose that the more closely related a new matrix is to a validated matrix, the higher the probability that the new matrix will perform similarly. The Matrix Extension Working Group has expended great effort to sub-categorize foods on the basis of their impact on microbial growth and recovery, as well as potential inhibitory effects, on rapid method technologies. These new sub-categorization schemes will be the basis for investigating proper protocols for matrix extension.

There are three situations to consider:

1. The new matrix is within the same sub-category or group (where there is no additional sub-category) as a validated matrix
2. The new matrix is in a new sub-category/group, but within the same class as a validated matrix
3. The new matrix is in a new class not previously validated

**Table 1. Data Requirements for Matrix Extension**

| If new matrix is: | Then data required are: | | |
|---|---|---|---|
| | For SLV Method | For MLV Method | For HCV Method |
| Situation 1: within the same sub-category/group as a validated matrix | None | None | None |
| Situation 2: in a new sub-category/group, but within the same class as a validated matrix | Verification | Verification | Single Lab Validation |
| Situation 3: in a new class not previously validated | Single Lab Validation | Multiple Lab Validation | Harmonized Collaborative Validation |

The data set required to extend a validated method to a new matrix is summarized in Table 1. The extension of a validated method to a new matrix in Situation 1 should be the most predictable and, therefore, require no further experimentation. Due to the proposed scheme of sub-categorization of foods, all foods within the same sub-category are expected to perform equivalently. Therefore, if the proposed new matrix falls into the same sub-category (see the appended tables) as a previously validated matrix, the proposed matrix does not require a verification or validation study. The method can be applied to the new matrix without further study. While formal verification is not required in situation 1, it is good laboratory practice to perform some preliminary experiments to demonstrate that the method performs as expected with any new matrices being analyzed by the laboratory.

Extending a method to a matrix in a different sub-category/group within the same class(Situation 2) is less predictable than Situation 1 and, therefore, would require a basic level of experimentation. In Situation 2, a limited study to verify, rather than validate, the utility of the method for that matrix would be sufficient for SLV or MLV methods Verification would reveal gross effects on method performance such as the presence of inhibitors. An HCV would require a Single Lab Validation study for matrices in Situation 2.

Situation 3, in which a new class is being examined, would require full validation for SLV, MLV or HCV methods. Thus, an SLV method would require an SLV study, an MLV method would require an MLV study, and an HCV method would require an HCV study to extend the method to the new matrix.

The verification of method performance with a new matrix is intended to assure the user that the new matrix will produce neither high false positive rates (matrix is free from cross reactive substances) nor high false negative rates (matrix is free of inhibitory substances). To this end, a protocol is proposed in which the new matrix is spiked with a single strain of target organism chosen from the attached Strain list (Appendix 9) or a

single toxin type at a level 10 to 50 times higher than the LOD for the most similar validated matrix. Six replicates of the inoculated matrix and six replicates of the uninoculated matrix are tested and confirmed by both the alternative and the reference method. If no false positive or false negative results are obtained, then the new matrix is verified. If either false positive or false negative results are observed, then the study is expanded to a Single Laboratory Validation to define the operating characteristics of the method with the new matrix.

The Single Laboratory Validation (SLV) should follow the study design from the original validation study and should measure the 50% LOD for the new matrix being studied. The spike levels should be adjusted according to the expected LOD for the assay being evaluated and for the new matrix such that fractionally positive results are obtained for at least one of the levels.

For MLV and HCV method extension to a new food category, a Single Laboratory Validation is first carried out to determine the 50% LOD of the method with the new matrix as described above. These data provide the basis for the MLV or HCV study.

When extending a method to foods containing preservatives such as sodium benzoate, it is recommended that at least one verification study be performed in all cases.

All studies should be carried out in parallel with a reference method, when one is available, in order to compare the $LOD_{50}$ values of the two methods. A test for statistical difference, such as Chi-Square, can be applied to compare the data sets where the same set of samples has been used for both methods (paired samples).

## D. ACCEPTANCE CRITERIA FOR METHOD MODIFICATION

Contract Objective being addressed:

Objective 12

*Can acceptance criteria be established for methods modification/substitution?*

Summary of Recommendation

It is logical to say that when a method is modified, its performance should be at least as good as the original method. Recognizing that the modification of a method may have benefits other than enhanced performance parameters, such as time to result or ease of use, a modified method cannot be required to perform better than the original. Further, since there are many applications for methods (screening, regulatory action, process control, etc.) a modified method used for a different application may be acceptable even though its performance may be inferior to the original method. The MEWG, therefore, defers the subject of acceptance criteria to the Steering Committee.

**Official Standards or Guidance Documents referenced:**

1. Philip Feldsine, Carlos Abeyta and Wallace H. Andrews. 2002. AOAC INTERNATIONAL Methods Committee Guidelines for Validation of Qualitative and Quantitative Food Microbiological Official Methods of Analysis. Journal of AOAC International 85 (5): 1188-1200.
2. ISO Standard 16140, *Protocol for the Validation of Alternative Methods*.
3. USDA Nutrient Data Laboratory http://www.nal.usda.gov/fnic/foodcomp/, April 2005

Appendix 1
Category 1 – Meat and Poultry

| Class | Sub-category | Examples |
|---|---|---|
| A<br>Water < 20% | None | Dehydrated Beef, Dehydrated Broth, |
| B<br>Water between 20 – 80% | B.1<br>(Protein < 10%) | Most prepared foods, containing large amount of carbohydrates (15-30%), e.g. Frozen Entrées |
| | B.2<br>(Protein 10-30%, Lipid 10-30%, Cooked) | Hot Dogs, Bologna, Corned Beef<br>Meat Patties |
| | B.3<br>(protein 10-20%, lipid 10-30%, raw) | Raw Chicken, Raw Beef,<br>Raw Pork Ground Beef |
| | B.4<br>(protein 10-20%, lipid 10-30%, Marinated or spiced raw) | Raw Chicken, Raw Beef,<br>Raw Pork |
| | B.5<br>(protein 10-35%, lipid < 10%, low fat, cooked) | Chicken Drumstick, Roast Beef- (Cured, Dried), Beef Brisket- Lean, Braised. |
| C<br>Water 80-90% | | Most Soups, Canned Baby Foods |
| D<br>Water >90% | | Most Broth. |

Category 2 – Fruits and Vegetables Appendix 2

| Class | Group | Sub-category | | |
|---|---|---|---|---|
| A: Fresh | *A.1: Low pH (<3.0-4.9)* most fruits, including citrus, berries, apples | A.1.1: Smooth product consistency | A.1.2: Rough/irregular product consistency | |
| | *A.2: Reduced pH (5.0-7.0)* melons, many vegetables | A.2.1: Smooth product consistency e.g. grapes, apples, squash | A.2.2: Rough/irregular product consistency e.g. berries, lettuce | |
| B: Frozen, and heat processed products | *B.1: Low pH (<3.0-4.9)* most fruits, including citrus, berries, apples | B.1.1: Smooth product consistency | B.1.2: Rough/irregular product consistency | |
| | *B.2: Reduced pH (5.0-7.0)* melons, many vegetables | B.2.1: Smooth product consistency e.g. grapes, apples, squash | B.2.2: Rough/irregular product consistency e.g. berries, lettuce | |
| C: Juice and Juice Concentrates | *C.1: Low pH (<3.0-4.9)* most fruit juices, including citrus, berries, apples, tomato | C.1.1: High °Brix (>60) high sugar fruit juice concentrates | C.1.2: Moderate °Brix (40-59) low sugar fruit juices | C.1.3: Low °Brix (<40) most fruit juices |
| | *C.2: Reduced pH (5.0-7.0)* most vegetable juices | C.2.1: High °Brix (>60) high sugar vegetable juice concentrates | C.2.2: Moderate °Brix (40-59) low sugar vegetable juices, | C.2.3: Low °Brix (<40) most vegetable juices |
| D: Dry and Low Moisture Products | D.1: Very low $a_w$ (≤0.60) (raisins, apricots) | | | |
| | D.2: Reduced $a_w$ (>0..60) (dried vegetables, dried apples) | | | |
| E: Fermented fruit and vegetable products | (e.g., sauerkraut) No further sub-categorization | | | |
| F: Nutmeats | No further sub-categorization | | | |

*Note: While compounds that can interfere with detection assays may be associated with many if not most food matrices, the inhibitory effect of fruit and vegetable matrices may be particularly troubling. Users are encouraged to consult the literature and perform preliminary experiments to demonstrate that the method performs as expected with new matrices of concern before routinely using the method on those matrices.

Category 3 – Dairy Products Appendix 3

| Class | Group by Water Content | Sub-category by Fat Content | Representative Examples |
|---|---|---|---|
| A. Fermented and Non-Fermented Products* | A.1: (<20%) | A.1.1 (<10%) | Milkshake powder, Buttermilk-dried, Dry non-fat milk, Dry Whey, casein** |
| | | A.1.2 (10-30%) | Dry, whole milk, Grated Parm. Cheese |
| | | A.1.3 (30-70%): | Powdered cream |
| | | A.1.4 (>70%): | Butter, margarine |
| | A.2: (20-50%) | A.2.1 (<10%) | Canned Condensed milk |
| | | A.2.2 (10-30%) | American cheese, pasteurized, Brie, Gouda, Monterey, Colby, Hard and Soft goat Cheese |
| | | A.2.3 (30-70%): | Margarine |
| | A.3: (50-80%) | A.3.1 (<10%) | Ice Cream, Low-fat Yogurt, Ricotta, Milkshake, Evap. Milk |
| | | A.3.2 (10-30%) | Sour Cream, Whipped cream, Mozzarella |
| | | A.3.3 (30-70%) | Heavy Cream |
| | A.4: (>80%) | A.4.1 (<10%) | Fat free Half and Half, Whey-fluid, Plain Yogurt, Cottage cheese (reg and low fat, Milk substitute, buttermilk, Milk |
| | | A.4.2 (>10%) | Half and Half reg. |

*To interpret the table, the user must first categorize the dairy product in question as fermented or non-fermented. Thereafter, the sub-categorization based on water and fat content can be used. Note that the representative examples are not meant to be exhaustive and there are many other products which might fit into any one subcategory.

**The detection of certain pathogens in some products may differ based on methods of manufacture (e.g. *Salmonella* detection in non-fat dry milk or casein products). Consult the literature before applying matrix extension in these particular applications.

Category 4 – Egg Products Appendix 4

| Class | Group | Examples |
|---|---|---|
| A | < 5% salt or sugar added | shell eggs, whole eggs, egg yolks, egg whites, dried whole egg, dried egg yolk, dried egg whites, egg substitutes |
| B | ≥ 5% salt or sugar added | whole eggs, yolks, or egg products |

Category 5 – Miscellaneous Foods Appendix 5

| Class | Group | Examples |
|---|---|---|
| A. Cereals and Grains | A.1<br>Flour and dry mixes | |
| | A.2<br>Unbaked, viable-yeast leavened products | |
| | A.3<br>Dough, batter, and baked products | |
| B. Chocolate* | B.1<br>Fat < 20% | Cocoa powders, all Confectionery products, Ingredients, Coatings, Chocolate bars |
| | B.2<br>Fat > 20% | Cocoa powders, all Confectionery products, Ingredients, Coatings, Chocolate bars |
| C. Pasta | C.1<br>Raw, Fresh | |
| | C.2<br>Raw, Dried | |
| | C.3<br>Cooked | |
| D. Dressings, Condiments and Marinades | D.1<br>Do not require refrigeration for microbiological safety | Contain preservatives, Aw <0.85 or pH < 4.0 |
| | D.2<br>Require refrigeration for microbiological safety | Specified by manufacturer (does not apply to products that need refrigeration after opening) |
| E. Soy Products | None | |

*Note: While compounds that can interfere with detection assays may be associated with many if not most food matrices, the inhibitory effect of chocolate and chocolate products may be particularly troubling. Users are encouraged to consult the literature and perform preliminary experiments to demonstrate that the method performs as expected with new matrices of concern before routinely using the method on those matrices.

Category 6 – Seafood Appendix 6

| Class | Group | Sub-category | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A. Finfish | A.1 Fresh Water | A.1.1 Raw Fresh | A.1.2 Raw Frozen | A.1.3 Cooked | A.1.4 Dried | A.1.5 Cold Smoked, Marinated or Cured | A.1.6 Carbon Monoxide (CO) Treated | A.1.7 Fermented |
| | A.2 Salt Water | A.2.1 Raw Fresh | A.2.2 Raw Frozen | A.2.3 Cooked | A.2.4 Dried | A.2.5 Smoked, Marinated or Cured | A.2.6 CO Treated | A.2.7 Fermented |
| B. Molluscan Shellfish* | | B.1 Raw Fresh | B.2 Raw Frozen | B.3 Cooked | B.4 Marinated or Hot Smoked | B.5 High Pressure Treated | | |
| C. Crustaceans | | C.1 Raw Fresh | C.2 Raw Frozen | C.3 Cooked | | | | |
| D. Squid/Octopus | | D.1 Raw Fresh | D.2 Raw Frozen | D.3 Cooked | | | | |

*Note: While compounds that can interfere with detection assays may be associated with many if not most food matrices, the inhibitory effect of molluscan shellfish is particularly well characterized. Users are encouraged to consult the literature and perform preliminary experiments to demonstrate that the method performs as expected with all molluscan shellfish matrices on which it is to be applied. In particular, the following are known to impact ability to recover target organisms: (1) differences (seasonal, storage, or processing related) in biochemical composition of the animal tissue; (2) differences in background flora arising from harvest water conditions (mostly seasonal) and temperature history of the product.

Category 7 – Animal Feed Appendix 7

| Class (Dry Matter) | Group (Crude Fiber) | Sub-category (Crude Protein) | Representative Examples |
|---|---|---|---|
| A. DM>75% | A.1 CF<10% | A.1.1 CP<20% | Cereal grains<br>Dried bakery waste<br>Dried whey |
| | | A.1.2 CP>20% | Bean varieties<br>Blood meal<br>Soybean meal<br>Distillers grains<br>Feather meal<br>Meat meal<br>Meat & bone meal<br>Poultry by-product |
| | A.2 CF>10% | A.2.1 CP<20% | Alfalfa hay<br>Clover hay<br>Barley hay<br>Cottonseed hulls<br>Dried beet pulp<br>Dried apple pomace<br>Wheat bran<br>Oat hulls |
| | | A.2.2 CP>20% | Canola meal<br>Sunflower meal<br>Cottonseed meal<br>Coconut meal<br>Avocado seed meal |
| B. DM<75% | B.1 CF<10% | B.1.1 CP<20% | Bread by-products<br>High moisture corn<br>Cane molasses<br>Beet molasses<br>Citrus molasses |
| | | B.1.2 CP>20% | Wet distillers grain (corn) |
| | B.2 CF>10% | B.2.1 CP<20% | Fresh alfalfa<br>Fresh clover<br>Wet Apple pomace<br>Wet beet pulp<br>Fresh grasses<br>Sugar beet tops<br>Ensiled forages |
| | | B.2.2 CP>20% | Wet distillers grain (sorghum, barley) |

Category 8 – Spices* Appendix 8

| Class | Group |
|---|---|
| 1 | Black Pepper, White Pepper, Caraway Anise, Celery, Cumin, Dill, Fennel, Nutmeg, Coriander, Ginger, Paprika |
| 2 | Onion, Garlic |
| 3 | Oregano, Cinnamon, Allspice |
| 4 | Thyme, Marjoram, Basil, Sage, Rosemary |
| 5 | Red Pepper, Chili Pepper |
| 6 | Cloves |

*Spices are a particularly troubling category as many contain uncharacterized naturally occurring compounds that can interfere with detection assays. Although this table can serve as a guideline, the user is strongly encouraged to consult the literature and perform preliminary experiments on each spice to demonstrate that the method performs as expected with new matrices before routinely using the method on those matrices.

Appendix 9 – See Excel Spreadsheet

**Appendix 9. Essential Reference Organisms and Toxins***

| Genus | Species/Serotype | Strain | Source | Additional Recommended Species/Strains: Species | Strain/Serotype | Source | Exclusivity Species |
|---|---|---|---|---|---|---|---|
| ***Aeromonas*** | *A. hydrophila* | ATCC® 49140™ | ATCC® | | | | |
| | *A. hydrophila* | ATCC® 7965™ | ATCC® | | | | |
| ***Bacillus anthracis*** | *B. anthracis* | AMC strain | BEI Resources | | | | |
| | *B. anthracis* | Ames strain | BEI Resources | | | | |
| Controlled Access | *B. anthracis* | Davis | BEI Resources | | | | |
| | *B. anthracis* | Kruger B1 | BEI Resources | | | | |
| ***Bacillus cereus*** | *B. cereus* | ATCC® 13061™ | ATCC® | | | | |
| | *B. cereus* | ATCC® 13061™ | ATCC® | | | | |
| | *B. cereus* | ATCC® 10876™ | ATCC® | | | | |
| | *B. cereus* | enterotoxigenic producing strains - FDA TJL-14 | FDA | | | | |
| | *B. cereus* | emetic toxin producers | Pending final method research | | | | |
| ***Brucella*** | *B. abortus* (CO2 dependent and independent) | | BEI Resources | | | | |
| | *B. canis* | | BEI Resources | | | | |
| Controlled Access | *B. melitensis* | | BEI Resources | | | | |
| | *B. suis* | | BEI Resources | | | | |
| ***Campylobacter*** | *C. fetus subsp. fetus* | ATCC® 15296™ | ATCC® | | | | |
| | *C. jejuni subsp. jejuni* | ATCC® 33291™ | ATCC® | | | | |
| | *C. coli* | ATCC® 43478™ | ATCC® | | | | |
| | *C. upsaliensis* | | | | | | |
| | *C. jejuni subsp. venerealis* | ATCC® 19438™ | ATCC® | | | | |
| | *C. lari* | ATCC® 35222™ NCTC 11457 | ATCC®/ NCTC | | | | |
| | *C. jejuni* subspecies *doylei* | ATCC® 49351™ NCTC 11924 | ATCC® /NCTC | | | | |
| | *C. jejuni* subspecies *jejuni* | CIP 702 | BEI Resources | | | | |
| | *C. jejuni* subpecies *jejuni* | NCTC 11168 | BEI Resources | | | | |
| ***Clostridium*** | *C. perfringens* | ATCC® 3624™ | ATCC(r) | | | | Requires Export permit outside of the US |
| | *C. perfringens* Hobbs serotype 2 | NCTC 8238 | | | | | Requires Export permit outside of the US |
| | *C. perfringens* Hobbs serotype 3 | NCTC 8239 | | | | | Requires Export permit outside of the US |
| | *C. perfringens* Hobbs serotype 13 | NCTC 10240 | | | | | Requires Export permit outside of the US |
| | *C. perfringens* | ATCC® 12919™ | ATCC® | | | | Requires Export permit outside of the US |
| | *C. botulinum* Type A | No longer available from ATCC® or by reference strain number | BEI Resources | | | | former ATCC® 25763™ |
| | *C. botulinum* Type B | No longer available from ATCC® or by reference strain number | BEI Resources | | | | former ATCC® 17848™ |
| | *C. botulinum* Type E | No longer available from ATCC® or by reference strain number | BEI Resources | | | | former ATCC® 9564™ |
| | *C. botulinum* Type F | No longer available from ATCC® or by reference strain number | BEI Resources | | | | former ATCC® 35415™ |
| | *C. butyricum* Type E | No longer available from ATCC® or by reference strain number | BEI Resources | | | | former ATCC® 43755™ |
| | *C. argentinense* Type G | No longer available from ATCC® or by reference strain number | BEI Resources | | | | former ATCC® 27322™ |
| ***Coxiella burnetii*** | | | | | | | |
| ***Cryptosporidium*** | *C. parvum* (For DNA only) | PRA-67D, IOWA strain | ATCC® or Waterborne, Inc. | | | | |
| | *C. parvum* (Oocyst) | IOWA strain | Bunch Grass Farm | | | | |
| | *C. parvum* (bovine genotype) | IOWA strain | | | | | |
| | *C. hominis* | | | | | | |
| ***Cyclospora*** | C. cayetenansis | Any isolates from feces from naturally infected individuals. Currently, no isolate or strain is maintained in a laboratory. | | | | | |
| ***Enterobacter sakazakii*** | *E. sakazakii* | ATCC® 51329™ | ATCC® | | | | |
| | *E. aerogenes* | ATCC® 13048™ | ATCC® | | | | |
| ***Escherichia coli*** | *E. coli* Biotype1 | ATCC® 11775™ | ATCC® | | | | |
| | *E. coli* Biotype1 | ATCC® 51813™ | ATCC® | | | | |
| | *E. coli* | ATCC® 25922™ | ATCC® | | | | |
| | *E.coli* | ATCC® 8739™ | ATCC® | | | | |
| **(Pathogenic)** | *E. coli* O157:H7 (toxin negative) | ATCC ® 43888™ | ATCC® | | | | |
| | *E. coli* O157:H7 (toxin positive) VT1 or VT2? | no longer available from ATCC® or by reference strain number | BEI Resources | | | | former ATCC® 43894 |

| Genus | Species/Serotype | Strain | Source | Additional Recommended Species/Strains: Species | Strain/Serotype | Source | Exclusivity Species |
|---|---|---|---|---|---|---|---|
| | *E. coli* O157:H7 (EDL 931)<br>Non O157:H7 EHEC strains<br>Shigella STX gene | no longer available from ATCC® or by reference strain number | | former ATCC® 35150 | | | |
| ***Giardia*** | *G. lambia*<br>*G. muris* | Human Isolate H3 P101 | Waterborne, Inc | | | | |
| ***Helicobacter pylori*** | | | | | | | |
| ***Listeria*** | *L. monocytogenes* ½ a | ATCC® 51772™ | ATCC® | | | | |
| | *L. monocytogenes* ½ a | ATCC® 51775™ | ATCC® | | | | |
| | *L. monocytogenes* ½ b | ATCC® 51780™ | ATCC® | | | | |
| | *L. monocytogenes* ½ c | ATCC® 51779™ | ATCC® | | | | |
| | *L. monocytogenes* 3a | ATCC® 51782™ | ATCC® | | | | |
| | *L. monocytogenes* 4b | Scott A | | | | | |
| | *L. monocytogenes* 4b | ATCC® 19115™ | ATCC® | | | | |
| | *L. monocytogenes* 4d | ATCC® 19117™ | ATCC® | | | | |
| | *L. monocytogenes* | ATCC® 19111™ | ATCC® | | | | |
| | *L. monocytogenes* (non-hemolytic) | ATCC® 15313™ | ATCC® | | | | |
| | *L. ivanovii* 5 | ATCC® 19119™ | ATCC® | | | | |
| | *L. innocua* 6a | ATCC® 33090™ | ATCC® | | | | |
| | *L. welshimeri* 6b | ATCC® 35897™ | ATCC® | | | | |
| | *L. seeligeri* | ATCC® 35967™ | ATCC® | | | | |
| | *L. grayi* | ATCC® 25400™ | ATCC® | | | | |
| | *L. grayi* | ATCC® 25401™ | ATCC® | | | | |
| ***Mycobacterium paratuberculosis*** | | | BEI Resources | Controllled Access Strain | | | |
| ***Norovirus*** | Norwalk – Group I<br>Snow Mountain Agent – Group II | Any isolates from feces from naturally infected individuals.<br>Currently, no isolate or strain is maintained in a laboratory. | | | | | |
| ***Salmonella*** | *Salmonella Typhi* | | | Representatives from somatic groups B-I | | | |
| | *Salmonella Paratyphi A* | | | Representatives from “further groups” | | | |
| | *Salmonella Paratyphi B* | | | | | | |
| | *Salmonella Paratyphi C* | | | | | | |
| | *Salmonella Sendai* | | | | | | |
| | *Salmonella Typhimurium* | ATCC® 13311™ | ATCC® | | | | |
| | *Salmonella Enteritidis* | ATCC® 13076™ | ATCC® | | | | |
| | *Salmonella choleraesuis* | ATCC® 10708™ | ATCC® | H2S negative strain (>48 hours) | | | |
| | 1. A minimum of 35 of the top 50 serotypes isolated in the United States from 1968 to 1998 (see attached table An Atlas of Salmonella in the United States, published by the CDC in 2000.<br><br>2. Representatives from somatic groups B-I (serotypes should be evenly distributed across the groups). A minimum of 30 serotypes seems appropriate.<br><br>Representatives from "further groups". These further groups should also include representative serotypes from the *S. enterica*, subspecies *salamae* (II), *arizonae* (IIIa), *diarizonae* (IIIb), *houtenae* (IV), and *indica* (VI). *S. bongori* should also be included if possible. | | | | | | |
| ***Shigella*** | *S. boydii* serotype 2 | NCTC 12985 | BEI Resources | | | | |
| | *S. dysenteriae* | NCTC 4837 | BEI Resources | | | | |
| | *S. flexneri* serotype 2a | 24570 | BEI Resources | | | | |
| | *S. sonnei* WRAIR I virulent | | BEI Resources | | | | |
| ***Staphylococcus*** | *S. aureus* | ATCC® 25923™ | ATCC® | | | | |
| | *S. aureus* | ATCC® 6538™ | ATCC® | | | | |
| | *S. epidermidis* | ATCC® 12228™ | ATCC® | | | | |
| **Staphylococcal Enterotoxins** | A | FRI-722 (FDA) | Toxin Technology | | | | |
| | B | 110-270USAMRIID | Toxin Technology | | | | |
| | C1 | FRI 137 (FDA) | Toxin Technology | | | | |
| | C2 | FRI 361 (FDA) | Toxin Technology | | | | |
| | C3 | FRI 1230 (FDA) | Toxin Technology | | | | |
| | D | FRI 1151 (FDA) | Toxin Technology | | | | |

| Genus | Species/Serotype | Strain | Source | Additional Recommended Species/Strains | | | Exclusivity Species |
|---|---|---|---|---|---|---|---|
| | | | | Species | Strain/Serotype | Source | |
| | E | FRI 326 (FDA) | Toxin Technology | | | | |
| | F (Toxic shock toxin) | RFDA 485 (A,B,D) (FDA) | Toxin Technology | | | | |
| | Non-toxigenic strain | FDA D87 | FDA | | | | |
| | | FDA D184 | FDA | | | | |
| | | *S. interimidis* | FDA | | | | |
| ***Vibrio*** | *V. cholerae* | 7th pandemic O1 strain | | | | | |
| | *V. cholerae* | O139 (7th pandemic strain that has mutated to have a capsule) | | | | | |
| | *V. cholerae* | non-O1/non-O139/non-toxigenic strain | | | | | |
| | *V. cholerae* | Gulf Coast O1 toxigenic strain (non-epidemic) | | | | | |
| | *V. cholerae* | classical *cholerae* strain | | | | | |
| | *V. cholerae* | O141 toxigenic strain that has recently emerged | | | | | |
| | *V. parahaemolyticus* O3:K6 | | | | | | |
| | *V. paraphaemolyticus* O4:K12 | tdh+/trh+ | | | | | |
| | *V. parahaemolyticus* | non-pathogenic strain (tdh-/trh-) | | | | | |
| | *V. parahaemolyticus* | clinical strain (tdh-/trh+) | | | | | |
| | *V. vulnificus*, biotype 1 (rRNA type B) | | | | | | |
| | *V. vulnificus*, biotype 2 (rRNA type A) | | | | | | |
| | *V. vulnificus* biotype 3 (rRNA type A/B) | | | | | | |
| ***Yersinia*** | *Y. enterocolitica* | ATCC® 27729™ | ATCC® | | | | |

* This table presents a reasonably comprehensive inclusivity list of reference pathogenic and related microbiological species and toxins that must be included in any validations study.
Other strains used for inclusivity or exclusivity must be characterized using nationally or internationally accepted reference methods.
The pre-collaborative and/or collaborative study protocols submitted by method devlopers or study directors for review, will be evaluted by an expxert panel to ensure that the strains selected for exclusivity include a suffcient number of appropriate "nearest-neighbours" to also challenge the inclusivity of the method.

# AOAC INTERNATIONAL
# Presidential Task Force on
# Best Practices for Microbiological Methodology

US FDA Contract #223-01-2464, Modification #12

Executive Summary
Sampling Working Group (SAWG)

## INTRODUCTION

As with any type of testing, an understanding of the sampling and measurement procedures for microbiological methods is necessary for gaining confidence that the obtained results "represent" the intended population or fulfill a study's purpose. The confidence of results can be undermined if care is not taken to control and minimize the variation of observed results due to sampling, sample preparation and measurement. To address this concern, the AOAC has asked the Sampling Working Group (SAWG) of the BPMM Task Force to identify and address the components of sampling and measurement variation – specifically, the factors that contribute to and must be controlled or understood in order to gain an understanding of results and thereby enhance their proper use. This would include identifying components across the whole process of sampling and measurement, including the method of measurement and the laboratory performance. Once these components of variation are understood, proper application of the method can be designed.

There has been significant work done by the International Commission on Microbiological Specifications for Foods (see ICMSF, 2002) to develop and provide guidance on the use of microbiological sampling plans for foods. The statistics underlying these sampling plans, however, are not well understood (Dahms, 2004). The components of variation, referred to above, were not considered in determining the operating characteristics of the plans; instead, rather idealized assumptions were made.

In view of these issues, the objective under consideration by the SAWG of the AOAC International Best Practices for Microbiological Methods (BPMM) Task Force is: (Contract question #3) What are reasonable performance standards (criteria) when microbiological methods are to be used for: 1)Attribute testing, 2) Variables testing, and 3) Process control testing.

## METHODOLOGY

The AOAC objective set forth is broad, and therefore the SAWG narrowed its scope to identify important areas that could lead to further investigation. Certain assumptions were made. One primary underlying assumption is that a statistically representative sample can be obtained and that if composite samples are to be used, then these composites will be "representative" from a unit or amalgamation of multiple units that they are to characterize. An indication that a set of samples is representative of the lot is that the variation between samples is less than the mean. It also follows from these

assumptions that outright "errors" due to mislabeling of samples, cross-contamination, incorrect readings from a machine, etc. would not be addressed. These possibilities are important to consider, and should be part of any well-designed laboratory standard operating procedure (SOP), but are beyond the scope of the SAWG.

The issues to be addressed by the group do not depend, per se, on whether the type of test being considered is an attribute or variable test. In other words, the recommendations presented below are being made with regard to qualitative tests as well quantitative tests that are more familiar to AOAC. In lieu of the above discussion, the SAWG considered the following tasks:

1) Identify and address performance components of variation relative to intra-laboratory, and inter-laboratory performance.
2) Identify and address components of variation of measurement error associated with the method within the laboratory.
3) Identify process control statistics and recommend a set of performance standards for statistical process control using microbiological measurements.

**I. Components of variation relative to intra-laboratory and inter-laboratory performance.**

The SAWG believes that to determine method performance, controlled inter-laboratory studies are needed. The recommendations are closely aligned with AOAC recommendations for collaborative studies of chemical analytical methods. The recommended performance standards are:

**1**. Ruggedness tests should be performed that attest to the robustness of the analytical procedure under expected normal operating procedures. Ideally 5-7 critical steps of the procedure should be identified, and the nominal, upper and lower specs for each step evaluated.

**2**. Microbial test validation should include estimates of test sensitivity, specificity, and accuracy.

**3**. A Collaborative study consisting of 5-10 laboratories should be conducted to determine reproducibility and repeatability standard deviation measures that cover the range of levels expected to be encountered and that are of regulatory interest. If this is not possible, then at least an intra-laboratory study, using more than one analyst, separated from each other, should be conducted. From these results, formulas predicting the standard deviations as a function of level should be estimated.

**4**. For QA purposes, laboratories should establish a range of acceptable results for individual samples based on confidence intervals using the repeatability standard deviations. Also, laboratories should establish process control

procedures, and use statistical process control methods for tracking performance over time.

**5.** When reporting results, the range given as the 95 percent confidence interval on the measurement should be stated.

**II. Identify components of variability within the lab**.

The SAWG focused on examples of method protocols to examine where the measurement error variation can occur. Enclosure A presents a detailed account of our identification of major sources of sampling variation that occur within the laboratory. We are recommending that laboratories develop a protocol for maintaining process control at critical points of the analytical procedures. The recommended performance standards for laboratories are:

**1**. Establish a process for listing sources that contribute to the variability of results in the laboratory (this should be developed).

**2**. Perform intra-lab repeatability studies to determine statistical distribution of results associated with the sources of variability.

**3.** Establish statistical process control procedures (based on split or check samples) within the laboratory to monitor performance.

**4.** For methods that involve confirmation of particular types of organisms where interfering organisms are expected, conduct a study to determine the proportions of targeted and interfering organisms in samples. This will help determine how many confirmations are needed to minimize false negative outcomes.

**III. Statistical Process Control (SPC)**.

SPC is a very broad area which SAWG believes is not well known to the scientific community. Consequently, for this task, the SAWG presents a general introductory discussion (Enclosure B) together with numerous examples. The suggested performance standards are general principles that should be followed, representing normative practice. These are:

**1**. Charts of plots of the output data are necessary for gaining the full benefit of doing SPC.

**2**. When the process is under control, the results plotted on a statistical process control chart should be normal or nearly normally distributed. In cases where this is not true and an alternative known distribution cannot be determined, transformations of the data should be considered.

**3**. During some "initial" period of time, when it is presumed the process is operating in a relatively stable manner – or is in control, the distribution of the measurements should be estimated and rules for evaluating the process should be formulated. Use of about 20-30 results (samples) or more for computing means and standard deviations or other summary statistics needed for distribution estimation is a desirable goal. However, this stipulation can be relaxed and thus should not hinder or limit the use of control charts if resources do not permit, in a timely fashion, analyzing this number of samples.

**4**. Rules for evaluating process control should be set with aids assessing the two types of errors: Type I ($\alpha$-probability), declaring the process out of control when it is not, and Type II ($\beta$- probability), not declaring a process out of control when it is. Typically there are two measures that are used for assessing these errors: 1) the probabilities of the two types of errors at a given time and 2) the average run length (ARL) or expected number of samples before an out of control signal (one of the rules being not met) is seen. When developing rules, the $\alpha$-probability (Type I error) should be kept low, for example, below 1%, or the ARL should exceed 100 (corresponding to less than 1% $\alpha$ - error).

**5**. When a process is thought to be "in control," the limits for assessing individual results are set at a distance from the mean (target), expressed as standard deviation units from the mean or process target value. The recommended and default distance is 3 standard deviations. Additionally, characteristics related to food safety may be targeted more than three standard deviations above or below critical limits, however statistical process control limits should still be placed 3 standard deviations from the target value.

**6**. There are numerous run/ trend rules that can be used, such as runs test, moving averages and CUSUMS, for detecting shifts in the process mean; and rules for detecting shifts in the process variation or other auto-correlated patterns that could be due to a systematic source of variation. The use of any of these may depend upon particular expected conditions when the process is out of control.

**7**. Specification Limits are not Statistical Process Control limits. Specifications are either customer, engineering, or regulatory related. Specification limits should not be placed on a control chart insofar as these might be considered as process goals thus influencing the efficacy of SPC procedures for ensuring a controlled process, and thereby undermining the safety of the product.

For more details concerning the specific performance criteria, please review the referenced Enclosure materials.

# Introduction - Sampling Working Group

As mentioned in the executive summary on sampling, an understanding of the sampling and measurement procedures is necessary for obtaining confidence that the obtained results "represent" the intended population or fulfill a study's purpose. The confidence of results can be undermined if care is not taken to control and minimize the variation of observed results due to sampling and measurement. To address this concern, the information below is presented as a foundation for and linkage to the two documents on measurement error (Enclosure A) and statistical process control sampling (Enclosure B).

## Sampling and Measurement

An important reason, and the one that is of interest for this Committee, for analyzing samples in the first place is to characterize some aspect of the distribution of the "true level", x, or most generally, to determine the distribution of x, within some well-defined population of product that the analyzed samples are "representing." The values of x could refer to levels or densities of some measurand or could refer to whether or not a pathogen is present in a sampled material. The results are a collection, $\{y_j, j = 1, \ldots n\}$ where n is the number of samples (here assumed randomly drawn for some population, with equal probability of selection). Thus values of y refer to the measured result, either a measurement of level or density of some measurand, or whether or not the pathogen was found. The value of y thus represents the "known" evidence, from which an inference is made regarding the possible values of x. In the interferential process there is always uncertainty associated with any conclusion or characterization made about possible value of x.

Mathematically this uncertainty can be represented by a "likelihood" function. This function can be derived in stages. First consider the probabilities of possible values of y for hypothetical values of x, $g(y|x)$. This is a function of the true value of x. However, x is not known, but rather y is known. The values of x, being unknown, are (next) assumed to occur with some probability density which can be labeled, f(x). With this supposition, the (full) probability relationship between y and x can be written down mathematically. To distinguish the case of y being known and x being unknown (from the case of x being known) the phrase "probability of y" is not used, but rather the phrase used is the "likelihood of y." More specifically, if the density of the distribution of x is f(x), then the likelihood (L) of obtaining a value of y can be expressed as a joint probability integral equation:

$$L(y) = \int g(y|x)f(x)dx \qquad (1)$$

This equation includes results reported as non-detects, ND, as a possible value. That is,

$$L(ND) = \int g(ND|x)f(x)dx \qquad (2)$$

where g(ND|x) is the probability of ND (of getting a non-detect, or a false negative) given a true value of x in the sample. An estimate of f(x) can be derived from the above integral equation, assuming g(y|x) is known. If f(x) is of known (or assumed to be a specific) mathematical form, parameterized with parameter vector, θ, of (often) unknown values, then using maximum likelihood (MLE) estimation or method of moments (MOM), estimates of values of θ can be obtained.

Often, the forms of f(x) and even g(y|x) will not be known, but their first two moments (mean and variance) can be estimated and be considered sufficient for many purposes. An example of this is with statistical process control (SPC), discussed in Enclosure B (SPC document), where SPC procedures depend upon specifying the mean and variance of the process. If it is assumed that the relationship between the expected value and variance of y given x and x is known, then the mean and variance of the distribution of x can be obtained. The relationship for the means and the variances are:

$$E(y) = E_f(E(y|x)) \quad (4)$$

$$Var(y) = var_f(E(y|x)) + E_f(var(y|x)) \quad (5)$$

where $E_f$ and $var_f$ refer to the expected value and variance of the distribution with density function f, and E and var, without subscripts, refer to expected value and variance of distribution g. The terms on the left are determined directly from the collection of $\{y_j, j = 1, \ldots n\}$of sample results; the terms E(y|x) and var(y|x) are assumed known functions of x, so that the above equations can be used to solve for $E_f(x)$ and $var_f(x)$.

A simple example is to assume that var(y|x) is some linear function of x: var(x) = ax + b, where a and b are constants. For some methods, such as methods of measuring densities of chemical residues, the coefficient of variability (CV) is assumed to be equal to 100(a + b/x), when x is the true level of some analyte, so that the variance, var(y|x), would be $(ax+b)^2$. Assuming that E(y|x) = x – that is, the method is unbiased - the above equations become:

$$E(y) = E_f(x) \quad (4a)$$

$$Var(y) = var_f(x) + E_f(ax+b) \text{ , or in the second case,} \quad (5a)$$

$$Var(y) =[1 + a^2]\, var_f(x) + [aE_f(x) + b]^2 \quad (5b)$$

If E(y) and Var(y) can be estimated from *a priori* information, for example, from inter- or intra-laboratory studies, then $E_f(x)$ and $var_f(x)$ can be estimated by solving the above equations.

Often there is a need for imputation or assigning a value of y when the imputed value is a non-detect value (ND). A standard procedure for imputation is to impute ½ the limit of detection (LOD) (EPA, 2000), and then compute the average and standard deviation using the imputed values for ND. A justification of this imputation procedure could be based on the

"principle of indifference,"[1] which here would invoke an assumption that the values of y that could have been measured would be uniformly distributed between 0 and L (where L = LOD). In other words, if it is thought that y represents an estimate of x on a sample, then the "best" estimate of x given that y is below the LOD = L, by the principle of indifference, is L/2. This is a confusing assumption and its very premise leads to contradictions, as is well known; for example, by the same "principle of indifference" applied to the square root of the true level, $x^{1/2}$, the imputed estimate would be $L^{1/2}/2$, so that for x, the imputed estimate would be the square of this value, specifically, L/4. Ideally if the true distribution were known (or assumed) then values for ND results could be derived using statistical estimation procedures. Based on assumptions for the distribution, procedures for imputation of results reported below the LOD have been proposed (Cohen, 1959; Persson and Rootzen, 1977; Singh and Nocerino, 2002). In any case, at least with chemical measurements, the LOD/2 imputation is commonly used (EPA, 2000) and would permit the above calculations to proceed.

## Importance for sampling

It might be (as is often the case) that the percentage of the variance component (of the total variance) due to measurement is small relative to the variance component due to sampling variation. However, even in this situation, the variance of individual results can be of such magnitude to affect significantly the confidence that is associated with individual results. For a simple example, assume that the distribution of APC counts is lognormal, and that the mean of the $\log_{10}$ of the sample values, y, is 3 and the sample standard deviation is 1. Since we are assuming that the distribution of the $\log_{10}(y)$ is normal, a 95% probability interval would be approximately 1 ($\log_{10}$) to 5 ($\log_{10}$). Consequently, if there were a specification that "permitted" no more than 4.5 $\log_{10}$ on a sample[2], then based on the normal distribution for the logarithm of the APC counts, assuming a mean value of 3 $\log_{10}$ and a standard deviation of 1, there is a probability of 6.7% that a sample value would exceed 4.5 $\log_{10}$ (the z-score corresponding to the limit, 4.5 $\log_{10}$ is z =( 4.5-3)/1 = 1.5, which has associated cumulative probability of 93.3%, so that probability of being greater than 4.5 is 6.7%).

For simplicity here, assume that the distribution of $\log_{10}(y)$, given a sample with a true level of x, such that the expected value of $\log_{10}(y)$ is $\log_{10}(x)$, and the standard deviation is 0.3 $\log_{10}$, independent of the value of x. From Equation 5a, $1^2 = \mathrm{Var}(\log_{10}(y)) = \mathrm{var}_f(\log_{10}(x)) + 0.3^2$, so that the population variance of $\log_{10}(x)$ is $1- 0.3^2 = 0.91$; and the standard deviation of $\log_{10}(x)$ is $(0.91)^{1/2} = 0.954$. Hence the 95% probability interval, symmetric about the mean of the $\log_{10}$ of the true levels for the population, is 1.13 to 4.87 $\log_{10}$ and there would be a 6.3% probability that a sample value would exceed 4.5 $\log_{10}$. The difference between the two

[1] Also referred to as the "principle of insufficient reason" developed in the 19th century, and later renamed 'priciple of indifference' by the economist John Maynard Keynes (http://en.wikipedia.org). It basically stipulates that lacking any other information one can assume equal probablites for a set of events. Where the events refer to values of continuous variables the principle leads to ambiguity as described within the text.

[2] In some situations, a specification would refer to the true level in a sample so that it would be necessary to know the measurement error to determine compliance. Some adjustment might be made then to account for measurement error.

intervals is not large (1 to 5 versus 1.13 to 4.87 $\log_{10}$). Now, if the measurement standard deviation were reduced by a factor of 2, to 0.15 $\log_{10}$, then the probability of a single result obtained on a randomly drawn sample being greater than 4.5 $\log_{10}$ would be about 6.0%, reduced from 6.3%, hardly a change at all.

On the other hand, the impression of the effect of reducing the standard deviation of the measurement error could be different when considering its impact on inferring a value for sample using single measurements. For a single measurement, a standard deviation of 0.3 $\log_{10}$ would imply that, the 95% confidence interval associated with that true sample value of $\log_{10}(x)$, would be $\log_{10}(y) - 0.588$, $\log_{10}(y) + 0.588$, a range of 1.176 $\log_{10}$, or a factor of about 15. If the true value for a sample was 4 $\log_{10}$, which is well below the specified limit amount of 4.5 $\log_{10}$, there would be about a 5% chance that a measured value would exceed 4.5, assuming a standard deviation of 0.3 $\log_{10}$ for the measured result. If the standard deviation were reduced by a factor of 2, then the range of the 95% confidence interval associated with a measured value would be $\log_{10}(y) - 0.294$, $\log_{10}(y) + 0.294$, a range of 0.588 $\log_{10}$, or a factor of about 3.9, a seemingly substantial reduction. The probability of a result being greater than 4.5 $\log_{10}$ given a true $\log_{10}$ value of 4 would be 0.043%, virtually zero, compared to the 5% when the measurement standard deviation is 0.3 $\log_{10}$. This could be considered a significant change.

Thus, overall, when considering the effect on sampling populations, reducing the measurement standard deviation from 0.30 to 0.15 does not amount to a significant change in the operating characteristic (OC) curve (which provides the probability of acceptable results given assumed true conditions (Juran, JM, 1951) when the results of the measurements are being used for assessing a distribution of levels within some population - in our example, the probability of failing was reduced from 6.3% to 6.0%, about a 5% reduction of the probability of obtaining failed samples. The effort needed to reduce the standard deviation by a factor of 2 would be at least 4 samples per analysis, and perhaps more, as discussed below. As shown by way of this example, it may not be worth the extra time and effort to increase the number of analyses per sample. However, when inferring a true value for a specific sample, perhaps in a legal setting, the reduction of the standard error of the mean might be significant, as illustrated by the above example.

In determining how many samples would be needed to reduce the standard error of the mean (compared to the standard deviation of a single result), the magnitude of the variance components associated with the sampling and measurements would need to be known. For example, very simply, the standard deviation may include significant day-to-day effects. In other words, samples analyzed on the same day would not be independent results, but rather would be correlated within the population of possible results that would be obtained for the sample if it were analyzed on different days with different reagents and so forth. This notion is expressed by identifying a parameter, $\delta$, called the intra-day correlation, which is the proportion of the between-day variance to the total variance - that is, the sum of the between-and within-day variance. For n samples analyzed per day for m days, the variance of the mean would be

$$\sigma_m^2 = \sigma_0^2\delta/m + \sigma_0^2(1-\delta)/(mn) \qquad (6)$$

where the first term on the right side represents the contribution of the between-day variance component (sampling for m days), and the second term represents the contribution of the within-day variance component (for mn samples). For example, for a value of $\delta$ of 0.3, the mean of 56 samples, analyzed over 8 days, 7 samples per day has the same variance as that of the mean of 98 samples analyzed over 7 days, 14 samples per day. An intra-day correlation of 0.3 is large, but not unbelievable, particularly for microbiological measurements wherein "causes" of contamination or high levels of organisms could vary day- to-day by substantial amounts.

Assume that a result needs to be obtained daily for some quality assurance or control purpose and thus results are analyzed in one day, so that m = 1. A question might be: how many samples are needed (in one day) in order that the standard error (of the mean) is a fraction r of the standard deviation, $\sigma_0$, of a single result? From Equation 6, assuming $\delta < r^2$, the number of samples needed would be:

$$n = \frac{1-\delta}{r^2-\delta} \qquad (7)$$

Thus, for example, if $\delta = 0.1$ and r = ½, 6 samples per day would be needed to have a variance of the mean be ½ the variance of a single result.

## Summary

***For microbiological measurements, true levels of the measurand are often highly variable over time, so that in general, given resources for a fixed number of samples, more samples over time with less samples per day, and more days of sampling is preferable if the purpose of sampling is to examine trends or get a good profile of the distribution of the measurand over time. However, if decisions are to be made on sample results for a given day, to ensure that product is safe then more samples per day might be needed.***

## Composite sampling

To minimize costs, composite sampling can be considered, when k samples (for example in one day) are divided into m composites of n samples (so that k =mn). The variance of the mean of the results obtained from the m composite samples would be:

$$\sigma_m^2 = (\sigma_0^2\delta + \sigma_a^2)/m + \sigma_0^2(1-\delta)/k = \sigma_0^2(1+\delta(n-1))/k + \sigma_a^2/m \qquad (8)$$

where, $\delta$ now refers to the intra-composite correlation, $\sigma_0^2$ is the between sample variance, ignoring measurement variance, and $\sigma_a^2$ is the pure analytical measurement (referred to as repeatability) variance. From Equation 8, it is seen that it is desirable that $\delta$ be small, which would be the case if it could be expected that true differences of levels between composite samples be negligible. Stratifying the population being sampled or selecting systematically from every $m^{th}$ sample to form composite samples (for example, from 12 samples, selecting the

first, fourth, seventh and 10th as the first composite, and so forth, for 3 composite samples consisting of 4 samples each) would effectively minimize the value of δ. Assuming δ is small and can be ignored in Equation 8, the variance of the mean would depend upon the relative magnitude of $\sigma_0^2$ and $\sigma_a^2$; if m (the number of composites) is small, then, even with n being large, the variance of the mean could be large since the term $\sigma_a^2$ /m could be large.

However, often microbiological analyses are not able to handle large samples, and thus there may be a limit to the size of the composite samples. The limiting factors regarding the size of composite samples are the container size required for a (for example) 1/10 dilution, the ability to homogenize large samples and incubator space. Some laboratories may be equipped to handle large size samples, using walk-in incubators and such; however, most laboratories do not have such equipment.

Consideration also needs to be given to the sensitivity of the analytical procedure as a function of the sample size. In other words, analyzing composite samples might introduce a bias if the sensitivity of recovery were affected by compositing. These considerations might lead to limiting the number of samples, n, within a composite sample. This in turn might make less innocuous the assumption of a small δ.

Suppose it is decided that M grams (or ml if liquid samples are being considered) is the size of the composite sample. That is, the number of individual samples, n, in a composite sample, times the weight, w, (or liquid volume) of each individual sample, nw, should be equal to M. The total number of samples, k = mM/w. Equation 8 for the standard error of the mean of m composite sample results becomes:

$$\sigma_m^2 = \frac{w\sigma_{0w}^2}{mM}(1+\delta(n-1)) + \frac{\sigma_a^2}{m} \qquad (9)$$

where the symbol $\sigma_{0w}$ refers to the between-sample variance for samples of weight (or volume) w. As w decreases (and thus increasing the number of samples, k) it would be expected that $\sigma_{0w}$ would increase. The relationship between the two quantities: w and $\sigma_{0w}$ would need to be explored in order to design an optimal composite sampling plan.

While composite sampling can lead to decrease of costs of sampling, it should be pointed out that the results obtained from composite sampling can mask information concerning the distribution of the levels of the measurand within the population being sampled. Information of the distribution of levels might be important for evaluating process control and for risk assessments that are primarily concerned with estimating risks typically associated with (occasional) high levels of some pathogen in food. Hence, for designing sampling plans, an understanding of how the results might be used is needed.

Designing sampling plans thus requires knowledge of variance components related to measurement and sampling variability associated with the sampling unit. In the SPC document (Enclosure B), the discussion does not address the effects of measurement error explicitly;

rather the document concentrates on the issues related directly to SPC, and estimated variances would include the contribution due to measurement. The total variability (due to sampling and analytical measurement) should be known or estimated in order to rationally design sampling plans– regarding the number of samples, composites, and repeat analyses that might be needed - and for constructing realistic OC curves. Information concerning specifics of this analysis can be found by reading Enclosures A, as well as reviewing the references in each case.

## References

Cohen Jr, A. C. 1959. Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics,* 1(3):217-237.

Juran, JM. 1951. Quality Control Handbook, Third edition, McGraw-Hill Book Company, NY.

Persson, T. and Rootzén, H., 1977. Simple and highly efficient estimators for a type I censored normal sample. *Biometrika,* 64:123-128.

Singh, A. and Nocerino, J., 2002. Robust estimation of mean and variance using environmental data sets with below detection limit of observations. *Chemometrics and Intelligent laboratory Systems*. 60:69-86.

United States Environmental Protection Agency (EPA). 2000. Assigning values of non-detected/non-quantified pesticide residues in human health exposure assessments. Office of the Pesticide Programs, item 6047.

# Enclosure A – Measurement Error

## Methodology

As mentioned in the executive summary, the AOAC questions posed are broad and the SAWG therefore narrowed its scope to identify important features that could lead to further development, as needed. Certain assumptions were made. One primary underlying assumption is that a "representative" sample can be obtained. Thus SAWG did not address outright "errors" due to mislabeling of samples, cross-contamination, incorrect readings from a machine, etc.

To address measurement error, the SAWG examined a publication that thoroughly attempts to quantify the measurement variation of basic microbiological methods (Niemela 2002). This document focuses on the measurement error associated with counting colonies and/or other discrete entities. The SAWG report identifies important factors that contribute to laboratory measurement variability and recommends that these factors be addressed with all methods for the purpose of controlling them and quantifying them, if possible.

The SAWG focused on examples of method protocols to examine where the measurement error variation can occur. Examination of these examples led to the identification of major sources of variation that the SAWG will consider:

A. Dilution
B. Recovery
C. Counting
D. Organism confirmation
E. Organism variability
F. Overall statistical considerations

A short description of the each of the sources of variation is given below, followed by section F that provides an example of how the operating characteristics for a plan could be constructed taking into consideration these sources of variation.

Both the enumeration of microbiological counts and the identification of microbes' genus and species involve a number of steps. Some of the steps include: sample collection, sample preparation including dilution(s), and (in some cases) maceration or mixing (or both). In determination of genus and species of pathogenic organisms, there is often an incubation step in a selective media prior to implementing one of the methods used for detection. All of these steps may have some error or variation associated with them.

A. Dilution error

Dilution errors are those errors associated with the sample preparation from the time a sample is gathered until the time the organisms are either counted or identified. Without proper training, the opportunity for error could be substantial, and the impact may vary from small to great. The first dilution error is initial sample size. If a sample of a particular size is to be placed into a fixed amount of diluent, the size of the initial sample (either larger or smaller than nominal) would cause under-dilution or over-dilution error, respectively. Similarly, if a surface area constituted the sample, then sampling an area larger than the specified area would result in an under-dilution, and an area smaller than the specified area (or incomplete swabbing of the specified area) will cause over-dilution.

With regard to setting up blanks for dilutions**,** some additional errors may occur. First, volumetric errors related to the use of graduated cylinders and pipettes could occur. Second, dilution blanks may be prepared volumetrically correct, but then autoclaved causing volume reduction. Errors may also be associated with incorrectly reading the meniscus in graduated cylinders and pipettes. Pipettes may differ in the way volumes are correctly measured. Plastic and glass pipettes' and cylinders' menisci are not read the same way, and may not have the same reliability.

Pipetting errors, of course, can occur for all dilutions after the original sample is placed into the original blank and the subsequent dilutions are made. Also, when micro-volumes of samples are pipetted into test containers, errors can occur due to pipetting technique. Additional errors may occur due to debris restricting the filling or emptying of the pipette or pipette tip, thus causing a non-representative sample to be placed in the testing container.

These errors do not include the obvious errors associated with spillage or leakage, but if unobserved or uncorrected, these factors could contribute greatly to the error associated with dilution.

B. Recovery error

The recovery rate for a microbiological method is the proportion of target cells (or spores) in the test sample that is presented to the detection method. With rare exception, recovery entails multiplication of the target cells to the high numbers required by the detection portion of the test. If multiplication does not occur or is impeded, the microbial count is underestimated. Factors confounding multiplication are as follows:

1. Incubation conditions: Test methods specify the time, temperature, and atmosphere of incubation. Ranges are generally provided for time and temperature. In many cases these ranges may be very broad and may compound in

methods that have multiple steps, e.g. liquid culture incubation for time ±2 hr, temp ± 1C, then subculture incubation for time ±2 hr, temp ± 1C, and finally agar plate culture for time ±2 hr, temp ± 1C. Validation studies rarely validate the extremes, as doing so is costly.

2. Media: Biological media is generally not calibrated from lot-to-lot or supplier-to-supplier, with perhaps the exception being Standard Methods Agar which is calibrated lot-to-lot per supplier. A great deal of variation can and does occur related to the components of basic culture media. Selective media adds another level of variation, since selective agents may vary in toxicity depending upon lot, preparation, and storage.

3. Product Matrix: Organisms may be in or on a product, in clumps or as single cells. Accurate enumeration requires full release from the matrix for repeatable enumeration. Some foods may be inhibitory to growth—spices are common examples. In addition, foods may contain competitive flora which may inhibit growth or outgrow the target population.

4. Target Flora: Recovery rates may vary among genera, species, or even subspecies and strains. The target flora may be injured and thus variation in recovery may increase.

C. Counting error

It is often believed (assumed) that the distribution of the results of a count follows the Poisson distribution. This is based on assuming that the cells are distributed uniformly so that, per unit of product, there is a single expected level, r. From this assumption, it can be shown (Jaynes, 2003: Probability Theory: The logic of Science Cambridge University Press) that the distribution of the number of cells in any volume, v, of product is a Poisson distribution with expected value rv, and standard deviation $(rv)^{1/2}$. However, cells, larvae or other types of microbiological contaminates are usually not distributed in nature as a Poisson, but rather are distributed in clumps, or colonies, either because there are factors that would attract microbes to cluster or because of the cell division process creating a tendency for colonies to form (e g, Campylobacter). Because the assumption of uniformity cannot be assumed, in order for the Poisson distribution to hold, it is necessary to homogenize the sample. Variation beyond that expected from the Poisson thus can be introduced when the sample is not homogeneous.

Additional variation is also introduced due to the non–exactness of the counting of colonies of a specified species. The counting may differ appreciably between persons for a given sample on a given medium. Familiarization thus with the counting procedures is an important requirement for analysts.

D. Organism confirmation error - selection and testing of colonies

During many analytical procedures colonies need to be selected for further testing. Methods usually specify that a certain number of colonies, or a certain proportion of the colonies meeting the description of the target species are selected for further testing. The sampling errors involved in this procedure depend upon the differential power of the primary isolation medium and upon the ratio of target species to non-target species that meet the description of colonies to be further selected.

**D.1. Differential power of primary isolation medium.** The differential power of the medium is the ability of the medium to cause target species to appear sufficiently different from non-target species, to facilitate the efficient selection of the target species for confirmatory testing (if required).

The differential power of the medium may be graded as follows:

Absolute: in which every colony on the plate is counted and no further testing is required. Examples: Standard Plate Count, Aerobic Plate Count.

Highly differential- in which we can be almost certain that colonies that meet the description belong to the target species, other colonies are clearly differentiated. Example: B. cereus on MYP or PREYPA or PEMBA.

Moderately differential- in which colonies meeting the description of typical strains may belong to the target species. Some non-target species may fail to be differentiated from the typical colonies of the target species and/or some atypical colonies may belong to the target group. Examples include: Salmonella sp. on BSA, Coagulase positive staphylococci on Baird Parker Agar, Listeria species on MOX.

Poorly differential- in which colonies of the target species and some (related) non-target species are not differentiated. Example: L. monocytogenes on Oxford Agar or PALCAM.

The magnitude of error that may be associated with this factor increases as the differential ability of the medium decreases.

**D.2. Ratio of target to non-target colonies on the primary isolation medium.** The ratio of target to non-target colonies affects the ability of the operator to select the most likely colonies for further testing. This factor operates in two different ways that may interact: the selectivity of the medium and/or the differential power of the medium.

The selectivity of the medium refers to the ability of the medium to suppress non-target species. If a medium is highly selective, it is more likely that a colony on the

agar will belong to the target species. In extreme cases, poorly selective media may allow non-target species to overgrow target species to the extent that the target species cannot be detected (for example, Citrobacter overgrowing *Salmonella* on XLD agar). In many cases, this reason is not a large problem if the medium is sufficiently differential.

The differential ability of the medium is discussed above. If poorly differential agars are used, then selecting colonies of the target species will depend entirely on the ratio of target and similarly-appearing non-target species. For example, both L. monocytogenes and L. innocua will have identical colonies on Oxford Agar, but only one of these species is the target. In qualitative tests: if the method requires a number of colonies to be selected, and only one (1) colony needs to be confirmed as positive for the target species to be reported as detected, then it will be possible to calculate the likelihood of selecting a colony of the target species depending upon the number of colonies of each of the target species and the similarly appearing non-target species. In quantitative tests, a number of colonies are selected and the proportion of these colonies found to be confirmed positive is used as a factor by which the presumptive positive count is multiplied to determine the confirmed positive count.

**D.3. Confirmatory Testing.** Approaches to confirmatory testing vary depending upon the target organism. In particular, the number of tests to be performed varies from target organism to target organism. Sometimes, only one test result is required (e.g., coagulase), whereas other times a range of confirmatory tests are required (e.g., BAM method for B. cereus). Each of these tests has its own characteristics (rates of false positive, false negative etc.).

E. Organism variability

Microbiological analytical tests exploit one or more microbial characteristic to differentiate between those microorganisms included within the group and those excluded. The breadth of the designated group can be large (Gram negatives, *Enterobacteriaceae*) or small (*Salmonella enterica* serotype Enteritidis phage type 4, *E. coli* O157:H7). An ideal test would detect every microorganism that is intended to be within the group (sensitivity) and ignore every microorganism intended to be excluded (selectivity). Failure to recognize a microorganism that should be in the group is termed a false negative; conversely, accepting a microorganism that should not be in the group is a false positive.

The variety of properties used to group microorganisms range from physical structure (rods, spore formers), to metabolic characteristics (ability to metabolize a particular sugar, production of hydrogen sulphide), to the ability to survive toxins (brilliant green agar, antibiotics), to production of antigenic proteins (ELISA tests), to the presence of plasmids, specific DNA sequences (PCR tests), and to the ability to produce a toxin (*C. botulinum*, Enterohemorrhagic *E. coli*).

Unfortunately for the consistent grouping of microbes, bacteria typically do not have a consistent set of characteristics within the group. Nearly all enterohemorrhagic *E. coli* O157:H7, for example, cannot ferment the sugar sorbitol (unlike nearly all other *E. coli* which can ferment it) and this characteristic is used in identification tests. However, there is a sorbitol-fermenting serotype of *E. coli* that can also produce the shiga toxins. Furthermore, microbial characteristics are not static as DNA exchange occurs between bacteria at a much higher level taxonomic level than species. Because of these inconsistencies in the presence of microbial characteristics, there is a trend to identify microorganisms of public health concern by the presence of the DNA that codes for the particular virulence factors. This attempts to include all microorganisms that can cause a particular illness, regardless of their conventional strain, species or even genus designations. But even this strategy is not without difficulties as pathogenicity is frequently the result of a cluster of virulence factors and not all pathogens that cause that illness may have an identical or complete set of the virulence factors. In addition, some strains may posses the DNA for the virulence factors but the genes are never expressed, making those strains non-pathogenic. The pre-test environment, whether in a food or an enrichment medium, can sometimes affect the expression of an identifying characteristic.

In the development of a method to detect specific organisms or groups of organisms, care must be taken to select a characteristic that is shared by all the organisms to be included and absent in those to be excluded. Validation of a test protocol by testing against a wide range of microorganisms of both groups is necessary. Quantifying the rate of false positives or false negatives, however, is difficult and rarely done. In actuality, virtually all microbial tests are not as sensitive or selective as desired. Microbiologists rely on subjective knowledge and experience of the appropriateness of most tests for the situation at hand. This is demonstrated by the classic "fecal coliform" test widely used to as an indicator of the presence of sewage contamination in shellfish and water. However, this test is inappropriate (not sufficiently selective) for detecting sewage contamination on vegetables as there frequently are non-pathogenic soil bacteria that would be declared positive by the test.

F. Overall statistical/distribution considerations

The statistical considerations for characterizing method performance can be described by taking an example method and considering the probability distributions that would be encountered at each of the steps for the method. The example that follows goes through this process.

Assumptions: It can be assumed that there is a probability distribution of levels, -cfu/ml – (or cfu/g), throughout the product being sampled. The concepts that are needed for designating distributions need to be discussed at greater length. But, for the moment, assume that any pathogen or interfering organism is uniformly distributed throughout the 100 ml of the sampled material.

In this example, it is assumed that there is one type of organism of concern - the target organism - distributed uniformly with level, $r_t$, and there is a another type of organism – an interfering organism - distributed uniformly with level, $r_i$.

Steps 1-2 Prepare a 1:5 dilution; spread 1 ml of material on three plates of one type of agar (for the moment consider just one type of agar).

It is assumed that the 0.2 ml of the 100 ml sample is randomly selected so that the number of cells, $n_x$, of the target or the interfering organisms is distributed as a Poisson distribution with parameter $0.2r_x$, where the subscript 'x' is either 't' or 'i'.

Step 3 Grow colonies.

The $n_x$ cells are assumed to develop into $m_x$ colonies. If $f_x$ is the probability that a cell will develop into a colony, and we assume that the events of these occurrences among the $n_x$ cells are independent, then the distribution of the number of colonies $m_x$ is a binomial with parameters, $n_x$ and $f_x$, so that the expected value of $m_x$, conditional on $n_x$ is $n_x f_x$. It turns out then that the number of colonies, $m_x$, is distributed as a Poisson distribution with parameter $\lambda_x = 0.2r_x f_x$, so that, unconditionally, the expected value of $m_x$ is $\lambda_x$.

Step 4 Select 5 colonies from the $m = m_t + m_i$ colonies for confirmation. If one or more is positive for the target organism then the sample is classified as positive for the target organism; otherwise it is not, and thus is classified as negative. For the moment, assume that any selected colony will be identified properly.

The distribution of the number of selected colonies of the target organism is hypergeometric. Let $k_t$ be the number of targeted selected colonies. The probability of a classified positive sample is the probability that $k_t > 0$. This can be written as:

$$P(k_t > 0) = 1 - \frac{m_i!(m_t + m_i - 5)!}{(m_i - 5)!((m_t + m_i)!},$$

where here it is assumed that $m_i \geq 5$. If $m_i$ is less than 5 then there is 100% probability of selecting at least one colony of the targeted organism (provided, of course, that $m_t > 0$).

If $m_i$ is large, and $m_t$ is not, then the above probability could be close to zero. If both are large, but at an expected certain ratio, say, $E(m_t) = gE(m_i)$ - that is, for every colony of the interfering kind there are expected g colonies of the targeted kind - then the percentage of colonies of the targeted kind is: $m_t/(m_t+m_i) \approx \alpha_t = g/(g+1)$ and the hypergeometric distribution can be approximated as a binomial distribution with parameters 5 and $\alpha_t$. The above expression - the probability of a classified positive sample – can then be approximated as:

$$P(k_t>0) \approx 1-(1+g)^{-5}.$$

Example (for large levels):

1. If g =1, (that is, there is an expected equal number of targeted and interfering organisms) then the probability of a positive sample is $1 - 2^{-5} = 1\text{-}1/32 \approx 0.97$, or 97%. Or, in other words, there would be a 3% false negative rate.

2. If g = 1/3 – that is, for every three interfering colonies there is one targeted colony, then the false negative rate would be $(1.3333)^{-5} = 0.132 = 23.7\%$.

3. If g = 0.5 – that is, for every two interfering colonies there is one targeted colony, then the false negative rate would be $(1.5)^{-5} = 0.132 = 13.2\%$.

4. If g = 2 – that is, for every interfering colonies there are 2 targeted colonies, then the false negative rate would be $3^{-5} = 0.004 = 0.4\%$.

5. If g = 2 but instead of 5 colonies only 3 colonies were tested for confirmation, then the false negative rate would be $3^{-3} = 3.7\%$.

   The above calculations indicate that the number of tested colonies can be important when there is a significant percentage of interfering colonies expected. Even when g = 2, the false negative rate is 0.4% when there are large numbers of both types of colonies, which could be considered large in some applications. The question that needs to be addressed is: <u>what values of g are possible or likely?</u>

   The types of uniformity or distributional assumptions made in these situations are paramount to the validity of the calculations. For the above scenario, it is assumed that the types of cells are distributed independently and uniformly within the sample. However, in reality it might be more realistic to assume an 'extreme' negative correlation of some sort between the types of cells, so that values of g are either close to 0 or 1.

<u>Example Calculations</u>:

a. Suppose, $r_t$ = 10 cfu/ml, and the likelihood for growth is $f_t$ = 75%, so that in a 0.2 ml sample, there would an expected 1.5 colonies of the targeted organisms. For the interfering organisms, assume that $r_i$ = 20 cfu/ml, and $f_i$ = 75% as well, so that there would an expected 3 colonies of the interfering organisms. Thus, the value of g would be 0.5. However, the expected number of cells in the 0.2 ml sample is small, so an exact calculation for determining the probability of a false negative would be needed. The probability of a (false) negative result is, Pn = 24.35% - almost 25% of the time, the results would be negative for the target organism.

b. Double $r_t$ (= 20 cfu/ml) and $r_i$, keeping everything else the same, Pn = 13.86%, close to the theoretical asymptotic result of 13.2% given above in 2).
c. $r_t$ = 10 cfu/ml and $r_i$ = 5 cfu/ml (so that g = 2) Pn = 22.32%.

d. $r_t$ = 20 cfu/ml and $r_i$ = 10 cfu/ml, then Pn = 5.1%.

e. $r_t$ = 30 cfu/ml and $r_i$ = 15 cfu/ml, then Pn = 1.35%.

f. $r_t$ = 40 cfu/ml and $r_i$ = 20 cfu/ml, then Pn = 0.6%.

g. $r_t$ = 60 cfu/ml and $r_i$ = 30 cfu/ml, then Pn = 0.4% (close to the asymptotic result).

h. $r_t$ = 100 cfu/ml and $r_i$ = 50 cfu/ml, then Pn = 0.4% (just for emphasis).

i. $r_t$ = 100 cfu/ml and $r_i$ = 200 cfu/ml, then Pn = 13.2%.

Figure 1 presents the operating characteristic (OC) curves for the probability of a negative finding, Pn, versus the assumed level of the target organism, $r_t$, assuming different levels of the interfering organism, $r_i$; where $r_t = gr_i$ . It is assumed same growth likelihood and recovery of 75% for both types of organisms, and 5 colonies are tested for confirmation.

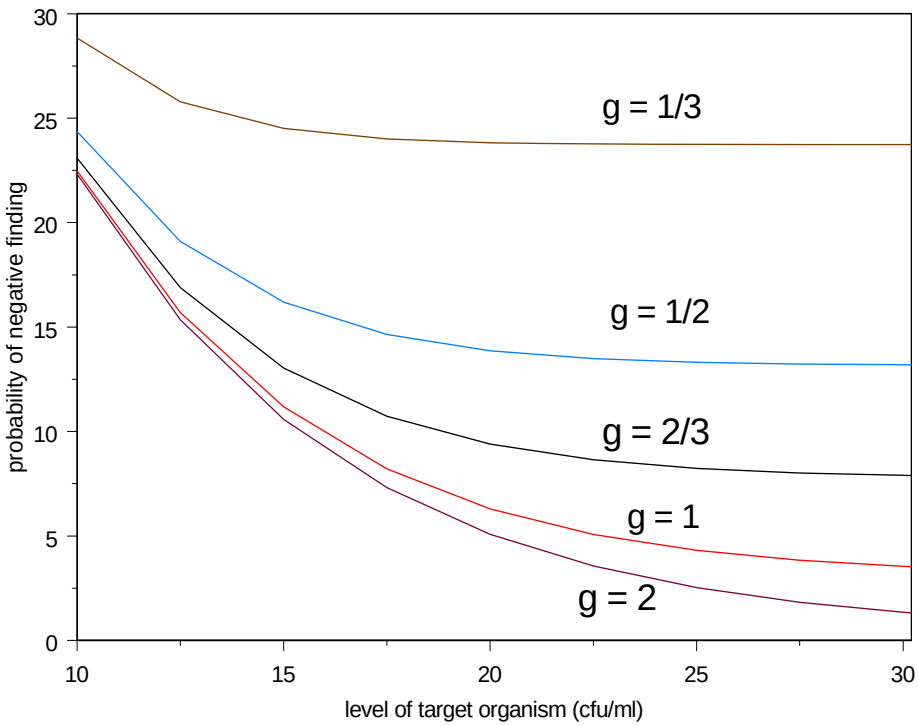



Figure 1: The OC curves for the probability of a negative finding at different levels of target and interfering organisms. For every colony of the interfering kind there are expected g colonies of the targeted, so that g = 1 means that there are expected the same number of interfering and targeted colonies; g = 2/3 means that for every 3 interfering colonies, there are an expected 2 targeted colonies

## Summary and conclusion

This review is a bottom-up approach that attempts to identify and quantify all potential errors of a laboratory method. From a mathematical characterization of the results associated with each source of variation, a derivation of the expected operating characteristics of a method can be made. However, there are other unidentified sources of variation that are not likely to be captured when studying the method through a bottom-up approach that are associated with human errors, for example pertaining to equipment settings and calibrations, as well as unexpected changes in environmental conditions that cannot be captured in small, controlled, laboratory studies. The importance of these might be determined through ruggedness tests, where the parameter specifications for critical steps of the method are changed slightly from their nominal values in order to determine the effect of small changes. Ideally methods that are used are rugged in that the results are not affected greatly by small changes of the method's specifications. To the degree that a test is rugged, the bottom –up approach for determining the magnitude of variation (of results) will capture a large portion of the total variation. Thus we have recommended ruggedness testing to be part of method validation (BPMM Task Force Report Executive Summary).

While inter-laboratory studies may be needed to develop reproducibility measures that basically validate methods to be used by qualified analyst , it is still critically important to identify the sources of variability in a method and quantify their effects within laboratory. These can be used for quality control monitoring. In addition, if definitive inter-laboratory studies providing reliable measures of method performance do not exist, then performance measures determined from a series of a bottom-up studies, identifying and quantifying variability associated with the critical steps of a process should be conducted that can be used for laboratory QA. It is possible that the performance operating characteristics estimated will be accurate when using a bottom-up approach particularly so if it can be shown that the method is rugged.

References

Dahms, Susanne. 2004. Microbiological sampling plans – statistical aspects. Mitt. Lebensm. Hyg. 95, 32–44.

International Commission on Microbiological Specifications of Foods (ICMSF). 2002. Microorganisms in foods 7: Microbiological testing in food safety management. Kluwer Academic/Plenum Publishers, New York, p. 118.

Jaynes, ET. 2003. Probability Theory: The logic of Science. Cambridge University Press, p170-171.

Niemela, Seppo, I. 2002. Uncertainty of quantitative determination derived by cultivation of microorganisms. Publication J3/2002, Center for Metrology and Accreditation, Advisory Commission for Metrology, Helsinki, Finland.

# Enclosure B - Statistical Process Control for process control of microbiological levels

## Executive Summary

As mentioned in Appendix D, the purpose of sampling and thus measuring something is to make some type of inference or evaluation of some population property. In Appendix D a short discussion of the effects of sample and measurement errors on evaluation was given. This appendix discusses in more detail one general sampling application: Statistical Process control (SPC), which is a type of quality control (QC) sampling used to control a process. Statistical Process Control (SPC) has been used in the manufacturing setting for many years for controlling the quality of produced items. Recently its applications have been extended to microbiological output for use, successfully we believe, in ensuring the safety of processed foods or other items that might present a hazard to consumers of the product. In addition SPC can be used by laboratories in helping ensure that the measurement process is "in control" – that is, that the measurement deviations from the true value, over time can be considered as being independent of time and within specifications that might have been determined from collaborative or inter-laboratory studies. This can be accomplished using split samples or check samples, and occasionally comparison with another more authoritative method.

There are two features that characterize SPC and differentiate it from other types of sampling, namely acceptance and survey sampling. These types of sampling involve taking samples from a well defined population of units, specifying an upper bound to the number of samples that would be taken, and, from the results obtained from these samples, making a decision or an evaluation abut the population that was sampled. As opposed to these types of sampling, SPC sampling does not involve specifying a fixed upper bound number of samples or necessarily identifying clearly a population of units. Rather, SPC involves sequential sampling over time, accompanied by a set of rules or criteria that are used to make decisions or evaluate, not so much a well defined set of units, but rather the process that is creating the units. The second feature that characterizes SPC is that the underlying values of parameters that are used to construct the rules are derived from results from sample units that were created by the process itself. In order for this to be done in meaningful way, the parameter values should be reflecting the process when it is in control. Thus, SPC as a subject matter, involves methodology for judging this – when can it be considered that a process is in control so that the rules that are to used for evaluating whether or not the process is or remains in control are valid. SPC involves evaluation of the process and not specifically whether produced units or obtained measurements are within some pre-defined specifications.

Laboratories can use QC procedures for assuring that the measured results being produced are within specifications that are defined by repeatability or reproducibility parameters. SPC though offers a degree of flexibility that takes into account the actual system or process of measurement, insofar as the criteria for evaluation are not derived from outside the process but are derived from within the process itself. The full application of SPC entails a continuous examination of the data with the purpose of not

only just judging whether or not a process is not producing as it should, but also that the process has the capability of producing better than it was initially thought it should, by helping identify areas of potential improvement. That is, evaluative criteria can change, taking into account the potential capability of the process.

Thus, this document has a twofold purpose. The first purpose, the primary one and the reason for the document, is to present "performance standards" regarding the application of SPC for microbiological output of a process. However, a second purpose is to provide a simple introductory paper that could serve as a beginning point for learning about SPC and its application for microbiological data. Thus examples in Appendix G1 are given that demonstrate principles that are rooted in the performance standards.

The sampling work group is recommending the following "performance standards" with respect to implementing SPC for microbiological data. The performance standards are not meant to prescribe procedures or criteria that should be used for evaluating processes; rather they are meant to provide guidance and a methodology to be used for developing a SPC sampling plan. Following the performance standards are discussions of them, a conclusion section, and specific examples (seven in all) given as Appendices of the report (BPMM report Appendix F.1). The examples include SPC for qualitative or attribute data, including binomial Poisson –like, and negative binomial distributions; continuous variable data of high levels of generic E. coli; and an example which uses SPC for tracking the occurrence of infrequent events such as the finding of E. coli O157:H7 on samples. Hopefully these examples will serve as useful material.

**Performance standards:**

1. Charts of plots of the output data over time are not only valuable for verifying calculations and having a visual picture of the variation exhibited by the process output, but also it is an integral tool to be used for identifying sources of unexpected variation in output leading to their elimination. Thus charting is a necessary tool needed to gain the full benefit of doing SPC.

2. Results to be plotted in a control chart, when the process is under control, used for statistical process control should be normal or nearly normally distributed. In cases where this is not true and an alternative known distribution cannot be assumed such as a Poisson, binomial, or negative binomial distributions[i], transformations such as the log transformation for microbiological counts, arcsine transformations for binomial data, or a square root transformation for data distributed nearly as a Poisson distribution should be considered.

3. During some "initial" period of time, it is assumed that the process is operating in a relatively stable manner – or is in control. During this period the distribution of the measurements should be estimated and rules for evaluating the process should be formulated. The statistical "rule of thumb" of using about 20-30 results or more for computing means and standard deviations or other summary statistics

needed to estimate the distribution of results and construct control limits is a recommended and desirable goal.

4. Rules for evaluating process control should be set with aids assessing the two types of errors: Type 1, declaring the process out of control when it is not, and Type 2, not declaring a process out of control when it is. Typically there are two measures, depending upon the nature of the rule, that are used for assessing these errors: 1) the probabilities of the two types of errors at a given time (referred to as $\alpha$- and $\beta$- probabilities, respectively); and 2) the average run length (ARL) – the expected number of samples before an out of control signal (one of the rules being not met) is seen.

5. When a process is thought to be "in control," the limits for assessing individual results are set at some distance from the average, expressed as standard deviation units from the mean or process target value. The default distance is 3 standard deviations[ii]. Limits other than these should be implemented when taking into consideration economic and public health costs of incorrect decisions regarding whether the process is in control. When developing rules, the $\alpha$-probability (for the Type 1 error) should be kept low, for example, below 1%.

6. There are numerous run/trend rules that can be used, such as runs test, moving averages and CUSUMS, for detecting shifts in the process mean; and rules for detecting shifts in the process variation or other auto-correlated patterns that could be due to systematic source of variation. The use of any of these may depend upon particular expected conditions that arise when the process is out of control, and the sensitivity desired for detecting such conditions. In assessing the use of these rules, one should consider the ARL. It is recommended, when the process is in control, that an ARL should exceed 100 (corresponding to a less than a 1% $\alpha$ - error).

7. Specification Limits are not Statistical Process Control limits; specifications are either customer, engineering, or regulatory related. Statistical Process Control limits are process related. Specification limits should not be placed on a control chart insofar as these might be considered as process goals thus influencing the efficacy of SPC procedures for ensuring a controlled process, and thereby undermining the safety of the product.

**Performance standard 1 – the necessity of charting**

Statistical process control (SPC) involves two aspects: use output data from a process to establish an expected distribution of values of some variable which is used for judging the control-status of a process when the process is (thought to be) in control; and a set of rules or criteria for which (future) output values from the process must satisfy in order not to declare, or declare presumptively, the process is out of control. In establishing the distribution to be used for determining the control status of the process, besides the output data, various other, regulative, type judgments are used that can affect the assumed distribution and the rules that are used for evaluating the process.

One feature that is included in the SPC methodology is charting – plotting of output process data values that are used for evaluating the process versus time or sample number, and examining the charted or plotted data. A question might arise is: why is this charting necessary? The implication of the question is that it may not be necessary, particular so with today's computer technology – all that is needed is to somehow feed the data into a computer program and the program would make the calculations, determine whether or not the rules were violated and thus provide the control-status of the process. Various answers to this question can be given. One answer could be that charting provides a confirmation of the calculations; however, with today's computer technology there are many other ways of ensuring that the calculations are correct to the extent that if there was a noted discrepancy between the plotted data and the computed results it more likely would be due to an error in plotting rather than in calculations. Thus, the answer to the question involving "looking" at a chart for the purposes of confirmation does not provide a good reason for the necessity of charting. Another answer might be based on psychology – the chart provides management with a visual picture of what is happening and this would give them a greater understanding of the process than what could be gained by examining sets of numbers and adherence of them to a set of rules. This answer by itself though would not provide a necessary reason for charting, at least not one in which a requirement of charting is recommended since there really would not appear to be a concrete gain from plotting.

However, this last answer is getting closer to the reason that compelled us to recommend, necessarily, charting, rather than just pointing out that charting is useful for the above stated reasons. The "seeing" of the chart can convey an understanding of the process that adherence to a set of rules cannot. Thus while the "looking" at charts can provide the confirmatory and psychological assurance, the "seeing" – meaning, a more in depth examination of the charted data - can provide additional information about certain aspects of the process that might have been unanticipated initially so that prior "rules" reflecting these aspects were not constructed. From "seeing" a chart, new insights might be gained that could show the inadequacy of the selected rules or could provide motivation for the development of new rules that lead to identifying unanticipated sources of error and an improvement of the process; on the other hand, however, it could lead to explorations that do not lead to improvements and thus could lead to an inefficient use of time and resources. Thus, to help prevent incorrect decisions statistical analysis (retrospectively) of data should be performed (See Appendix 2). The "look and see"

approach to charting is emphasized in SPC, notwithstanding possible pitfalls associated with this.

**Performance standard 2 – The control distribution**

Statistical Process Control, (SPC) has been used successfully to control quality and costs of manufactured products since the late 1920's. This statistical tracking system used for monitoring processes performance was developed by Dr. Walter Shewhart[iii]. He discovered that variation observed in manufacturing output was visually "different" from the variation that he would expect to see for similar type characteristics in nature for a stable system. Dr. Shewhart speculated that the variation that was not expected was due to processing errors by either labor or management. In other words, if the process was "under control," the deviations from a mean value of statistical measurements that "track" some feature or output of the process would be distributed in a "random" looking fashion without any clear patterns, "unimodally" or at least displaying some degree of "regularity" or "stability" with very few outlier values. Further, it was assumed that the errors would be symmetrically, or nearly symmetrically, distributed around the mean value. In other words, normality, or near normality, is a natural distribution to assume when a process is under control since it is then assumed that the deviations are "caused" by many, inherently uncontrolled factors, each contributing only a small amount to the magnitude of the deviation. Historically then, in the manufacturing setting, rules or control limits for assessing a process to be out of control were set symmetrically with respect to the mean value – the assumption being that a result could be equally likely above as below the mean value. Thus, ,the distribution of the plotted values for the control chart was assumed to be normal and the operating characteristics of the rules - the probability of declaring the process out of control as a function of the true process mean - were evaluated assuming the underlying distribution of results is the normal distribution.

For microbiological data the above assumptions may not be true – rather, often (explicit examples are given in Appendix F.1) distributions seen will not be symmetric. If the non-symmetric distribution is known, then it is possible to use this distribution directly with the accompanying mathematical calculations to derive control limits with certain desirable operating characteristics. In such a situation parameters of these distributions can be estimated by maximum likelihood estimation or other statistical procedures and control plans can be determined directly using estimated distribution. However, often these specialized assumptions cannot be made, since with processing and measurement there would be expected unavoidable differences over time that could be caused by factors related to slight variations of equipment settings, environmental conditions and personnel that cannot easily be controlled or completely eliminated. For example, it might be assumed that under ideal conditions, the plate count distribution would be Poisson, with a parameter, $\lambda$ - representing, in this case, the expected value. But value of this parameter may not be constant from day to day, or sample to sample, rather, $\lambda$ itself would be a random variable, taking on possibly different values for different samples. Because of this ($\lambda$ being a random variable), the total variation seen in the obtained results would not be expected to be equal to the expected variation of results seen from a Poisson distribution. The distribution of the results thus might be represented

well as a mixture of Poisson distributions. One such distribution is the negative binomial distribution, which has two parameters.

In general though, the expected distribution when the process is in control may not be known other than it most likely would not be symmetric. And for the classical SPC control procedures (as described below for Performance Standard 3), the limits are set using sample mean and standard deviation values for results on sample collected from a process assumed to be in control or nearly so, as if the distribution of these results were generated from a nearly normal distribution. If the distribution of results is not nearly symmetric, then transformations of the output variable, for example, taking the logarithm of microbial plate counts, may induce a more symmetric looking distribution. There is often another advantage of using the transformed variable: namely, the expected standard deviation would be less dependent on the expected mean value of the particular result. Thus, if plate counts were thought to be distributed as nearly lognormal, then a log transformation would make the distribution nearly normal and the variances of each transformed result would be nearly uniform for the data. Similarly if the data results were thought to Poisson-like distributed, a square root transformation of the results would make the results more symmetrical and make the variance more uniform (Appendix 3); for the binomial distribution, the arcsine transformation, $\sin^{-1}[(x/N)^{1/2}]$; and for the negative binomial, the inverse hyperbolic sine transformation, $N^{1/2}\sinh^{-1}[(x/N)^{1/2}]$ would make the distribution more symmetric and the variance more uniform (Johnson and Kotz, 1969).

While a normal distribution of the deviations from the mean value is not an absolute necessity for applying the control techniques discussed in this paper, historically the stated probabilities describing the operating characteristics of the control plan are computed assuming normal distributions and used for motivating decision rules. As a result of these considerations, performance standard 2 is recommended.

.

**Performance standards 3 and 4 – Establishing the control distribution and rules for process evaluation**

SPC is applied as follows:

1) During some "initial" period of time, it is presumed that the process is operating in a relatively stable manner, as described in the preceding paragraphs. This is a very important presumption and in actuality to reach this point when the process controls and parameter values are set, it may be needed an extended period of experimentation or trials. Whenever possible, independent validation of the presumption of process control should be made by other means, different from the statistical process control planning to be used, such as, for laboratory QC, the use of reference standards or cultures with known characteristics. If the distribution of results is expected to be nearly normal, then during this period statistical measurements should be distributed randomly around a mean value, $\mu$ with a standard deviation, $\sigma$. Values for these parameters are estimated during this time.

2) Over time, the statistical measurements are plotted on a graph, called a Shewhart chart (see Appendices for examples), showing the distribution of the statistical measurements. The Shewhart chart is basically the plot of the measured values versus sample number, starting with some sample labeled 1.

3) If the plotted statistical measurements do not meet any one of a set of criteria the process is considered to be "out of control."

The criteria are chosen to reflect different manifestations of "out of control" of interest to the producer. Particular types of "out of control" signals are: a) "short term" non-systematic errors that might occur that result in an unacceptable product for a given day or lot; b) persistent errors that cause a systematic deviation from the pre-designated target value, $\mu$; and c) persistent errors that cause an increase in variability ($\sigma$) of process output.

Decision errors in regard to deciding whether or not a process was under control are similar to decision errors guarded against by the use of statistical procedures when testing two competing hypotheses in science. That is, a Type 1 error is made by deciding that the process is out of control when, in fact the process is in control and thus would not require adjusting; and a Type 2 error occurs when a process is not adjusted (actually is out of control) but it is decided that it is not out of control and the process is left as is. The probabilities of these errors are, respectively, referred to as $\alpha$ - and $\beta$- probabilities. Both of these errors could contribute to processing inefficiencies.

Processes can be affected by Type 1 and 2 errors because management and hourly workers often make adjustments that should not be made or fail to make adjustments that should be made in the attempt to "improve" the process output. The psychological forces that lead to changes or no-changes and thus errors influence the output of a process. A belief could develop, particularly the more one gains experience with the process, that ad-hoc adjustments based on one's expert judgment would lead to a better process and output than just relying on pre-set rules as implied by charting and SPC. While in certain circumstances this may be true, often times it would not be so, and such a belief (of the advantages of following expert judgment) is not a reason to resist placing control charts on a process and using SPC. If nothing else, the use of control charts and SPC helps establish objective criteria for making adjustments (once the limits are established).

In other words, SPC and the use of rules for evaluating the process, determining $\alpha$ - or $\beta$-probabilities of the rules are not meant to eliminate expert judgment; rather these activities should be viewed as an aid for making judgments helping to prevent unwarranted actions that lead to a Type 1 or Type 2 error. "Out of control" signals can be considered "presumptive" regarding whether the process is out of control; and the examination of the data once plotted can lead to judgments of "an out of control process" that the charting "rules" have not reflected. Thus performance standards 3 and 4 we consider to be necessary for preventing the dominance of expert judgment in the evaluation of a process, but is not meant to eliminate it.

Consequently, SPC in its fullest sense involves preliminary analyses or testing of the process to a point where process parameters have been determined, such that it is believed that when the process is operating in accordance with the parameter specifications, the distribution of the measured output that is being used for evaluating process control would have the characteristics described above (random, stable, nearly symmetric or with some other designated distribution). In the developmental stage of the process this assumption may not be true. The SPC and plotting techniques described here can also be used in the developmental stages; however, in this situation the criteria for out of control may need to be changed.

**Performance standard 5 – Control limits for individual values**

Dr. Shewhart understood that in order for a control system to work effectively the rules or criteria used for determining the control-status of a process should meet a couple of requirements. These requirements include:

1. The $\alpha$-probabilities (of incorrectly saying a process was out of control when it was not) must be low enough so to not unnecessarily create delays in processing (which could be costly) and fatigue workers and management from looking for causes of variation that do not exist;

2. The criteria must be robust enough so that a number of probability distributions can be accommodated by the procedures; and

3. The criteria should be simple and easily "seen" on a graph[iv].

As a consequence of the above considerations, Dr. Shewhart settled on his most well-known criterion that placed, what he termed, "Control Limits" a distance of three standard deviations from the process average; that is, control limits were set such that if a single measured value, (labeled often as $X_i$), was either greater than $\mu + 3\sigma$ or less than $\mu - 3\sigma$, with $\mu$ being the process average or intended process target then the process was to be presumed out of control. When the underlying distribution is normal, then the probability of exceeding one of the limits is 0.135%, so that the two-sided $\alpha$-error is 0.27%. For most distributions expected for processes under control, the likelihood of seeing measured observations that do not satisfy these criteria is small[v] thus satisfying requirements 1 and 2 above. Also, the third requirement is clearly met because the limits are just horizontal lines on the chart, and it can be easily seen if a plotted point is not between the two lines, indicating "out of control." This criterion would "catch" a processing error that might not be systematic, and, when not met, would imply that there is some aspect of processing that might not be controlled.

## Performance standard 6

### Tracking trends or shifts in the process mean value

There are many ways to evaluate "systematic" errors that would cause the mean value of the process to change. One very simple way, which can be easily seen on a Shewhart chart, is to use "run" tests, for example, to declare a process "out of control" when 8 consecutive points fall on the same side of the target value (e.g. process mean) - a run of length 8. When the process mean value equals $\mu$, such a pattern is highly unlikely, (assuming here a symmetrical distribution of measured values) so that when such a pattern is seen it is likely that the process mean is not equal to $\mu$. A run of 8 consecutive results above or below the targeted mean value has, for those 8 results, a 2 in 256 chance of occurring (accounting for the two possibilities of 8 results above the mean or 8 results below the mean) or about a 0.8% probability, ($2(0.5^8) = 0.0078$). However, with such tests, it takes one result to break the pattern. A criterion might be set as: if at least 7 of 8 consecutive results are above or below the mean value, then the process would be considered as out of control (or presumptively out of control, pending further investigation). The probability of at least 7 out of 8 observations above or below the mean value has a probability of 7% (from the Binomial probability distribution with an incidence parameter with a value of 0.5). That is to say, it would be expected that 7% of any 8 consecutive sample results to have at least 7 of the results above or below the mean value when there is for each result a 50% chance of being above or below the mean value. Thus if a criterion of at least 7 of 8 results are above or below the mean value the process would be presumed out of control, the $\alpha$-probability would be about 7%. For a one-sided test that is, for example, a test for which the concern is only with a process change that results in an increase of the process mean value, the $\alpha$-probability would be 3.5%. This percentage is usually considered too high, given the costs associated with investigating a presumptive out of control signal.

Because just one result can "break" the pattern, runs tests are not very "powerful" for detecting small or even moderate shifts (relative to the standard deviation) – that is the $\beta$-error may be large. For example, if the mean increased by one standard deviation unit, so that the true process mean changed to: $\mu + \sigma$, then the probability of an individual result being below the target value, $\mu$, is 16%, (84% of the values will be above the target). For 8 consecutive results, the probability of having at least one result below the target value of $\mu$ is about 75%, so that the $\beta$ error associated with this criterion would be 0.75, (for a single run of 8 values). The criterion for the upper control limit would not help much: there is an 83% probability that all 8 results would be below the upper control limit of $\mu + 3\sigma$.

For this reason, moving averages and CUSUMs are often used for monitoring processes, where a moving average is the average of the results in a group of consecutive samples and a CUSUM is a procedure that accumulates iteratively deviations from the target value. Moving averages are more difficult to compute because the samples used for computing averages are always changing and thus at any time the results over a (changing) set of samples need to be known. Also there is an issue of how many samples

to use in computing the moving average (the window length of the moving average). The CUSUM (Johnson and Leone[vi], 1964; Juran 1988[vii]) control procedure avoids these problems and is thus simpler to compute and to design. The CUSUM value is basically updated at each sample by adding the deviation: $X_i - \mu$, to the previous value, where $\mu$ is the target value, for example, the (expected) process mean[viii]. When CUSUM values are plotted versus sample number, evidence for a shift[ix] in the process mean value is easily seen when the graph of the points steadily increases or decreases. For a simple charting of CUSUM[x] where a control limit can be depicted such that a process out of control signal would be given when the CUSUM value exceeded the limit, L, a slight modification of the CUSUM as described above can be made, namely: computing, $S_i = \max(0, S_{i-1} + X_i - \mu)$, where $S_i$ is the CUSUM value for the $i^{th}$ sample, and $S_0 = 0$ (or some other value for a' quick start'). When $S_i > L$, this would imply that there was a positive shift in the process mean some time in the recent past. (An example of the calculations is given in Appendix 2.) A similar type CUSUM can be constructed for negative shifts of the process mean. Developing limits for these is more complicated and is out of the scope of this document, however, it is encouraged that these procedures be considered when designing control charts.

To measure the effectiveness of sampling plans and sampling criteria (or rules), another parameter, called the average run length (ARL) is used. The run length, RL, is a random variable that counts the number of samples (starting at some specific sample) before the first signal for "process out of control" is given; in other words, the number of samples until at least one of the sampling plan's criteria or rules for declaring the process out of control is obtained. By convention, the sample for which there is a signal is counted, thus all run lengths are greater than or equal to 1. Control plans and their criteria are often evaluated by characterizing the distribution of the run lengths, and in particular by the expected number of samples – the average run length (ARL) - before the first "out of process" signal given, starting from some specific sample. When a process is in control, the desire is that the ARL should be large; and of course, when the process is out of control the desire is to have a small ARL. ARL has become a traditional parameter to consider, but other parameters of the run length distribution could be considered as well, for example, selected percentiles of the RL distribution.

A simple example, presented below, shows comparisons between a CUSUM rule and the "eight in a row" rule described above. In the example, a CUSUM rule is given that has a comparable ARL when the process is in control to that of the "eight in a row" rule discussed above. The following table gives the average and median run lengths. The values given in the table were determined from 20,000 simulations[xi]. The assumption for the underlying distribution is normal, with mean equal to $\mu$ and standard deviation = 1. Control is when the mean, $\mu = 0$.

Table 1: Results: each entry determined from 20,000 simulations.

CSUSM calculations: $S_k = \max(0, S_{k-1} + x)$, where x is distributed normal with mean = $\mu$ and standard deviation = 1, and $S_0 = 0$. The process in control mean = 0; when $\mu > 0$ the process is out of control. Out of control signal when $S_k > 21.5$. The parameter "mu" is the number of standard deviation the process has drifted from the target ($\mu$).

| mu | CUSUM mean ARL | CUSUM median ARL | 8 in row mean ARL | 8 in row median ARL |
|---|---|---|---|---|
| 0.0 | 510.7 | 392 | 506.9 | 351 |
| 0.1 | 177.1 | 152 | 299.9 | 210 |
| 0.2 | 100.8 | 91 | 183.0 | 129 |
| 0.3 | 69.6 | 65 | 121.2 | 86 |
| 0.4 | 53.8 | 51 | 82.6 | 59 |
| 0.5 | 43.2 | 41 | 58.9 | 43 |
| 0.6 | 36.5 | 35 | 43.7 | 32 |
| 0.7 | 31.4 | 30 | 33.7 | 25 |
| 0.8 | 27.5 | 27 | 26.8 | 20 |
| 0.9 | 24.7 | 24 | 22.2 | 17 |
| 1.0 | 22.2 | 22 | 18.9 | 15 |
| 1.1 | 20.3 | 20 | 16.4 | 13 |
| 1.2 | 18.6 | 18 | 14.4 | 11 |
| 1.3 | 17.2 | 17 | 13.0 | 10 |
| 1.4 | 16.0 | 16 | 11.9 | 8 |
| 1.5 | 15.0 | 15 | 11.1 | 8 |

The chart shows that for a value of $\mu$ less than 0.7 standard deviation units but greater than zero, the CUSUM has smaller ARL than that for the "8 in a row rule." For shifts in the mean value of ¼ standard deviation units, the ARL for the CUSUM is nearly ½ that for the "8 in a row" rule.

For the above control plans, when the process is in control, the ARL is about 500 corresponding, in a sense, to an $\alpha$ - probability of 0.2% (since for a rule on individual results with this probability of a signal, the expected value of the number of samples before the first signal using a geometric distribution would be about 500). The median value for this geometric distribution is about 345, corresponding reasonably close to the median run lengths shown above for when the process is in control.

There is a great deal of literature on designing moving averages and CUSUMs, and much useful information about these can be found in the above mentioned books or even on the internet (from a reputable organization such as the US National Institute of Standards and Technology).

**Tracking process variability**

The process standard deviation, $\sigma$, also can be tracked in many ways. A very simple criterion based on the absolute value of the differences of consecutive measured values, MR – for moving range, can be used to track the process standard deviation. This is not a very robust measure; better might be to group more results, and compute the moving standard deviations or moving ranges for the results in a group, however, these statistics are not as easily computed and plotted and may have little meaning for

processes where data are relatively rare such as microbiological data. The MR on the other hand involves only having knowledge of the most recent and second most recent results. Another option when possible is to subgroup data into discrete subgroups and to calculate the range, (high minus low observation within a subgroup). However, for data that are expensive to obtain, hard to gather and or relatively rare, charting of MR values and using them to get a visual understanding of the process has become popular.

**Performance standard 7 – Specification limits**

Other, non-process control related values, such as, specification limits should not be placed on control charts. The reason for this is psychology. Specifications are something that all individuals who deal with a customer are accustomed to meeting. These specifications may be engineering specifications, customer requirements or regulatory critical limits, to name just a few examples. Since people are accustomed to meeting these values, and, as a consequence, specifications are given a higher priority than control limits. From a process control stand point this is not the ideal situation, insofar as the goal of process control is to achieve the best control possible for given resource constraints. Reducing variation is a particularly important goal when microbiological quantities that could represent a hazard to human health are the object of the control procedure. For example, if specification limits are "looser" than the obtainable SPC limits for a particular process and one were to make adjustments based on the specification limits rather than the control limits then adjustments would be made less often than the SPC limits would require, thus creating type 2 errors. This lack of control of a process could mask undiscovered sources of error, which if persistent could result in a product that is unsatisfactory or unsafe. In other words, a process which is not controlled and thus for which there may be unidentified sources of variation, by the mere fact of there being unidentified sources of variation not being controlled may, without the producer being aware of it, result in unsafe product. Sampling, per se, cannot be counted on for assuring to customers a product is within specifications when the process is not in control – rather only good process control can assure that. For these reasons it is advised that only process control related values be placed on the control chart.

**Conclusion**

The SPC chart can be an important aid in identifying when and where an investigation for a cause for the process being out of control should commence. The low $\alpha$-probability does not imply that, when a process is in control, "out of control" signals would not occur. However, since these occurrences are not expected frequently, the occurrence of one encourages an examination of the process in search for "Assignable Causes" for each out of control signal. However, if out of control signals occur more frequently than what would be implied by the $\alpha$-probability, random chance –the unlucky draw- should be ruled out as a possible reason for the signals, and that there is an "assignable cause" for the excessive variation in the process output and/or one or more of the process parameters are incorrectly set. This would then call for a more rigorous review or further study of the process. If the plot of the data shows an abrupt change from consistently being in control to consistently being out of control, then it can be

concluded with high confidence that there has been an enduring failure somewhere in the process that requires immediate remediation. The plotting of the process may reveal a gradual, progress loss of control over a series of lots or production units. This pattern could result, for example, from a piece of equipment steadily becoming out of adjustment or a progressive environmental contamination resulting from an inadequate sanitation program. Another pattern could show a transitory but reoccurring or cyclical loss of control, e.g., every Monday morning. While no explicit criteria are given for detecting these types of cyclical patterns, one could use the "run rules" of 8 in a row (discussed above), e.g., if for 8 Mondays, the plotted point is above the target value, it would be suggested that for some reason results for Monday are "out of control." The SPC plots can also document improvement in process control resulting from deliberate alterations or added mitigations. The lower levels due to the process alterations are used to establish new process standards.

When the limits for declaring a process out of control are exceeded too frequently, a producer always has the option to accept the implied non-desirable or optimal processing. Whether this option is taken depends upon 'costs' (technical feasibility, monetary) of fixing the problem, e.g., taking measures that would reduce either the process mean or the process variation. For example, the likelihood of a process being declared out of control with respect to some microbiological indicator variable could be reduced by increasing the heat processing temperature. However, this mitigation requires more energy consumption and may reduce sensory and nutritional quality of the food product. Reducing the variation might be accomplished by simply improving the air circulation within the oven or the one-time expense of a new oven. This mitigation would likely have an additional benefit of reducing the proportion of product that was over cooked, thereby improving the sensory and nutritional quality. This example shows a general rule: it is generally more advantageous to reduce variation first. If that is not successful, then a process step(s) may need to be redesigned to lower the entire distribution by lowering the process mean.

In the past almost eighty years the genius of Dr. Shewhart's methods have proven themselves, and are as effective today as four score earlier. Although Dr. Shewhart used some biological examples in his book, "Economic Control of Quality of Manufactured Product, (D. Van Nostrand Company, Inc., New York, 1931)," he did not make reference to their use with regards to microbiological data. Another classic book for Quality Control that provides various statistical process control procedures is Juran, JM, 1974 Quality Control Handbook, third edition. McGraw-Hill Book Co. NY.

The brief summary presented in this document will not adequately cover the subject matter of SPC and quality control charting procedures. To aid the reader in gaining an understanding of SPC, this document includes seven examples – presented as Appendices – that cover some microbiological uses of standard SPC Charts and variations of the standard Shewhart chart which uses Shewhart's $\alpha$-level for setting control limits and other out of control rules. The examples are based on computer generated simulated data, or, in one case, constructed data; the primary purpose of these examples is just to illustrate procedures and approaches for analyzing the data and

implementing SPC. From these examples, it is hoped the reader will get an idea of the uses of control charts and will be motivated to pursue the subject matter further.

There are 7 examples, all found in Appendix F.1 – SAWG SPC Appendices:

Appendix 1: Classical SPC – generic E. coli levels, treated as a variable (continuous) data.

Appendix 2: Counts, using a Poisson distribution (C- chart).

Appendix 3: Counts, not using a Poisson distribution, but rather comparing a negative binomial distribution and a square root transformation.

Appendix 4: Proportions, using a binomial distribution (NP- chart).

Appendix 5: Proportions, using a binomial distribution with different numbers of units per sample (P- chart).

Appendix 6: Counts, using a Poisson distribution, with different sample sizes (U-chart).

Appendix 7: Infrequent events, based on exponential distribution (F- chart).

---

[i] If such distributions are assumed then goodness-of -fit statistics should be given for verification.

[ii] Low values for microbiological measures would not be considered as undesirable or that, necessarily, the process is out of control. Rather consistent low values could be considered as evidence of that an improvement in the process could be made.

[iii] Walter A. Shewhart, "Economic Control of Quality of Manufactured Product, 1931"

[iv] We have not seen this requirement attributed to Dr. Shewhart, but it is certainly implied by his emphasis on charting and plotting data points.

[v] For all unimodal distributions, likely to be seen, using these criteria, the two-sided $\alpha$ -probability is reported to be below 5% (Vysochanskii and Petunin, 1980[v]). Since, primarily with microbiological data, the concern is with an out of control process leading to high values, this would imply that the one-sided $\alpha\alpha$ - probability would be even smaller.

[vi] Johnson, Norman L. and Leone, Fred C. (1964). Statistics and Experimental Design, Vol. 1. John Wiley & Sons, New York.

[vii] Juran, J.M. (1988) Juran's Quality Control Handbook. 4th ed. McGraw-Hill, New York.

[viii] Actually, more generally, X - $\mu$ - k is used where k is a constant that can be chosen to provide operating characteristic desired by the designer.

[ix] It is assumed the process was initially in control with a process mean equal to $\mu$. If the CUSM signaled after a few samples, then this assumption would be questioned.

[x] See for example: http://www.itl.nist.gov/div898/handbook/pmc/section3/pmc323.htm.

[xi] The simulations were run on Statistical Analysis Systems (SAS – release 8.0) using their normal generator: normal(0).

## Appendix 1: Control Charts for Variables Data – classical Shewhart control chart:

When plate counts provide estimates of large levels of organisms, the estimated levels (cfu/ml or cfu/g) can be considered as variables data and the classical control chart procedures can be used. Here it is assumed that the probability of a non-detect is virtually zero. For these types of microbiological data, a log base 10 transformation is used to remove the correlation between means and variances that have been observed often for these types of data and to make the distribution of the output variable used for tracking the process more symmetric than the measured count data[1].

There are several control charts that may be used to control variables type data. Some of these charts are: the Xi and MR, (Individual and moving range) $\overline{X}$ and R, (Average and Range), CUSUM, (Cumulative Sum) and $\overline{X}$ and s, (Average and Standard Deviation).

This example includes the Xi and MR charts. The Xi chart just involves plotting the individual results over time. The MR chart involves a slightly more complicated calculation involving taking the difference between the present sample result, $X_i$ and the previous sample result. $X_{i-1}$. Thus, the points that are plotted are: $MR_i = X_i - X_{i-1}$, for values of i = 2, …, n. These charts were chosen to be shown here because they are easy to construct and are common charts used to monitor processes for which control with respect to levels of microbiological organisms is desired.

**$X_i$ and MR Chart:** The example briefly described here is for $\log_{(10)}$ transformed generic *E. coli*.

Steps required for developing $X_i$ and MR charts are:

1. Define the characteristic …… Generic *E. coli* levels measured from a 25 gram sample, using 3M Petrifilm™
2. Determine sample size (number of samples) …….. 1 for $X_i$ and 2 for MR charts
3. Log(10) transform the data
4. Calculate mean from control data
5. Calculate Moving Ranges
6. Calculate $X_i$ and MR control limits
7. Place control limits on charts with baseline data
8. Plot baseline data and connect consecutive points with a line
9. Place control limits on a blank chart
10. Collect new data
11. Plot new $X_i$ and MR values as they are collected
12. Connect each point to the previous point with a straight line.
13. View both the $X_i$ and MR charts after each point for out of control signals.

After defining the characteristic and deciding that the $X_i$ and MR is the appropriate chart for one's particular situation, baseline data are collected. Baseline data, both $X_i$

values, (Log(10) transformed) and MR values are placed on a baseline control chart prior to calculating control limits, (Figure 1).

After collecting about 30 $X_i$ values and 29 $MR_i$ values, (there is one less MR than $X_i$ since the first MR is not calculated until the second $X_i$ is collected), one can calculate control limits. Before calculating control limits one must first calculate the average moving range, ( $\overline{MR}$ ) and the average of the Xi values, ( $\overline{X}$ ).

$$\overline{MR} = \frac{\sum MR's}{(Number\_of\_Xi) - 1} = \frac{27.11}{29} = 0.93$$

$$\overline{X} = \frac{\sum Xi's}{(Number\_of\_Xi)} = \frac{42.78}{30} = 1.426$$

Note: For computing the standard deviation, σ, needed to establish upper and lower control limits, or other criteria used for evaluating a process, a relationship between $\overline{MR}$ and σ when the results are distributed as a normal distribution. The relationship is simply σ = $\overline{MR}$ /$d_2$, where $d_2$ is a constant = 1.128. Thus, for example, the upper Shewhart control limit is the mean plus 3 times an average $\overline{MR}$ divided by 1.128.

**MR Chart Control Limit Formulae:**

$UCL_{MR}$= $D_4$ x $\overline{MR}$ = 3.267 x 0.93 = 3.04

$LCL_{MR}$= D3 x $\overline{MR}$ = 0 x 0.93 = 0

Center Line, ( $\overline{MR}$ ) = 0.93

Values of $D_4$ and $D_3$, can be found in material on quality control published by professional organizations, for example as given in the endnotes[2] .

**$X_i$ Chart Control Limit Formulae:**

$UCL_{Xi}$ = ( $\overline{X}$ or Target) + (3 x ( $\frac{\overline{MR}}{d_2}$ )) = 1.43 + (3 x ( $\frac{0.93}{1.128}$ )) = 3.90

$LCL_{Xi}$ = ( $\overline{X}$ or Target) - (3 x ( $\frac{\overline{MR}}{d_2}$ )) = 1.43 - (3 x ( $\frac{0.93}{1.128}$ )) = -1.06

Center Line, ( $\overline{X}$ ) = 1.43

(3 x ( $\frac{\overline{MR}}{d_2}$ )) may be replaced with (2.66 x $\overline{MR}$ )

Note: If data are put into a computer, calculator, or spreadsheet, it might be simpler to compute the sample standard deviation, using the "usual" formula, $s = \left[ \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1} \right]^{1/2}$

where n is the number of samples (n = 30). The upper and lower limits are set at ( $\overline{X} \pm$ 3s). However, this method may provide wider limits than those calculated using the first method (Wheeler and Chambers, 1992).

After calculating control limits and the central value (the mean), horizontal lines at these values are placed to the control chart and the baseline data are plotted, $X_i$ and $MR_i$ versus i (Figure 2).

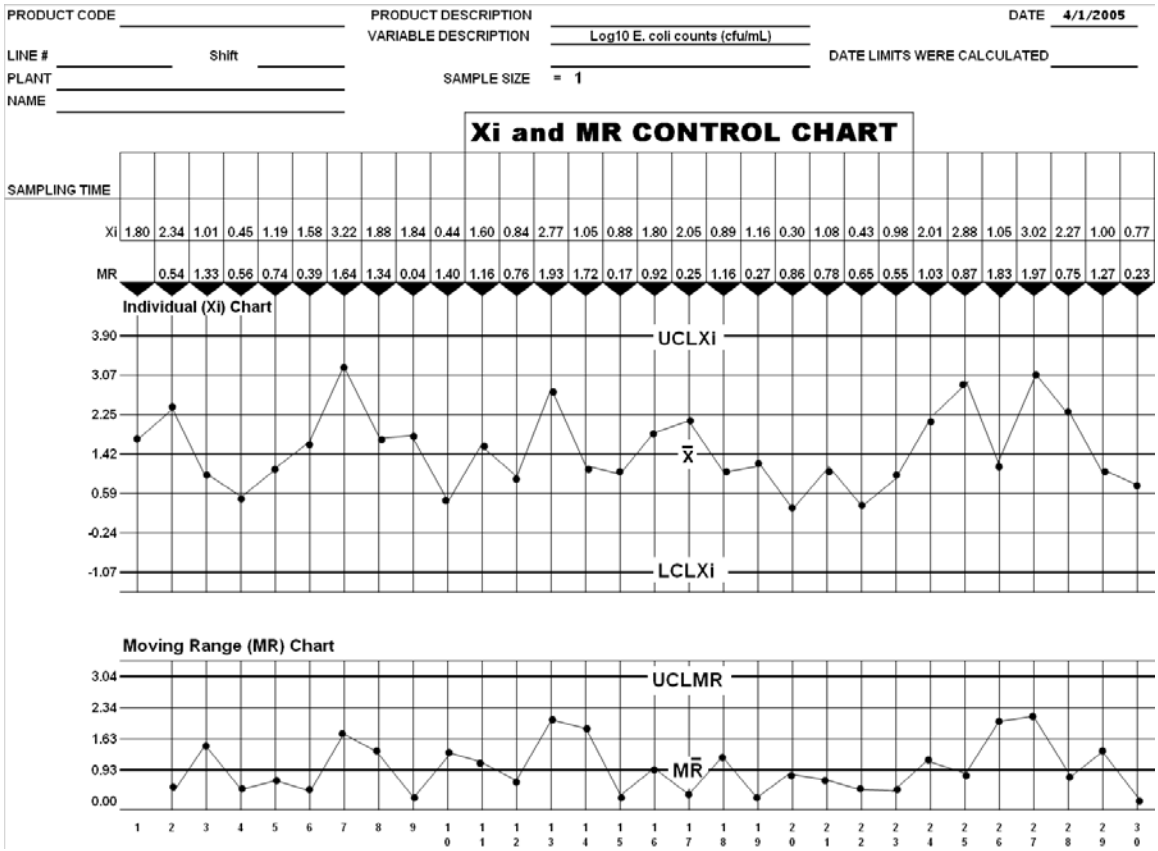



**Figure 1: Baseline data plotted on the control charts with limits derived from the baseline data.**

The baseline data, when plotted, produce a "stable appearing" process. The limits are then transferred to a blank control chart and $X_i$ and $MR_i$ values are plotted as they are collected. After the $X_i$ and $MR_i$ are plotted and connected to the previous point with a straight line, both the $X_i$ and MR charts are viewed for out of control sequences. Pyzdek (1974) suggests the following out of control rules be used:

$X_i$ Chart:

1. Any point exceeding a control limit
2. Eight consecutive points on the same side of the average, ( $\overline{X}$ )

MR Chart:

1. Any point exceeding a control limit
2. Eight consecutive points on the same side of the average, ( $\overline{MR}$ )

Figure 2 demonstrates a process with a positive shift in E. coli counts. At about point number 19 the process showed a positive shift. This was identified after the eighth consecutive point above average on the $X_i$ chart, and confirmed by the out of control point exceeding the $UCL_{Xi}$ on point number 27. Although the process average had shifted up there is no indication that the variation had increased, (Figure 2). If this were the case, then the reason for the out of control pattern would be systemic, affecting the processing within the plant, and would not be from a source which would only affect a portion of the process output, such as a supplier effect. That is, certain possible causes could be eliminated from consideration.

.

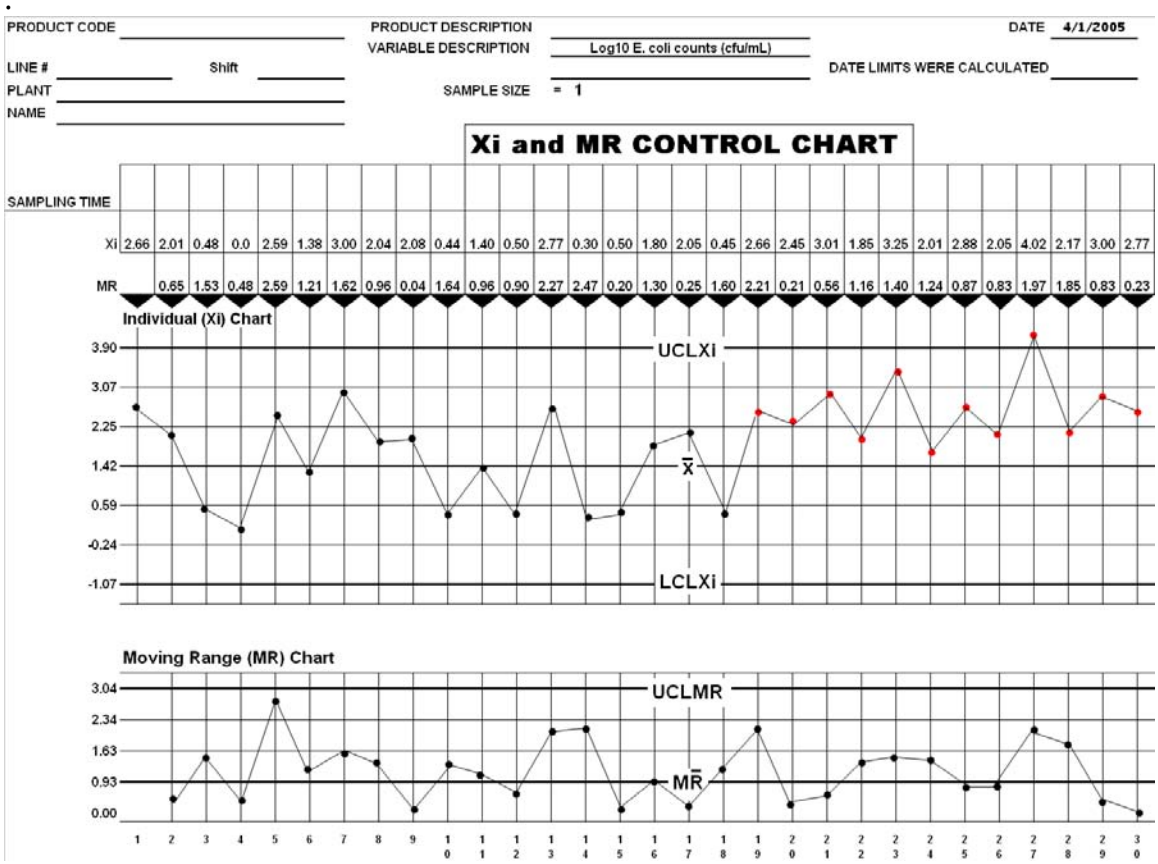



**Figure 2: Log(10) *E. coli* counts illustrating an increase in CFU per ml**

A large "movement" between two consecutive sample values in the $X_i$ chart may cause the MR chart to exceed the $UCL_{MR}$ while both $X_i$ points were within the control limits on the $X_i$ chart. Such a signal may signify a shift in the process mean: a positive shift if the latter result were the large one, for example, (see Figure 3). Or, if a shift in the mean did not occur, then the signal in the MR chart could imply that, while in a systematic way the process is not out of control, there could still be some factor associated with one of the two samples affecting results. Consequently, an investigation of the sources associated

with the two samples might provide a clue of an uncontrolled factor that could be contributing to process variation or could lead to actions that could lead to an improvement of the process. If this were to happen with some regularity, the motivation to investigate would be increased.

If possible a moving range average using more than 2 results might provide a more accurate detection of short term variation. The more terms used though the harder it would be to identify probable causal factors. With computers, it is possible that more than one type of moving range could be computed; for example the two-term moving range, and a 5-term moving range. The ideas presented here remain the same regardless of the number of terms used in the moving range – just different parameters values for $D_4$ and $D_3$ would be used.

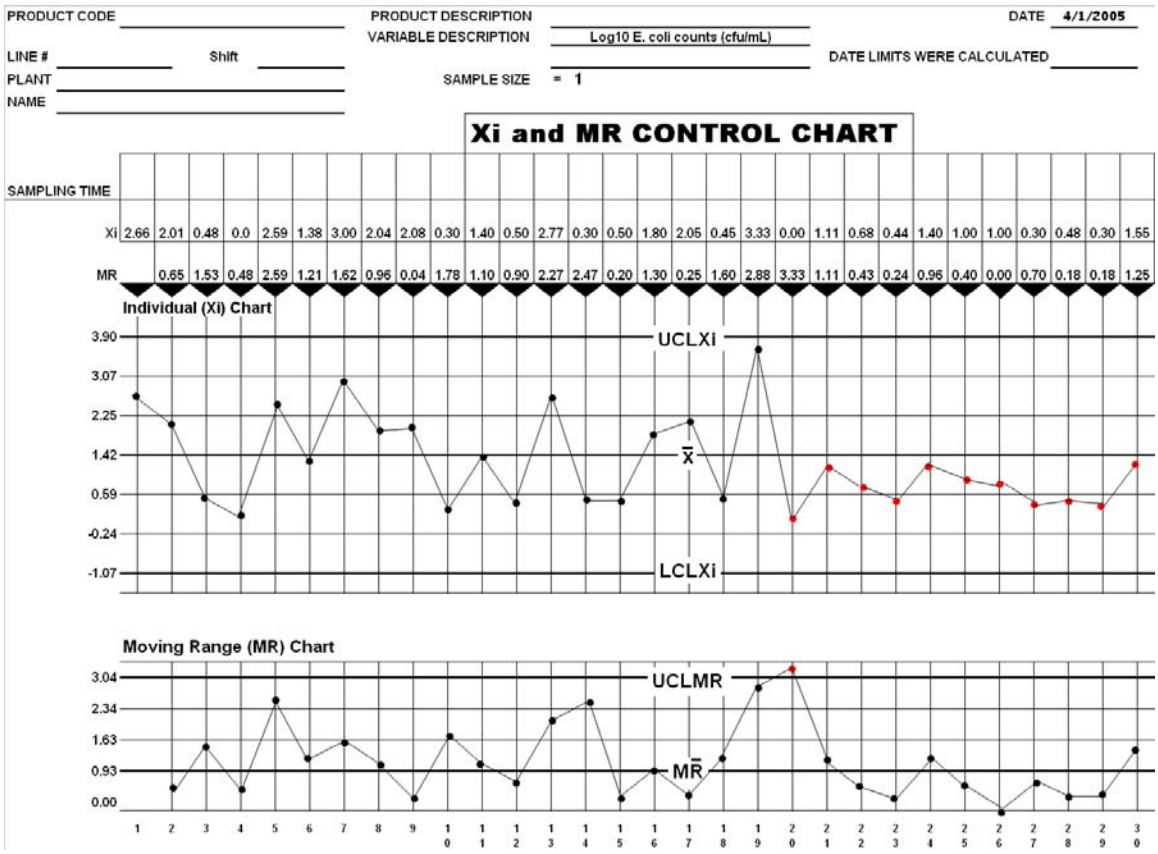



**Figure 3: A down shift in CFUs per ml. first picked up by an out of control signal on the MR chart.**

## Appendix 2: Control Chart for Poisson Distributed Characteristics, with one sample size - the C Chart

When counts are not high and there is a non-trivial probability of not detecting any colony forming units (recorded as ND), the counts seen on a plate can not readily be considered as variable data, as in the previous example (Appendix 1). In this case, a discrete distribution, such as a Poisson distribution or negative binomial distribution can be considered de modeling the distribution of counts (where ND is zero). Microbiological examples which fit Poisson-like probability distributions are not as common as those which lend themselves to the binomial or normal distributions. The Poisson distribution is characterized completely by the value of one parameter, which is the expected value of the distribution. The variance of the Poisson is equal to the expected value, and since the lowest possible value is zero, and there is no limit for the highest values, the distribution is positive skewed. Poisson distributions arise under very specialized conditions, when an assumption of "pure" or simple uniformity is appropriate. However, often this assumption is not appropriate; rather there are many factors that can affect the results, all acting simultaneously so that pure or simple uniformity is not appropriate. Consequently, two parameter distributions such as a negative binomial or even binomial distribution, under certain circumstances can fit data well (Appendix 3). However, the Poisson distribution is an important one, and it some circumstances it might provide a good fit to the data. Thus, this example is being given.

Data for this example were generated using a Poisson distribution, so the Poisson distribution will provide a good fit to the data. A procedure for determining this is given. The example continues with a retrospective analysis, demonstrating one of the features (and possible pitfalls) of such an analysis.

The C chart is used when sample size (number of units or amount of material, being sampled for one analysis) is constant for all samples; the U chart is used for circumstances where sample size may vary. Without loss of generality, it is assumed that the sample size is 1; that is, the direct counts for some material are being recorded.

A word of caution: the Poisson distribution is a skewed distribution, thus $\alpha$- and $\beta$-probabilities need to be calculated taking into consideration when the probability of being above or below the target value is not 50%.

### Control Chart for Poisson distribution with a constant sample size=1

For this example the number of organisms that appear on an aerobic plate count (APC), Petrifilm™, from pre-operational food contact surfaces swabs, (1 square inch, 2.54 cm x 2.54 cm) are counted and expressed as Colony forming units, (CFUs), per square inch, (sampling area). For this example the area is swabbed with moistened cotton tipped swab. The swab is used to plate the results on Petrifilm™. The number of colonies is counted after a 48 hour incubation period.

The data from the one hundred swabs are given in Table 1. (Note: these data are not actual data, but were generated using a Poisson distribution). The table includes the observed frequency of results, the predicted frequency assuming a Poisson distribution estimated from the data using the maximum likelihood estimate (MLE)[3]; the likelihood ratio contribution for the observed result, the likelihood ratio contribution for the observed result when combining the results greater than 6 into one category, and the chi-square statistic for the observed result, which is the square of the difference between the observed and expected frequencies divided by the expected frequencies, $(f-e)^2/e$, where f is the observed frequency, e is the expected or predicted frequency derived from, using the estimated Poisson distribution. The likelihood ratio contribution is minus twice the product of the observed frequency and the difference of the natural logarithms of the observed and expected frequencies, or, symbolically:

$$-2f[\ln(f) - \ln(e)],$$

where "ln()" is the natural logarithm.

. The sum of the chi-square and likelihood-ratio contributions are statistics that are asymptotically distributed as a chi-square with k-1 degrees of freedom, where k is the number of distinct results (or categories) for which estimates are made. The likelihood ratio when combining results greater than 6 has 8 categories so that the chi-square approximation is based on 7 degrees of freedom. The results do not indicate any severe lack of fit, notwithstanding the 13 negative results that were observed when only 9 were predicted, and the large observed value of 11, which would not be predicted to be seen very often.

**Table 1: Results from the 100 preoperational swab, counts, frequency of results, and predicted frequency using maximum likelihood estimate for Poisson distribution**

| count | observed frequency | predicted frequency | likelihood ratio | likelihood ratio combined 7 df | chi square 7 df |
|---|---|---|---|---|---|
| 0 | 13 | 9.1 | 9.35 | 9.35 | 1.701 |
| 1 | 20 | 21.8 | -3.40 | -3.40 | 0.144 |
| 2 | 27 | 26.1 | 1.78 | 1.78 | 0.029 |
| 3 | 19 | 20.9 | -3.62 | -3.62 | 0.173 |
| 4 | 10 | 12.5 | -4.53 | -4.53 | 0.515 |
| 5 | 6 | 6.0 | -0.04 | -0.04 | 0.000 |
| 6 | 2 | 2.4 | -0.74 | -0.74 | 0.069 |
| 7 | 1 | 0.8 | 0.38 | 5.70 | 1.129 |
| 8 | 0 | 0.2 | . | . | . |
| 9 | 1 | 0.1 | 5.43 | . | . |
| 11 | 1 | 0.0 | 11.33 | . | . . |
| sum | 100 | 100 | 16.0 | 4.50 | 3.76 |

Accepting that the underlying distribution of results is a Poisson distribution with expected value of 2.4, the steps involved in using a C control chart are:

1. Define the characteristic… APC counts per sampling location
2. Determine sample size …. 1 square inch area, 1 sampling area
3. Collect baseline data
4. Calculate Control Limits
5. Place Control Limits on chart of baseline data
6. Plot the baseline data
7. Connect consecutive plotted points with a straight line
8. Place control limits on a new chart
9. Collect and plot data as collected
10. Connect each point to previous point with a straight line
11. Observe chart for out of control signals after each point

The formula for C Center line and control limits are:

Average Count = $\overline{C} = \frac{Total\_Number\_of\_CFUs}{Number\_of\_Samples}$

where: Sample_size = 1 and Number_of_Samples = k = 100 (in this example).

$\overline{C} = \frac{240}{100} = 2.40$

Center Line = $\overline{C} = 2.40$

Control Limit Calculations:

Standard Deviation: $\sigma = \sqrt{\overline{C}}$ - that is, it is assumed that the distribution is a Poisson distribution.

Upper Control Limit C:

$UCL_c = \overline{C} + \left(3 \times \left(\sqrt{\overline{C}}\right)\right)$

$UCL_c = 2.40 + \left(3 \times \left(\sqrt{2.40}\right)\right) = 7.05$

Lower Control Limit C:

$LCL_c = \overline{C} - \left(3 \times \left(\sqrt{\overline{C}}\right)\right)$

$LCL_c = 2.40 - \left(3 \times \left(\sqrt{2.40}\right)\right) = -2.25$ so $LCL_c = 0$

For the upper limit, if the underlying distribution was a Poisson with expected value equal to 2.40, then the probability of a result greater than 6 is 1.16%; and greater than 7 is 0.334%. This depending upon the $\alpha$ -probability desired, either a result greater than or equal to 7 or 8 would be considered as presumptive evidence of an out of control process.

For the lower limit, the probability of 0 is 9.1%, which for the $\alpha$ -probability (in this document) would be considered as too high. Thus, a single value of 0 would not, by itself be considered as an indication of a process change.

If an actual Poisson distribution was not assumed or any other distribution could not be found to fit the data, then the square root transformation, or its variations $(x+3/8)^{1/2}$, could be used for plotting and construction a control chart.

In this example, the SPC chart of Figure 1 is constructed, with the upper limit of 7 and the lower limit of 0, derived assuming that the results are distributed as a Poisson distribution. The last 40 points are shown on the chart.

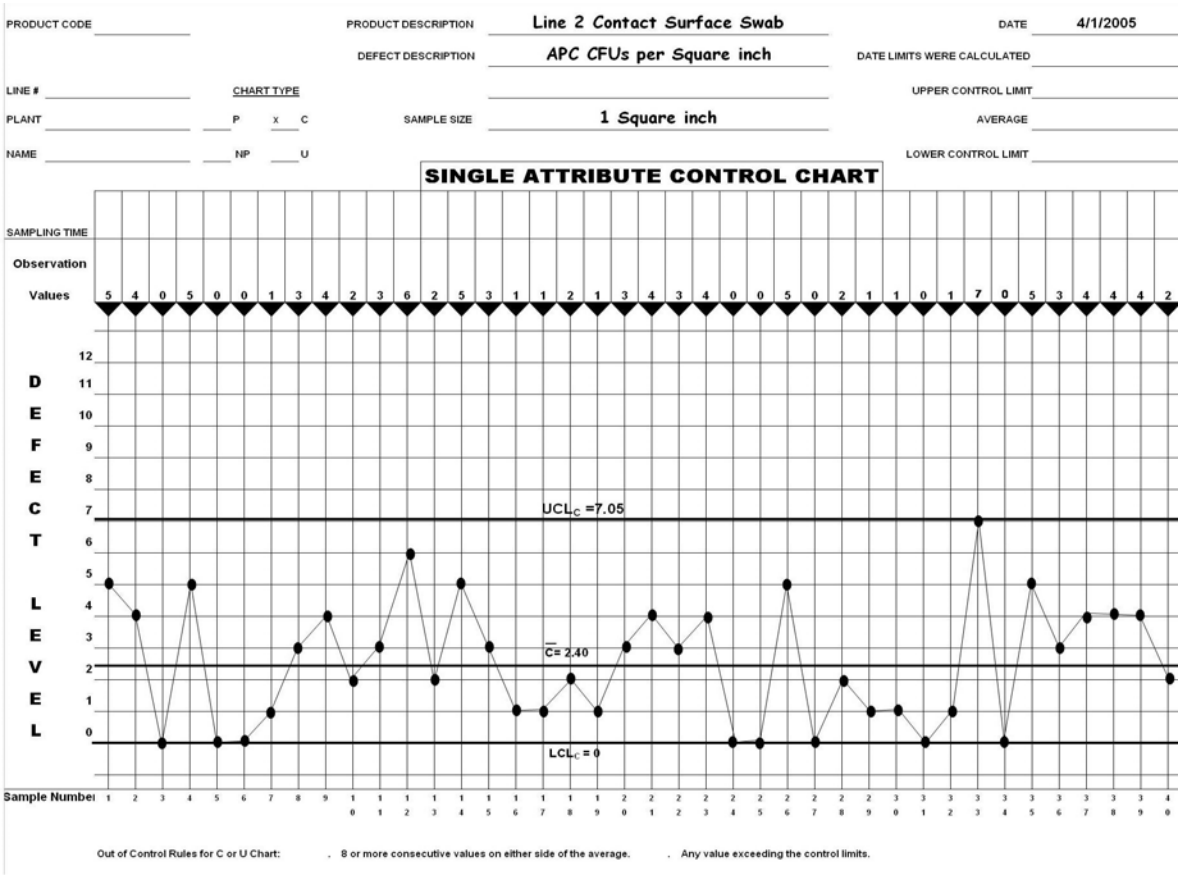



**Figure 1: A C chart showing the last 40 data points of the baseline data plotted**

Upon an examination of this chart, out of the 40 points, only one had a value of 7 (or more), and there were 8 zero results. Moreover there does not seem to be a consistent trend (in actuality CUSUMS or moving averages would be computed as well to judge trends). By these usual rules, it might be reasonable to conclude that the 40 points plotted represent a set of data for which the process is in control. Only one result reached the value of 7, but that, by itself, would not be a reason to suspect an out-of control situation.

However, a further examination might lead to some questions and further exploration of the process. The probability of 8 or more zero results from 40, when the underlying distribution is Poisson is about 2.5%; however, what may be of interest is that the 8 values seemed to be "clustered" with respect to time, where 3 zero results occurred within the first 6 times of sampling, and the other 5 zero results occurred within a span of 11 samples 18 samples later.

To explore the possibility that this pattern represents a possible source of unexplained variation, consider a one-sided CUSUM for the occurrence of negative results. That is, consider a CUSUM, $S_k$, where, $S_0 = 0$, $S_k = \max(0, S_{k-1} + \delta_k - p)$, $\delta_k$ is the Kronecker delta function for a negative result for the $k^{th}$ sample ( = 1 when the result is negative, otherwise equal to 0), and p is a constant equal to the probability of a negative result ( = 0.091). Let the signal for "out of control" – meaning that the probability of a negative, somewhere, was greater than 9.1% - be 5. That is, when the CUSUM value is equal to or greater than 5 the CUSUM signals and the null hypothesis that the probability of a negative was equal to or less than 9.1% over the 40 samples would be rejected in favor of the alternative hypothesis that for some time the actual probability was greater than 9.1%. The reason for selecting the signal limit of 5 is: the ARL for this CUSUM with a signal limit of 5 is about 346 when the underlying probability of a negative result is 9.1%, as predicated from the assumed Poisson distribution with parameter value equal to 2.4. Thus, if this CUSUM rule had been constructed before the samples were being collected, the one sided $\alpha$ -probability would have an assigned value of 0.29%.

Calculating this CUSUM with these 40 samples, at the $34^{th}$ sample, the CUSUM exceeded 5.0 so that a signal would have occurred. Table 2 provides an example of a spread sheet that can be used for the calculations of CUSUM. The above formula for the CUSUM implies that the CUSUM value at the $k^{th}$ sample depends on the CUSUM value for the $(k-1)^{th}$ sample and an increment value, which is the value of the $k^{th}$ sample minus the target constant, $\mu$. The updated value of the CUSUM for the $k^{th}$ sample is the sum of the previous CUSUM value plus the increment, provided this sum is not less than zero; otherwise it is set equal to zero. Thus the formula for the CUSUM can be written as:

$$CUSUM_k = \max(0, CUSUM_{k-1} + increment_k)$$

where the value of the increment = $\delta_k$ - p. Table 2 presents the calculations of the CUSUM for the 40 samples using the above formula.

As can be seen from Table 2, the CUSUM value exceeds the value of 5 at the $34^{th}$ sample, thereby suggesting that the process mean probability of a negative exceeded its target value of 9.1% at least some time at or before the $34^{th}$ sample. However, the probability of a Type 1 error with respect to the statistical question, given the 40 data points, is not the same as the assigned $\alpha$-error probability given above (based on the ARL). The statistical question involves deciding between two hypotheses: the null hypotheses, $H_0$: the probability of negative results is never larger than 9.1% over the period of time that the samples were collected; versus the alternative, $H_A$, that at some time, the probability exceeded 9.1%. To help evaluate possible answers, the probability of seeing a signal

within 40 samples is needed, when the probability of a negative is equal to 9.1% over the 40 samples, thereby assuming that the null hypothesis is true. The 2nd percentile of the distribution of run lengths (estimated by simulation using the binomial random generator for SAS®, release 8.0) is about 40 samples, indicating that the retrospective $\alpha$ -probability of a Type 1 error is about 2% (that is, of having a signal within the first 40 samples when the probability was 9.1% of a negative result).

**Table 2: Calculation spreadsheet for CUSM.**

| sample number | sample result | target fraction positive | increment (result-target) | CUSUM |
|---|---|---|---|---|
| 1 | 0 | 0.09 | -0.09 | 0.00 |
| 2 | 0 | 0.09 | -0.09 | 0.00 |
| 3 | 1 | 0.09 | 0.91 | 0.91 |
| 4 | 0 | 0.09 | -0.09 | 0.82 |
| 5 | 1 | 0.09 | 0.91 | 1.73 |
| 6 | 1 | 0.09 | 0.91 | 2.64 |
| 7 | 0 | 0.09 | -0.09 | 2.55 |
| 8 | 0 | 0.09 | -0.09 | 2.46 |
| 9 | 0 | 0.09 | -0.09 | 2.36 |
| 10 | 0 | 0.09 | -0.09 | 2.27 |
| 11 | 0 | 0.09 | -0.09 | 2.18 |
| 12 | 0 | 0.09 | -0.09 | 2.09 |
| 13 | 0 | 0.09 | -0.09 | 2.00 |
| 14 | 0 | 0.09 | -0.09 | 1.91 |
| 15 | 0 | 0.09 | -0.09 | 1.82 |
| 16 | 0 | 0.09 | -0.09 | 1.73 |
| 17 | 0 | 0.09 | -0.09 | 1.64 |
| 18 | 0 | 0.09 | -0.09 | 1.55 |
| 19 | 0 | 0.09 | -0.09 | 1.46 |
| 20 | 0 | 0.09 | -0.09 | 1.37 |
| 21 | 0 | 0.09 | -0.09 | 1.28 |
| 22 | 0 | 0.09 | -0.09 | 1.18 |
| 23 | 0 | 0.09 | -0.09 | 1.09 |
| 24 | 1 | 0.09 | 0.91 | 2.00 |
| 25 | 1 | 0.09 | 0.91 | 2.91 |
| 26 | 0 | 0.09 | -0.09 | 2.82 |
| 27 | 1 | 0.09 | 0.91 | 3.73 |
| 28 | 0 | 0.09 | -0.09 | 3.64 |
| 29 | 0 | 0.09 | -0.09 | 3.55 |
| 30 | 0 | 0.09 | -0.09 | 3.46 |
| 31 | 1 | 0.09 | 0.91 | 4.37 |
| 32 | 0 | 0.09 | -0.09 | 4.28 |
| 33 | 0 | 0.09 | -0.09 | 4.19 |
| 34 | 1 | 0.09 | 0.91 | 5.09 |
| 35 | 0 | 0.09 | -0.09 | 5.00 |
| 36 | 0 | 0.09 | -0.09 | 4.91 |
| 37 | 0 | 0.09 | -0.09 | 4.82 |
| 38 | 0 | 0.09 | -0.09 | 4.73 |
| 39 | 0 | 0.09 | -0.09 | 4.64 |
| 40 | 0 | 0.09 | -0.09 | 4.55 |

Retrospective analyses such as this one are fraught with problems regarding the "true" magnitude of the $\alpha$ - and $\beta$- probabilities. The only reason that this CUSUM was designed, after the fact, was because of the pattern of negative results that were observed and the higher than expected frequency of them in the 40 samples. However, any truly "random" sequence of numbers could turn up patterns that are suggestive of possible non-randomness suggesting a more complex generator (of the numbers), which, when statistically tested for, would result in a low calculated $\alpha$ -probability. Notwithstanding these types of problems, the results of this analysis might suggest or provide evidence of the existence of a factor that is causing excess process variation, and that further examination of the process would be worthwhile; for example, the producer might explore to see if there were any common sources peculiar to the initial 6 and the later 11 data points. If there were, then particular sharper criteria or rules related to the possible common sources could be constructed; if there were not, then the observed pattern could be considered as arising due to statistical variation.

**Appendix 3: Count Data that is not Poisson distributed, with many non-detect values**.

This example presents data that do not seem to have arisen from a Poisson distribution.

**Table 1: Obtained counts, together with goodness-of-fit statistics for a Poisson distribution.**

| count | observed frequency | predicted frequency | likelihood ratio | likelihood ratio combined 7 df | chi square 7 df |
|---|---|---|---|---|---|
| 0 | 15 | 4.9 | 33.39 | 33.39 | 20.576 |
| 1 | 19 | 14.8 | 9.40 | 9.40 | 1.168 |
| 2 | 22 | 22.3 | -0.65 | -0.65 | 0.005 |
| 3 | 15 | 22.4 | -12.0 | -12.0 | 2.447 |
| 4 | 10 | 16.9 | -10.4 | -10.4 | 2.790 |
| 5 | 6 | 10.1 | -6.31 | -6.31 | 1.696 |
| 6 | 4 | 5.1 | -1.93 | -1.93 | 0.234 |
| 7 | 2 | 2.2 | -0.36 | 13.68 | 2.643 |
| 8 | 0 | 0.8 | . | . | . |
| 9 | 1 | 0.3 | 2.58 | . | . |
| 10 | 3 | 0.1 | 21.53 | . | . |
| 11 | 1 | 0.0 | 7.57 | . | . |
| 15 | 1 | 0.0 | 19.55 | . | . |
| 20 | 1 | 0.0 | 37.40 | . | . . |
| Sum | 100 | 100 | 99.68 | 25.10 | 31.56 |

With 7 degrees of freedom, the likelihood ratio test and the chi-square test for lack of fit are significant, with significance levels less than 0.01, suggesting that the fitted Poisson distribution does not fit the data well.

A two parameter distribution, the negative binomial is a distribution which is often used to model the distribution generating the data when the Poisson does not provide a good fit. The probability density, f(k|p, n) of the negative binomial is:

$$f(k \mid p, n) = K(n, k)p^{n}(1-p)^{k}, \quad k = 0,1,....$$

where n and p are parameters whose values are to be estimated and K(n, k) is a binomial-like coefficient. The above data were fit to a negative binomial, using a maximum likelihood estimates (MLE) of the parameter values. The MLE of p and n were 0.3655 1.7342, respectively. Table 2 provides summary statistics for the MLE fitted negative binomial. The likelihood ratio with 8 degrees of freedom was obtained by pooling all the results greater than 7 into one category.

**Table 2: Obtained counts, together with goodness-of-fit statistics for a negative binomial distribution**

| count | observed frequency | predicted frequency | likelihood ratio | likelihood ratio combined 8 df | chi square 8 df |
|---|---|---|---|---|---|
| 0 | 15 | 17.5 | -4.56 | -4.56 | 0.347 |
| 1 | 19 | 19.2 | -0.42 | -0.42 | 0.002 |
| 2 | 22 | 16.7 | 12.23 | 12.23 | 1.709 |
| 3 | 15 | 13.2 | 3.93 | 3.93 | 0.257 |
| 4 | 10 | 9.9 | 0.24 | 0.24 | 0.001 |
| 5 | 6 | 7.2 | -2.17 | -2.17 | 0.197 |
| 6 | 4 | 5.1 | -1.97 | -1.97 | 0.245 |
| 7 | 2 | 3.6 | -2.34 | -2.34 | 0.704 |
| 8 | 0 | 2.5 | . | -1.38 | 0.075 |
| 9 | 1 | 1.7 | -1.07 | . | . |
| 10 | 3 | 1.2 | 5.69 | . | . |
| 11 | 1 | 0.8 | 0.48 | . | . |
| 15 | 1 | 0.2 | 3.69 | . | . |
| 20 | 1 | 0.0 | 7.84 | . | . . |
| sum | 100 | 100[a] | 21.57 | 3.55 | 3.54 |

[a]) The sum includes predicted numbers for counts not shown, for example, a count of 12, up to a count of 20.

From Table 2 it appears that the fitted negative binomial distribution fits the data well. Assuming that the distribution generating the count data is the estimated negative binomial distribution, the probability that an individual count would be greater than or equal to 18 is 0.133%. Thus, the individual limit, corresponding to the Shewhart limit of 3 standard deviation units above the mean for a normal distribution, would be 18, using the negative binomial distribution.

The assumption that these data were collected for a process under control is important here. The 15 non-detects may suggest a measurement problem, insofar as the number of these seems high compared to what might have been expected if it were believed the distribution of counts would be Poisson distributed. This might be one area of further, retrospective, exploration. Many other distributions could be fit to these data directly, for example Poisson with added zeros, or other types of distributions.

As suggested in this document, another possible way of constructing a SPC plan and chart is to consider transformations of the data in an attempt to make the data more symmetric and nearly normal. Figure 1 is a comparison of the box-plots, of the square root transformed counts with the raw counts. The square root transformed counts are multiplied by 2, so that the means of the two sets of numbers are nearly the same. As is clearly seen, the square root transformed results provide a more symmetric distribution than that of the raw counts. The means and standard deviations for the raw counts and twice the square root of the counts, and the control limits derived from them are given in Table 3.

**Table 3: Means, standard deviations and control limits when using raw counts and square root transformed counts.**

| type | mean | std dev | limit: 3 std dev above mean | limit using square root |
|---|---|---|---|---|
| raw counts | 3.01 | 3.20 | 12.62 | . |
| twice square root | 2.97 | 1.79 | 8.36 | 17.47 |

The Shewhart limit of 3 standard deviations above the mean using the square root transformation is 18 (rounding up from the 17.47 given in Table 3), the same as that derived using the fitted negative binomial. From the 100 raw counts, only 1 was above 18.

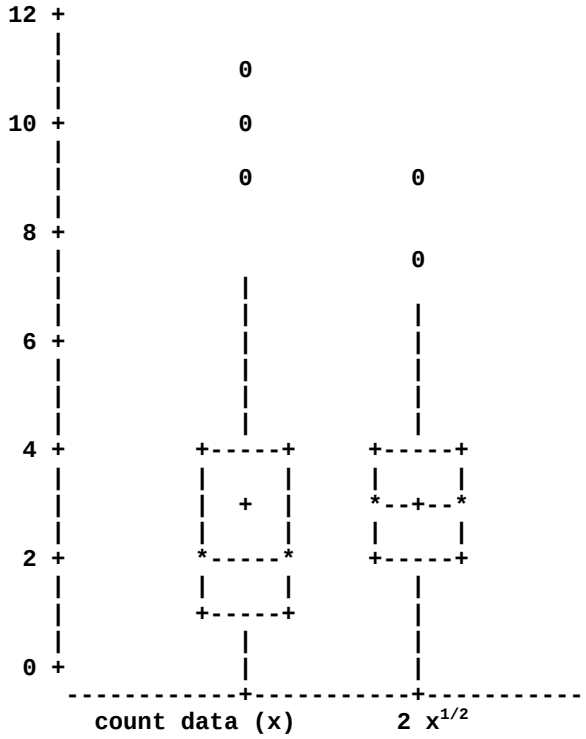



**Figure 1: Box-plots of counts and twice the square root of the counts. The high values of 15 and 20 for the count data are not shown.**

## Appendix 4 - Control Chart for Binomially Distributed Data, with one sample size

This example is a very important one since often qualitative analyses, looking just for the presence of some pathogen in samples, are performed. A chart that can be used to track the control of a process with respect to the presence of some pathogen on samples is called a "NP" - control chart. Generally such charts can be used for a binomially - like distributed characteristics (a two-class attribute test), for example, the detecting of *Salmonella* spp on samples. One of the classifications is assigned the name "defective" or positive, and it is that classification for which process control is measured. P refers to the percentage or probability of "defective" units (positive units of some product); the magnitude of P is to be controlled (usually to be low as possible).

The letters "NP" are used as a mnemonic for the plotting of the number of "positive results"; the expected value of the number of positive results is equal to the sample size, N, times the assumed proportion of positive samples, P – or, symbolically, NP.

In this example, it is assumed a sample size of 50 product tests, constituting one sample, for which the number of positive results is the output. A NP-chart is a plot of the number of "positive" test results within a sample over time. The example given in Appendix 5 provides methodology that can be used when the sample sizes are not the same (using a P-chart or a transformation of the results).

### Control Chart for Binomially Distributed Data Plotted as the Number of Positive Outcomes from an Inspection

For the NP - control chart, "N" indicates sample size, (often an upper case N is used to symbolize population size, but it is the SPC convention to use an upper case N which stands for sample size – the number of units being considered together as one sample), and P represents the proportion of the units that are "defective," as described above. The set of N units is referred to as a "sample", so that the first N-unit set is labeled sample 1; the second N-unit set is labeled sample 2, and so forth. An NP-chart is simply a plot of $X_i$ = the number of defective units in the $i^{th}$ sample, versus sample index value (or some other appropriate time measure), with lines connected between successive data points. The steps involved in determining the control limits for a NP-control are:

1. Define the characteristic… Presence of *Salmonella* spp. in 25 grams of product
2. Determine sample size….. Sample size = 50, 25-gram units
3. Collect baseline data
4. Calculate Control Limits
5. Place Control Limits on chart of baseline data
6. Plot the baseline data
7. Connect consecutive plotted points with a straight line
8. Place control limits on a new chart
9. Collect and plot data as collected
10. Connect each point to previous point with a straight line
11. Observe chart for out of control signals after each point

For this example, Figure 1 is the number of *Salmonella* spp. positive units identified in a sample of 50 units of product is plotted versus sample number (or time of sampling) on the NP control chart. After about 30 data points have been collected, (a rule of thumb for normal or nearly normal data is that about 30 data points should used for control limit calculations) the control limits are ready to be calculated.

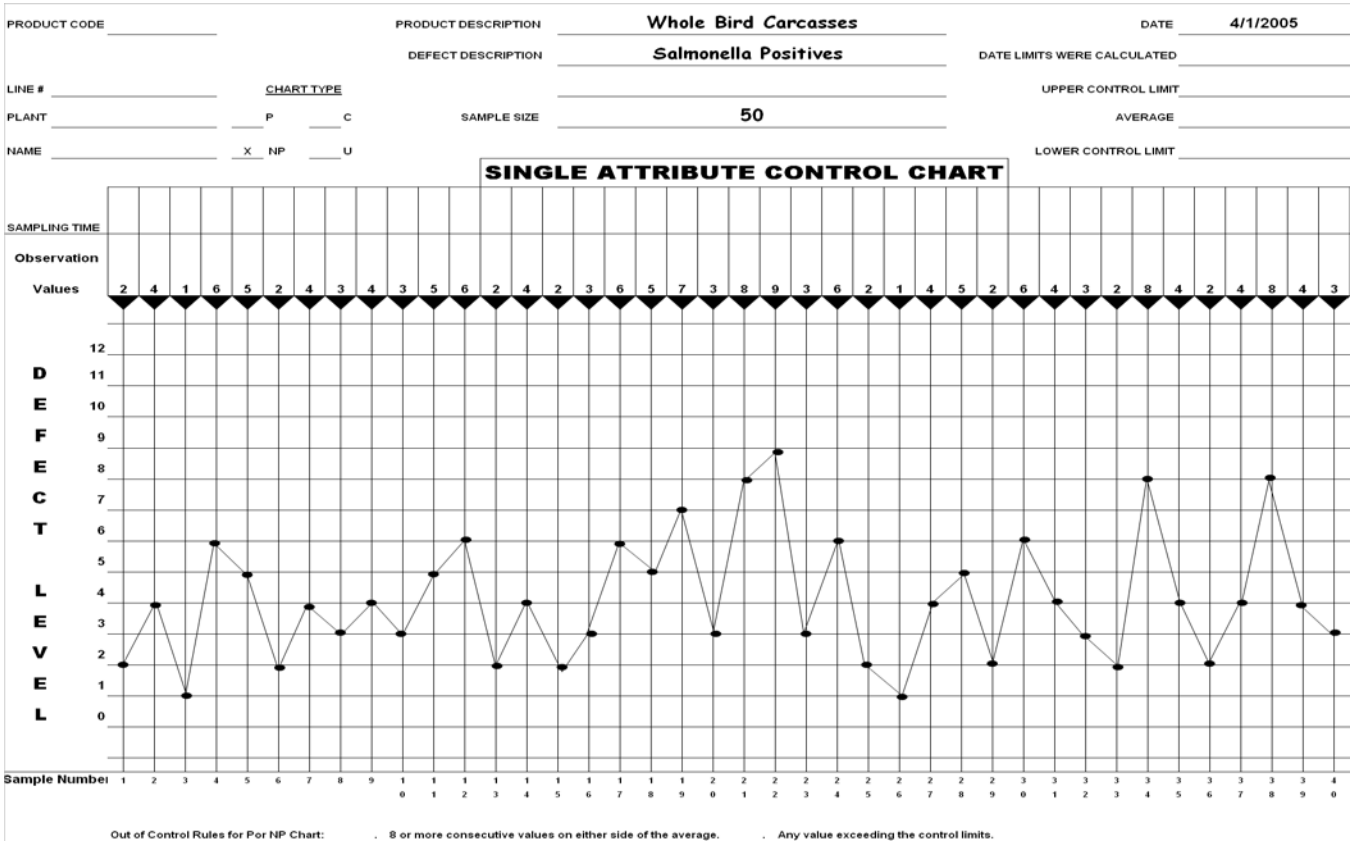



**Figure 1. Base line *Salmonella* spp. data collected in sample size = 50.**

The formula for NP Center line and control limits are:

Average Proportion Positive = $\overline{P} = \dfrac{\sum_{i=1}^{k} Xi}{Sample\_size \times Number\_of\_Samples}$

where Sample_size = N = 50; and Number_of_Samples = k = 40 (in this example).

$$\overline{P} = \frac{165}{50 \times 40} = 0.0825 = 8.25\%$$

Center Line = $N\overline{P}$ = Sample Size x $\overline{P}$ = 50 x 0.0825 = 4.125
(This gives the expected number of positive results per sample).

Control Limit Calculations:

Standard Deviation: $\sigma = \sqrt{N\bar{P} \times (1-\bar{P})}$ - that is, it is assumed that distribution is a binomial distribution. In actuality this may not be true even when the process is in control because of inherent intra-sample correlations that would cause the expected value of the number of positive results to vary by sample.

Note: An estimate of the standard deviation can be computed as: $\sigma = \sqrt{\frac{N\sum_{i-1}^{k}(P_i - \bar{P})^2}{k-1}}$,

where k is the number of samples (= 40) and $P_i$ is the fraction of positive results (of the N analyses) for the $i^{th}$ sample.

In this example, it is assumed that "the best" control is achieved so that the deviations from the expected value follow a binomial distribution, implying that the standard deviation is proportional $(p(1-p))^{1/2}$, where p is the expected percentage of positive units. If this assumption is incorrect and that the expected value of the probability of defects on a unit changes from day to day or sample to sample, a condition known as over-dispersion may exist. In this case, the standard formula for standard deviation, or the MR statistic discussed in Appendix 1, could be used, for the number of positive results, or for the arc-sine transformation: $y_i = N \sin^{-1}(P_i^{1/2})$. Evidence of this condition may be identified by plotting the baseline data on the chart with control limits calculated in the manner shown and observing many point either "out of control" or at least near the extremes, but that the deviations seem random and symmetric. More formal statistical tests for "over-dispersion" can be made using statistical programs such as PROC GENMOD of SAS[4]. Such a pattern might arise, when there are uncontrollable factors, such as day-to-day variations attributable to environment or slight, but uncontrollable differences, of supply input quality. In this case, the process standard deviation can be computed using the above standard formula. However, if this case does exist, the producing establishment should strive to eliminate some of these sources that contribute to the variability of the process, particularly when the establishment determines that the characteristic should be binomially distributed. Obtaining better control and eliminating factors that cause over dispersion usually leads to improvements, (reduction in percentage of defective units in this example) which means new control limits will need to be calculated once the improvement is documented.

In the following it is assumed a binomial distribution describes the number of positive results.

Upper Control Limit NP:

$UCL_{NP} = N\bar{P} + \left(3 \times \left(\sqrt{N\bar{P} \times (1-\bar{P})}\right)\right)$

$1-\bar{P} = 1 - 0.0825 = 0.9175$

$UCL_{NP} = 4.125 + \left(3 \times \left(\sqrt{4.125 \times 0.9175}\right)\right) = 9.96$

Lower Control Limit NP:

$LCL_{NP} = N\overline{P} - \left(3 \times \left(\sqrt{N\overline{P} \times (1 - \overline{P})}\right)\right)$

$LCL_{NP} = 4.125 - \left(3 \times \left(\sqrt{4.125 \times 0.9175}\right)\right) = -1.71 = 0$

From the above data, the average proportion of *Salmonella* spp. positive was 0.0825. If the expected proportion of positive results for each 50-set sample was 8.25%, then the average or expected number of positive results in a 50-set sample is 4.125, ($N\overline{P} = 4.125$), and the control limits would be 9.96 and 0 for the $UCL_{NP}$ and $LCL_{NP}$, respectively. A value of 10 then would be considered as out of control. (The actual probability of 10 or more positive units in 50 units, when the probability of a positive is 8.25 % and the true underlying distribution is binomial, is 0.70%, which provides a reasonably low α-error rate.) Since the calculated value of $LCL_{NP}$ is a negative value and since one can never find fewer than zero positives in a sample, the $LCL_{NP}$ defaults to zero. (A result of zero may not be considered by itself to be an out of control event, since the probability of no positive results from 50 samples, given a true percentage of 8.25% and a binomial distribution, is 1.35 %, which is substantially larger than the nominal value of α of 0.135% used by Shewhart, when the underlying distribution is normal). By placing the control limits and center line on the baseline data run chart a control chart is produced (Figure 2).

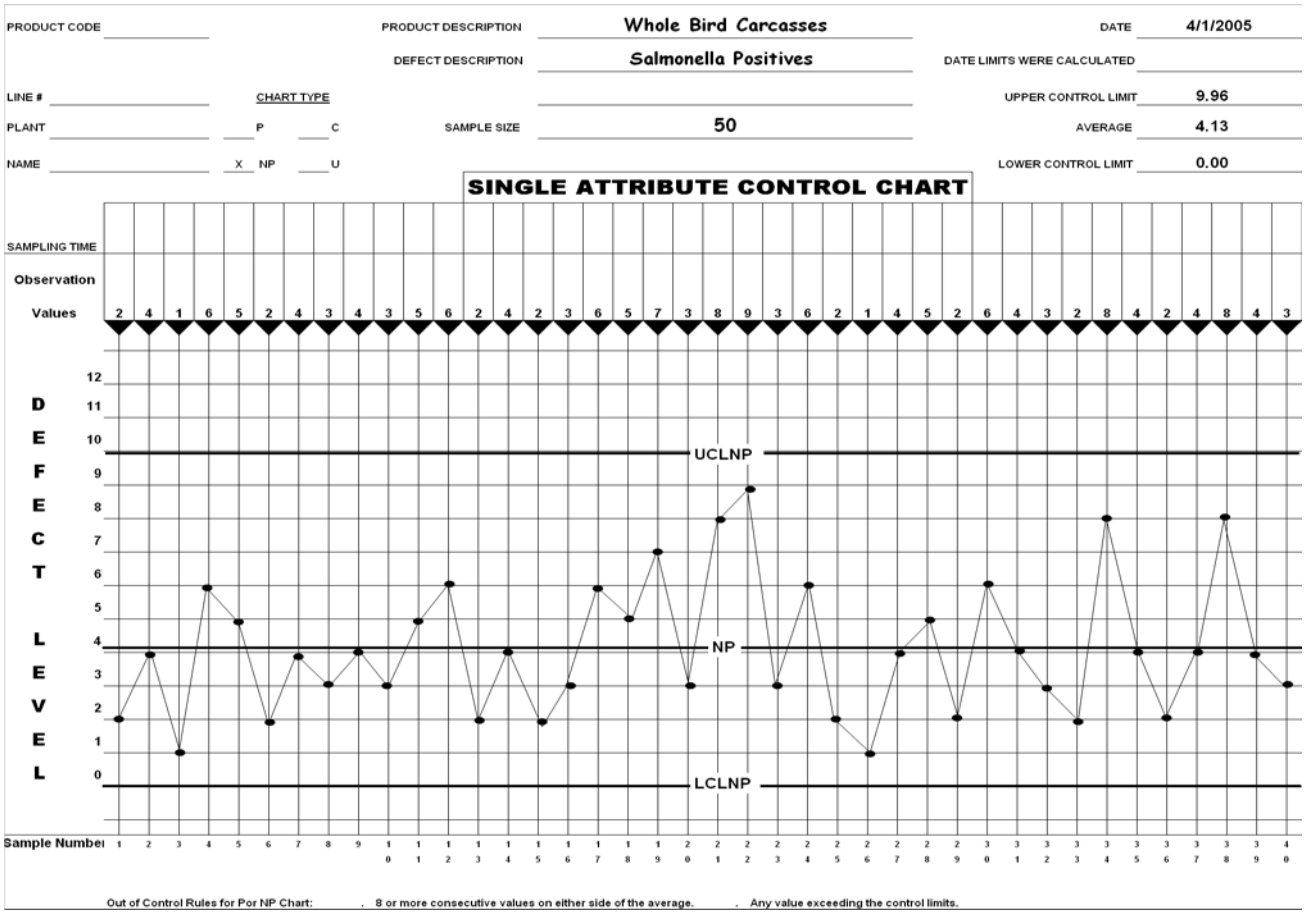



**Figure 2. NP control chart of baseline data**

The control limits are calculated and the plot of the baseline data confirms the control limits appear reasonable, as in Figure 2. The control limits are then transferred to a blank control chart, or extended from the control chart containing the baseline data, and data points are plotted and connect to the previous plotted point with a straight line. Figures 3 and 4 demonstrate two control charts that show a period of expected performance followed by an out of control period. Figure 3 shows a process with the incidence of *Salmonella* spp. increasing, while Figure 4 shows a reduction in *Salmonella* spp. Both of the charts thus show out of control conditions, because both of these red point series are 'unexpected' from what would be seen, based on the baseline data and the assumption that the process was in control.

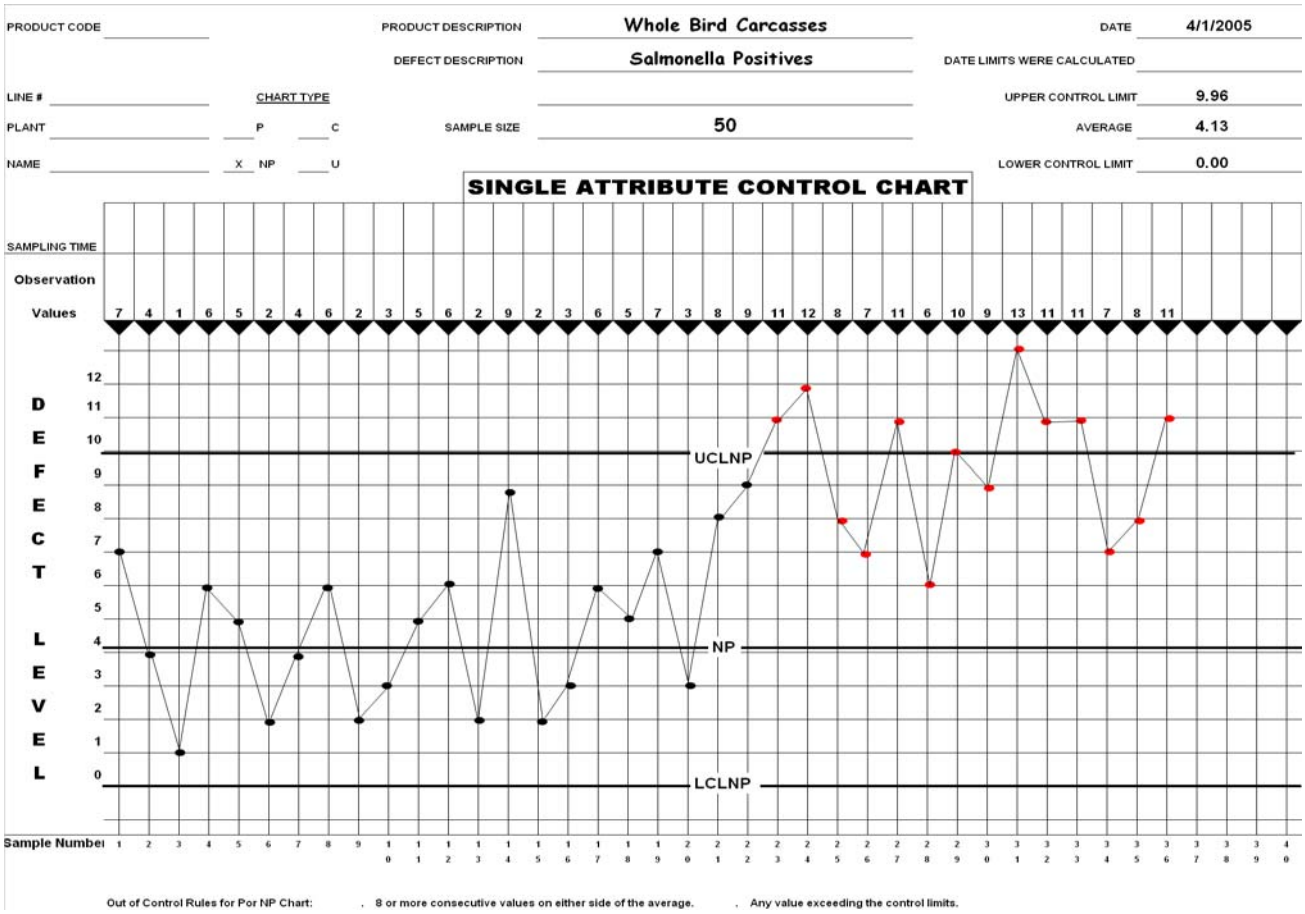



**Figure 3. NP chart showing an increase in *Salmonella* spp.**

The identification of the out of control signal suggests an investigation as to what caused the out of control condition. It is through the identification and control of factors that adversely affect the process that leads to process improvement. Once these are controlled often a new level of quality is achieved. A new level of quality in this case can be identified by an "out of control" condition where an unexpectedly high number of consecutive points fall below average (Figure 4). The informed manufacturer at this point would identify the

conditions required to produce these better results and after a brief time calculate new control limits based on the new level of quality that the process is producing. There are methods described for computing new control limits, see Wheeler and Chamber, 1992[5], for examples.

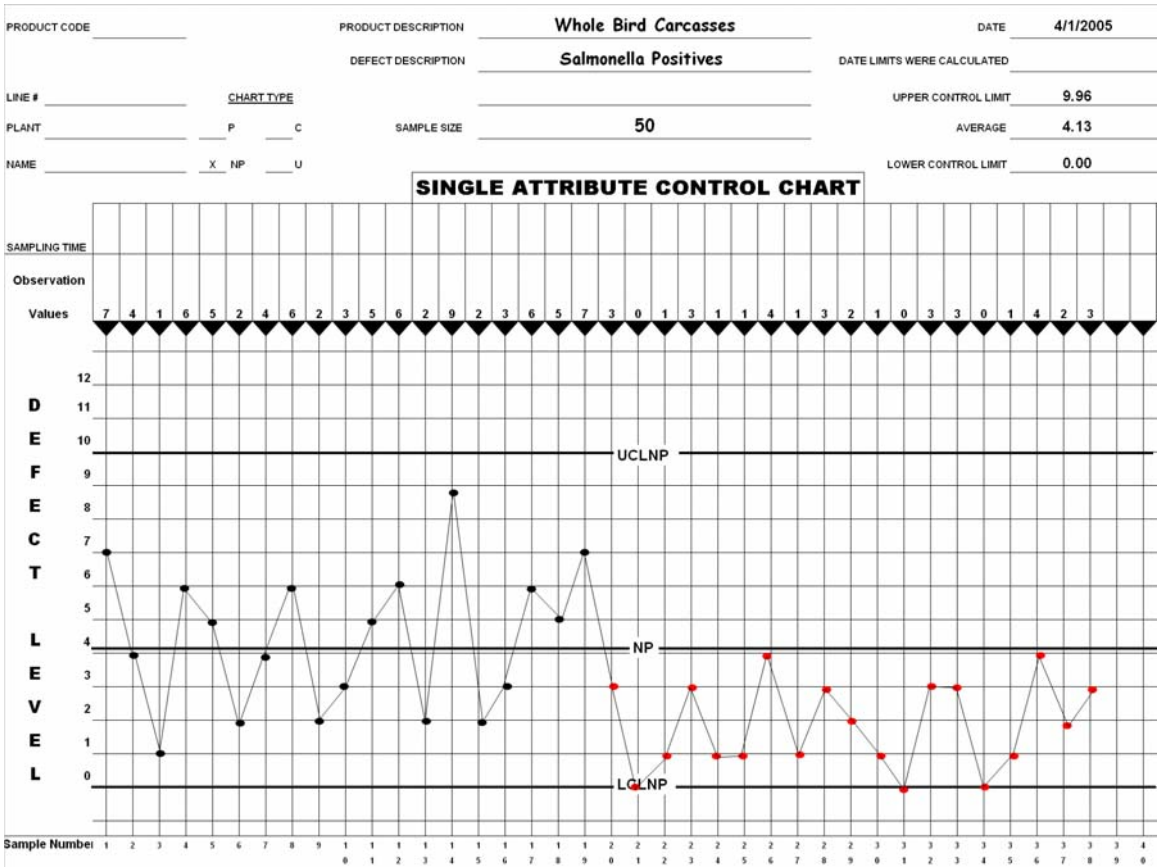



**Figure 4. NP chart showing a reduction in *Salmonella* spp.**

## Appendix 5 - Control Chart for Binomially Distributed Data Plotted as Proportions (P Chart), with varying sample sizes

The example given in Appendix 4, of the binomial control chart plotted the number of positive results out of a sample of 50 units. The characteristic feature in that example was that the number of units per sample was fixed (= 50), so that the expected number of positive results per sample (of 50 units) was the same. However, in many situations the sample size is not the same and thus the expected number of positive results would not be the same. Thus plotting the number of positive results is not appropriate for a control chart since the underlying assumption for the data to be used for plotting, namely, that the results are from a common distribution when the process is under control, would not be satisfied.

A simple adjustment might be to plot the proportion of positive results, $P_i$ rather than the number of positive results; however, while the expected value would be same for all samples, the expected variances of the results will no longer be the same. Thus, such data would not be usable for plotting for the reason given above. However, one possible way of correcting this is to plot: $Z_i = \sqrt{N_i}\ (P_i - P)$, where P is the assumed true proportion of positive results and $N_i$ is the sample size for the $i^{th}$ sample. In this case, the expected value of Z is zero, and the standard deviation of Z is $[P(1-P)]^{1/2}$. For sufficiently large $N_i$, the distribution would be the same (approximately normal) for each plotted data point, so that the $Z_i$ could be used for plotting a control chart. A control chart for Z would have Shewart control limits of $\pm 3[P(1-P)]^{1/2}$. CUSUMS and moving averages could be constructed with the $Z_i$ values. Or, if the sample sizes were not that large, an arcsine transformation: $y_i = \sin^{-1}(P_i^{1/2})$ could be used, setting $Z_i = \sqrt{N_i}\ (y_i - \overline{y})$.

If the number of distinct values of $N_i$ is small (say two or three) it would be possible to just plot the $P_i$, and have two or three Shewhart limits depicted on the same chart. The following is an example of a P-chart with two Shewhart limits for two values of $N_i$ ( = 50 or 100). .

Similar as above for the other charts, the steps involved are:

1. Define the characteristic…… Presence of *Salmonella* spp.
2. Determine sample size or sizes … Sample size = 50 or 100 25-gram units
3. Collect baseline data
4. Calculate Control Limits
5. Place Control Limits on chart of baseline data
6. Convert observations to proportions
7. Plot the baseline data
8. Connect consecutive plotted points with a straight line
9. Place control limits on a new chart
10. Collect and plot data as collected
11. Connect each point to previous point with a straight line
12. Observe chart for out of control signals after each point

For this example, Figure 1 is the number of *Salmonella* spp. positive units identified in a sample of 50 units of product is plotted versus sample number (or time of sampling) on the NP control chart. After about 30 data points (results from 30 samples) have been collected, (a rule of thumb for normal or nearly normal data is that 20-30 data points should used for control limit calculations) the control limits are ready to be calculated. These data will be used to calculate control limits when these values are converted to proportions.

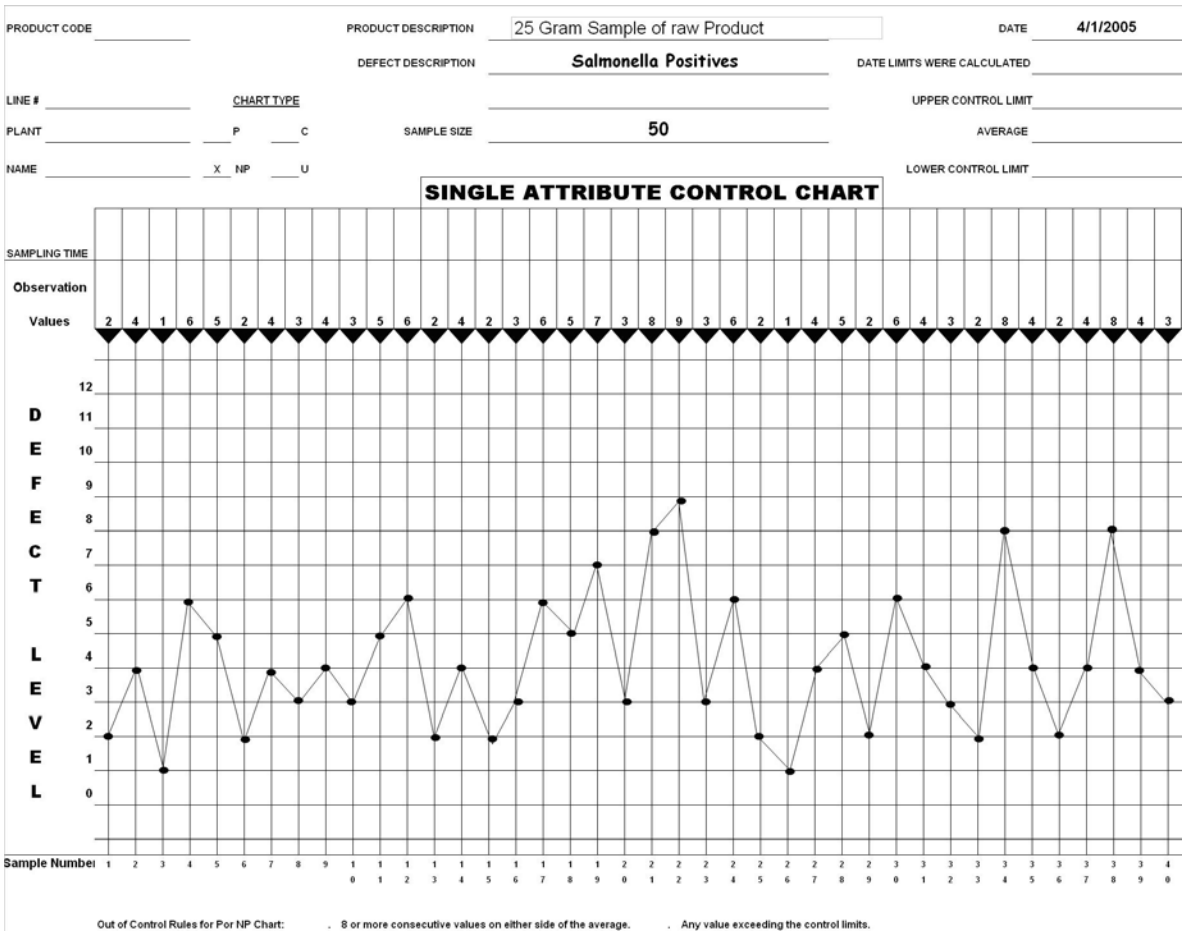



**Figure 1. Base line *Salmonella* spp. data collected in sample size = 50**

The formula for P Center line and control limits are:

Average Proportion Positive = $\overline{P} = \dfrac{\sum_{i=1}^{k} Xi}{Sample\_size \times Number\_of\_Samples}$

where Sample_size = $N_i$ = 50; and Number_of_Samples = k = 40 (in this example).

$$\overline{P} = \frac{165}{50 \times 40} = 0.0825 = 8.25\%$$

Center Line = $\overline{P} = 0.0825$

Control Limit Calculations:

Standard Deviation: σ = $\sqrt{\frac{\overline{P} \times (1-\overline{P})}{n}}$ - that is, it is assumed that distribution is a binomial distribution.

When $N_i = 50$

Upper Control Limit P:

$$UCL_P = \overline{P} + \left(3 \times \left(\sqrt{\frac{\overline{P} \times (1-\overline{P})}{n}}\right)\right)$$

$1-\overline{P} = 1 - 0.0825 = 0.9175$

$$UCL_P = 0.0825 + \left(3 \times \left(\sqrt{\frac{0.0825 \times 0.9175}{50}}\right)\right) = 0.1992$$

Lower Control Limit P:

$$LCL_P = \overline{P} - \left(3 \times \left(\sqrt{\frac{\overline{P} \times (1-\overline{P})}{n}}\right)\right)$$

$$LCL_P = 0.0825 - \left(3 \times \left(\sqrt{\frac{0.0825 \times 0.9175}{50}}\right)\right) = -0.0342 \text{ so } LCL_P = 0$$

For the process that the data were collected from above the average proportion of *Salmonella* spp. positive was 0.0825. The average or expected proportion of positive samples in a sample size of 50 is 0.0825, ($\overline{P} = 0.0825$), and the control limits calculate to be 0.1992 and 0 for the $UCL_P$ and $LCL_P$, respectively. By placing the control limits and center line on a P control chart and then transforming the actual baseline data into proportion by dividing by the sample size, (50 for the baseline data) and plotting these data on the control chart, the baseline data can be viewed for stability (Figure 2). Figure 3 gives a schematic showing the relationship of NP-charts and P-charts when the sample size is the same (=50).

When the sample size is 100, the Shewhart control limits are determined using the same formulas as above, expect substituting the sample size 100 for 50. The target mean remains the same at 0.0825; the upper limit decreases to 0.165; and the lower limit would be zero. For a sample size of 100, the probability of no positive samples is 0.018%, well below the α-probability of 0.135%. The probability of one or zero positive results in a 100-sample set is 0.18%, so that even one positive result could be used as the lower limit when the sample size is 100.

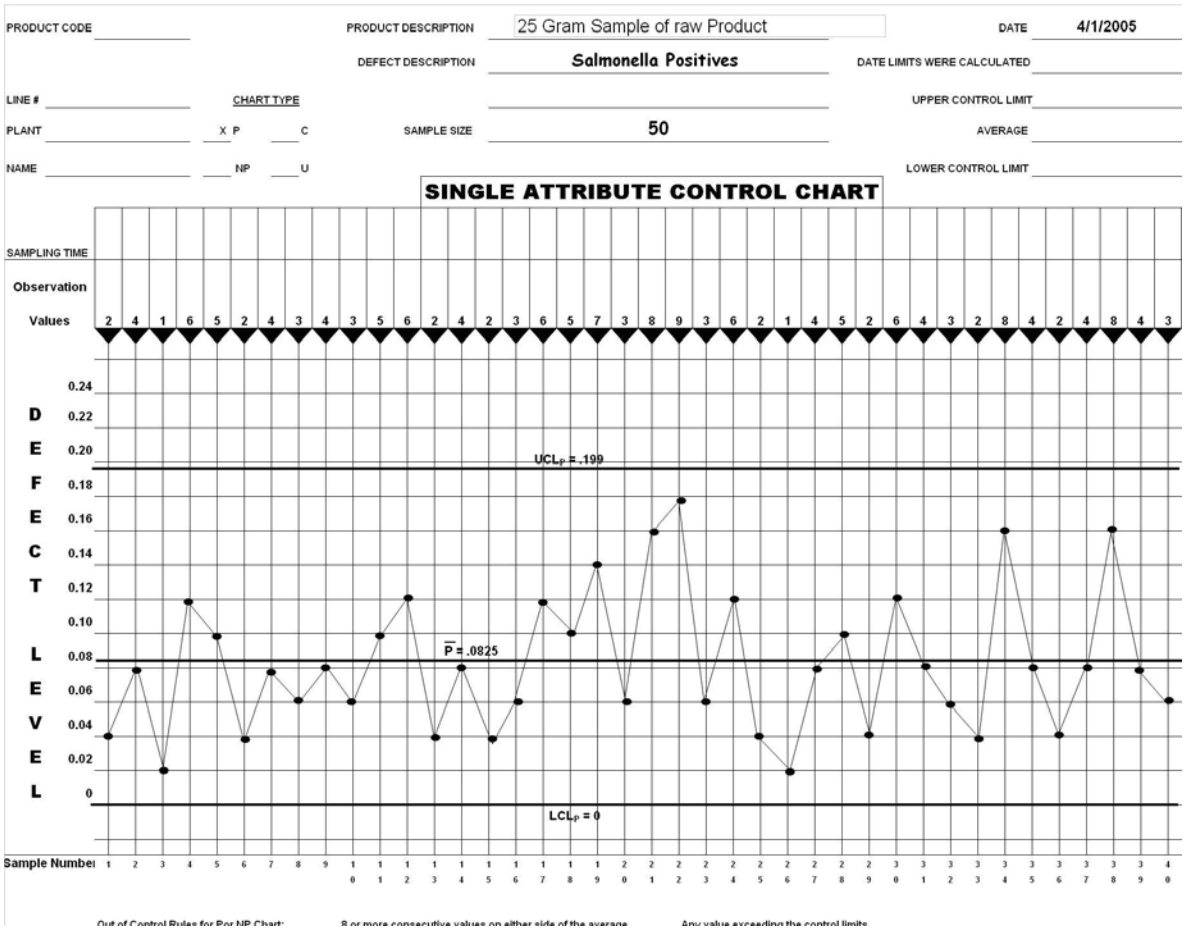



**Figure 2. P control chart of baseline data**

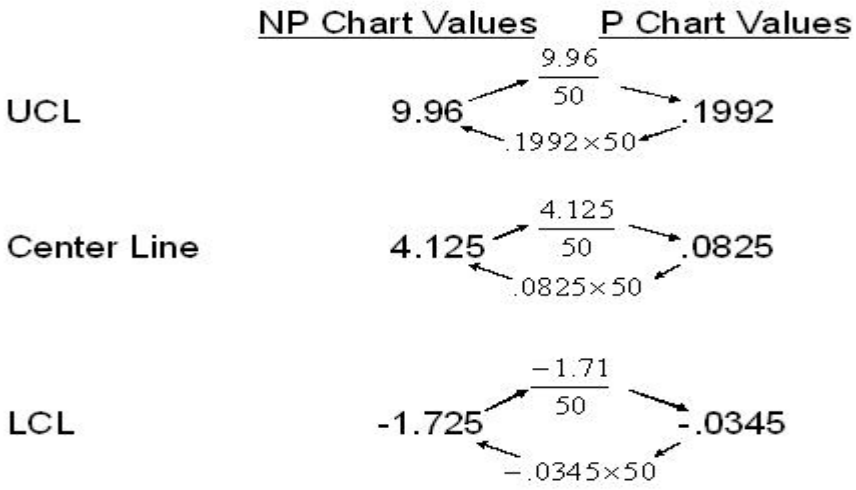



**Figure 3. A comparison of center line and control limits for an NP and P-charts.**

A control chart can be set up to accommodate two sample sizes. The new control chart has one lower control limit, (zero), one center line and two upper control limits, one for sample size 50 and one for sample size 100 (Figure 4). Data for both sample sizes are plotted on the chart. On this chart observations from a sample of size 50 are shown in black and observations from a sample size of 100 are shown on blue. All observed values are divided by their respective sample sizes before plotting. Notice how the data are connected in chronological order and sample size differences can be handled.

An interpretation of this chart is as follows: 1. the process was operating in a relatively stable manner for the first third of the chart. The result at sample number nine, where the data point fell between the two control limits, does not indicate an out of control signal since this point is associated with a sample size equal to 50 and this point is below the $UCL_P$ for sample size 50. 2. Even though the criteria for declaring a process out of control did not occur, the pattern of results suggested that improvements could be made. Suppose such an attempt to change the process was made in order to try to decrease the process mean, and the data points when the processing was conducted after the change was plotted. In this example (Figure 4), the change seems to have made a difference as the level of positives dropped after the change was made. 3. Further suppose that after additional data were collected so that there are about 30 values (in this example 28) new control limits were established for this process. Then the (data corresponding to the points in the circle would be used for the calculations of new control limits using the above formulas with sample size equal to 100 or 50).

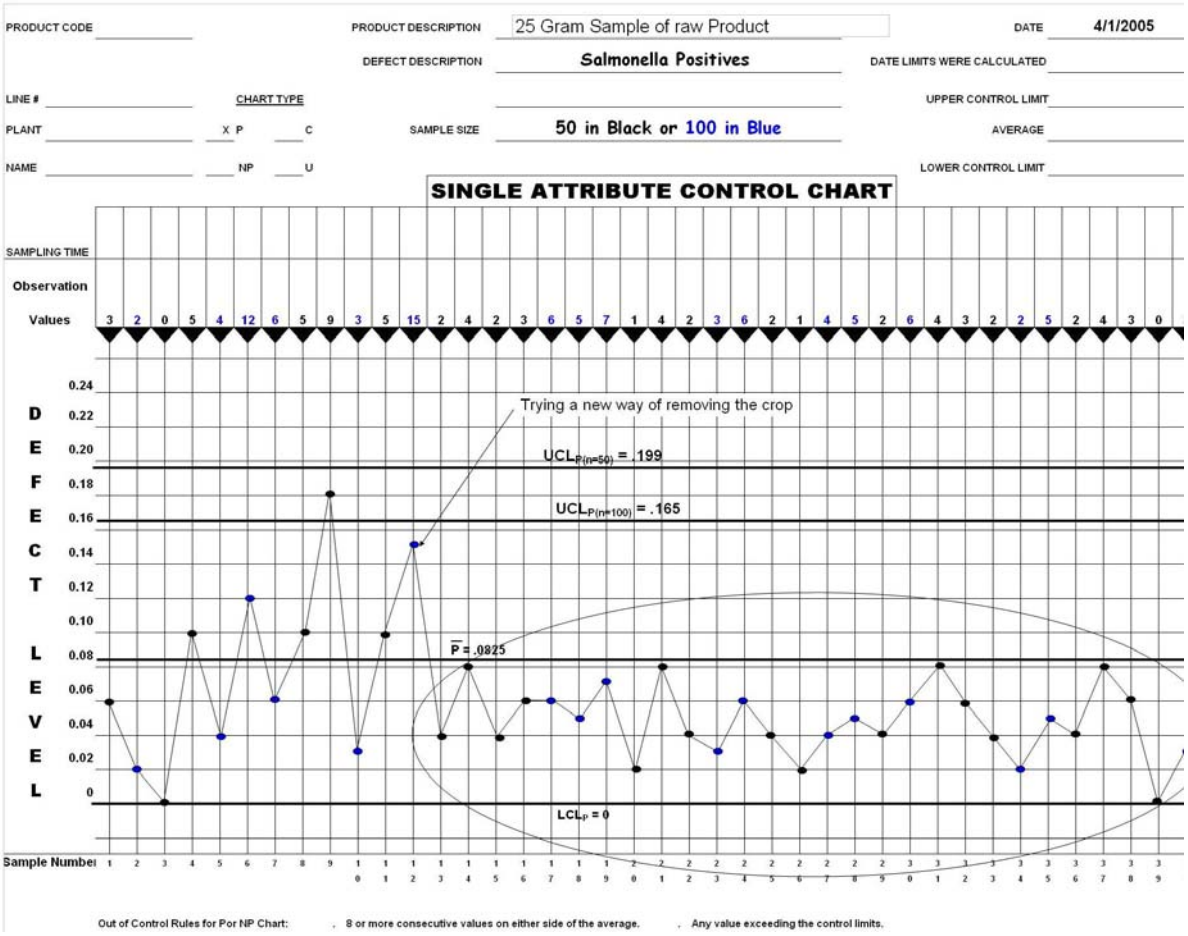



**Figure 4. P Chart developed to accommodate two sample sizes**

## Appendix 6: Control Chart for Poisson distribution for more than one sample size or when one expresses results in a unit size not equal to sample size - the U Chart

When data are collected using more than one sample size (amount of material sampled) and an underlying Poisson – like distribution can be assumed, a U chart may be used to do process control. A U chart is a plot of observation per sample, normalized to a fixed unit size.

In this example APC counts are measured for a fully cooked product. Briefly, a 10 % dilution is prepared by removing10 grams of fully cooked product from post-packaged product aseptically and placing it in 90 ml of diluent, stomached for 60 seconds and then 1 ml is plated and incubated for 48 hours. After the incubation CFUs are counted and data are reported as CFUs per gram. Since the actual amount of product in the one ml of plated diluent is actually 0.10 gram a situation is presented where count data are reported in units other than that equal to the sample size, and counts are low. At later time, the size if the same changes to a value different than 0.10 grams. In this case, all data could be standardized to be expressed in per gram units, where control limits would depend on the same size, in a similar fashion as given in Appendix 4, for P- charts. The standardization is just dividing the counts by the sample size: U = C/sample size. For these reasons a U chart is chosen as the chart to use for process control for this characteristic.
Actual Counts of the last 100 samples were:

**Table 1. Results from the previous 100 APC, (CFUs) per ml and the frequency of results**

| Count per ml. | Frequency |
|---|---|
| **0** | **49** |
| **1** | **29** |
| **2** | **13** |
| **3** | **3** |
| **4** | **3** |
| **5** | **1** |
| **8** | **1** |
| **11** | **1** |

For graphing purposes, each observation is divided by the sample size to express the results as CFUs per ml, (a 1 CFU outcome is reported as 1/.1 = 10 CFUs per ml

The steps involved in using a U control chart are:

1. Define the characteristic… APC counts per gram
2. Determine sample size …. 1 ml, .1 gram
3. Collect baseline data
4. Calculate Control Limits
5. Place Control Limits on chart of baseline data
6. Standardized the data by dividing by sample size
7. Plot the baseline data

8. Connect consecutive plotted points with a straight line
9. Place control limits on a new chart
10. Collect and plot data as collected
11. Connect each point to previous point with a straight line
12. Observe chart for out of control signals after each point

The formula for U Center line and control limits are:

Average Count = $\overline{U} = \frac{Total\_Number\_of\_CFUs}{Sample\_Size \times Number\_of\_Samples}$

Where: Sample_size =0 .1 and Number_of_Samples = k = 100 (in this example).

$\overline{U} = \frac{100}{0.1 \times 100} = 10$ CFUs per ml

Center Line = $\overline{U} = 10$

Control Limit Calculations:

Standard Deviation: σ = $\sqrt{\frac{U}{n}}$ - that is, it is assumed that distribution is a Poisson distribution.

The control limits for the U chart are calculated as:

Upper Control Limit U:

$UCL_U = \overline{U} + \left(3 \times \left(\sqrt{\frac{\overline{U}}{n}}\right)\right)$

$UCL_U = 10 + \left(3 \times \left(\sqrt{\frac{10}{0.1}}\right)\right) = 40$

Lower Control Limit L:

$LCL_U = \overline{U} - \left(3 \times \left(\sqrt{\frac{\overline{U}}{n}}\right)\right)$

$LCL_U = 10 + \left(3 \times \left(\sqrt{\frac{10}{0.1}}\right)\right) = -20$ so $LCL_U = 0$

These control limits were then placed on a single attribute, (U) control chart and the last 40 data points are plotted to view how some of the baseline fit on the chart (Figure 1).

The control chart illustrates how the user divides each observation by the sample size and plots the standardized results, (for the first observation 1 CFU is divided by 0.1 gram which provides a value of 10, so 10 is plotted as the first point, representing 10 cfu/g).

If the sample were to change to 0.5 instead of 0.1, then the control limits would change: the upper control limit would equal: $10 + 3(10/0.5)^{0.5} = 23.4$, and the lower control limit would equal 0, since $10 - 3(10/0.5)^{0.5} < 0$.

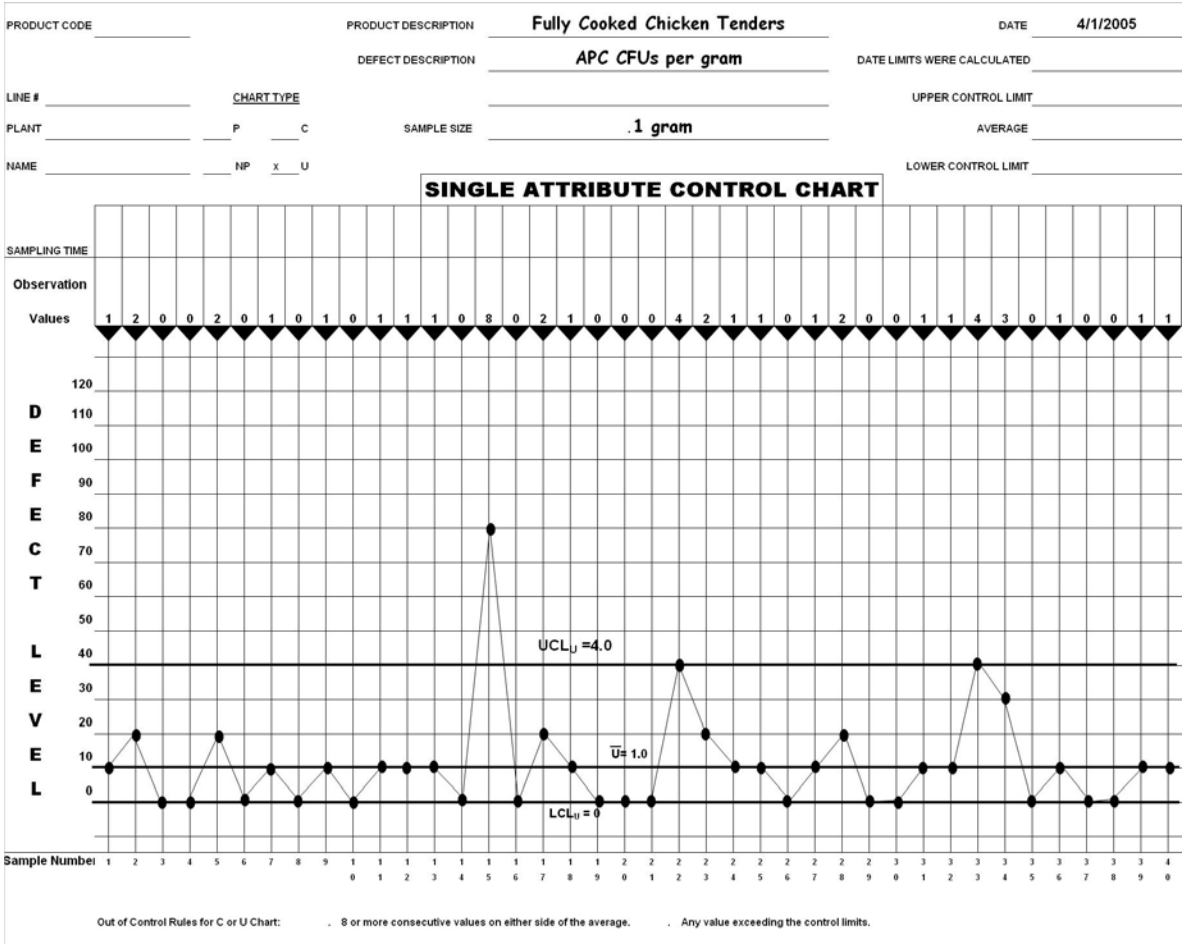



**Figure 1: U Chart of CFUs per gram of fully cooked product**

As with the P chart, a U chart can also accommodate more than one sample size, and, as with the P chart, the larger the sample size the closer the control limits are to the center line.

## Appendix 7: Control Charts for Rare Events:

### The Failure Control Chart, F Chart:

For events that are rare, to the extent that reasonable subgroup sample sizes would yield many zero values, a Failure Control Chart (F Chart) is an effective method for gaining an understanding as to whether the rate of the event is increasing, decreasing or remaining approximately stable. This particular chart was in fact developed to help answer this question. When the process is in control, the duration of time between events would be expected to follow an exponential probability distribution, which is described by a single parameter, given everything else being equal. In other words, it can be assumed that the number of failures expected over any number of times would be proportional to the number of samples, and the distribution of the number of failures would be binomial. The approximation made for determining the control limits is that the number of samples can be considered a continuous variable, associated with time. This assumption is reasonable when it is assumed that the failure rate is small. When the process is in control, it is assumed that the value of the failure rate parameter is constant over time.

In order to develop an F Chart the average time (number of samples) between events must be estimated. After an event, the time or number of samples since the last event are determined. The distribution of the times between events is assumed to be exponential distribution when the process is under control. Again the rule of thumb of observing 20- 30 or more events, to obtain a good estimate of the average time between events is recommended. The average time is referred to as "Mean Time Between Failure," MTBF, to keep consistent with reliability engineering convention. The probability of having not failed based on the current MTBF is calculated as follows:

> Reliability: R = Probability of not failing = $e^{\frac{-t}{MTBF}}$ where t is the number of samples since the previous failure.

High values of R imply low values of t, which would be undesirable.

The follows from the definition of the exponential distribution, which has the cumulative distribution function, $cdf(t) = 1- \exp(-t/\beta)$, where $\beta$ is a constant. The expected value of a random variable distributed as an exponential distribution with parameter $\beta$ is $\beta$. Thus the value of MTBF is estimated from baseline data consisting of many samples by just dividing the number of samples by the number of failures, assuming that this last value is not zero. To get a reasonable accurate estimate of MTBF, following the normal convention, the number of samples collected should provide about 20-30 failures. However, the standard error of this estimate assuming that the number of samples between failures is distributed as an exponential distribution is $MTBF/n^{0.5}$, where n is the number of failures. The error CV thus is $100\%/n^{0.5}$. To have an error CV of less than 20% would require more than 25 failures; to have CV of less than 10% would require more than 100 failures. If it is anticipated that the failure rate would be low, so that MTBF would be large, this latter number of failures might be difficult to obtain. However, an error CV larger than 20% could impact on the accuracy of the control charting. Thus it seems that at

least 20-30 positive results, and possibly more, should be used when computing MTBF for a control chart

For example, during a previous year, a plant collected 4,400 *E. coli* 0157:H7 samples, of those samples, 44 samples tested positive. The MTBF can be determined by dividing the number of sample by the number of positives:

$$\text{MTBF} = \frac{4400}{44} = 100$$

Time is a continuous variable and sample number is a discrete variable. This discrepancy may cause some problems when the MTBF is "small." In the example being presented, MTBF = 100, so that the probability t =1 is 1%, since $R = e^{-1/100} = 0.99$ is the probability of not failing, so that the probability of a positive sample is 1-0.99 = 0.01. The implication of this is that two consecutive positive samples, providing an observation of t = 1, is not enough to signal "out of control" if the control limits are set where the $\alpha$-probability is to be low, about 0.135%, based on the normal distribution assumption for the Shewhart (one-sided) control limit of $\mu + 3\sigma$. In order to have a $\alpha$-probability that low or lower, the MTBF must be no less than 750 samples.

There are many ways this "problem" can be dealt with. The easiest is just to count the number of samples between positive results, exclusively, so that the above example would provide an observation of t = 0, (two consecutive events would mean no negative results between events), and thus would automatically (regardless of the value of MTBF) provide an "out of control" signal. This is a "conservative" approach insofar as it assigns the number of days the minimum it could be assuming that time was a continuous variable and what is being measured is that time when a "failure" takes place. In practice this should not create a serious bias in the $\alpha$- and $\beta$-probabilities, but has the effect of increasing the $\alpha$-probability slightly while decreasing the $\beta$-probability slightly over actual values. Thus, the time, t in the above formula is, (t), equal to number of samples since last positive -1.

Steps required to develop an effective F Charts are:

1. Define the event of interest…….Positive finding of *E. coli* 0157:H7
2. Calculate the MTBF …….. From last 20- 30 events
3. After an event determine the time or number of samples since the last event.
4. Determine t, by subtracting 1 from the answer in 3.
   (or combining 3 and 4, determine the number of samples between positive events , exclusive of the positive samples)
5. Determine the 'probability', R, of having not failed, using t, computed in 4.
6. Plot this probability of having not failed.
7. Connect the plotted point to the previously plotted point.
8. Review the chart for out of control signals after each point.

Note that the MTBF is not directly placed on the F-Chart, rather the F-chart is scaled from zero to 100% with limits set at a probability equal to those historically set by Shewhart at 0.13% and 99.87%. These probabilities correspond closely to the upper and lower control limits set at $\mu \pm 3\sigma$ for the usual Shewhart control chart, discussed above. The target line corresponding to the mean is at 50% (labeled $\tilde{F}$). When the times between events are distributed as an exponential distribution, the probabilities, R, will be distributed as a uniform distribution between 0 and 1, and thus the data points, R, would be randomly distributed around the center line (0.5) rather than being distributed non- symmetrically as would be the case if the times themselves were plotted. Thus R could be used for constructing moving averages, CUSUMS, or other trend statistics, keeping in mind though that the underlying assumption is a uniform distribution rather than a normal distribution.

After each event, the time since the last event or number of samples since the last event are placed in the space labeled, Time Since Last Failure - 1 (t), and the probability of not having a failure is calculated and entered in the space labeled Probability of not Having a Failure ( R ). An F Chart with a MTBF of 100 may look like that illustrated in Figure 1.

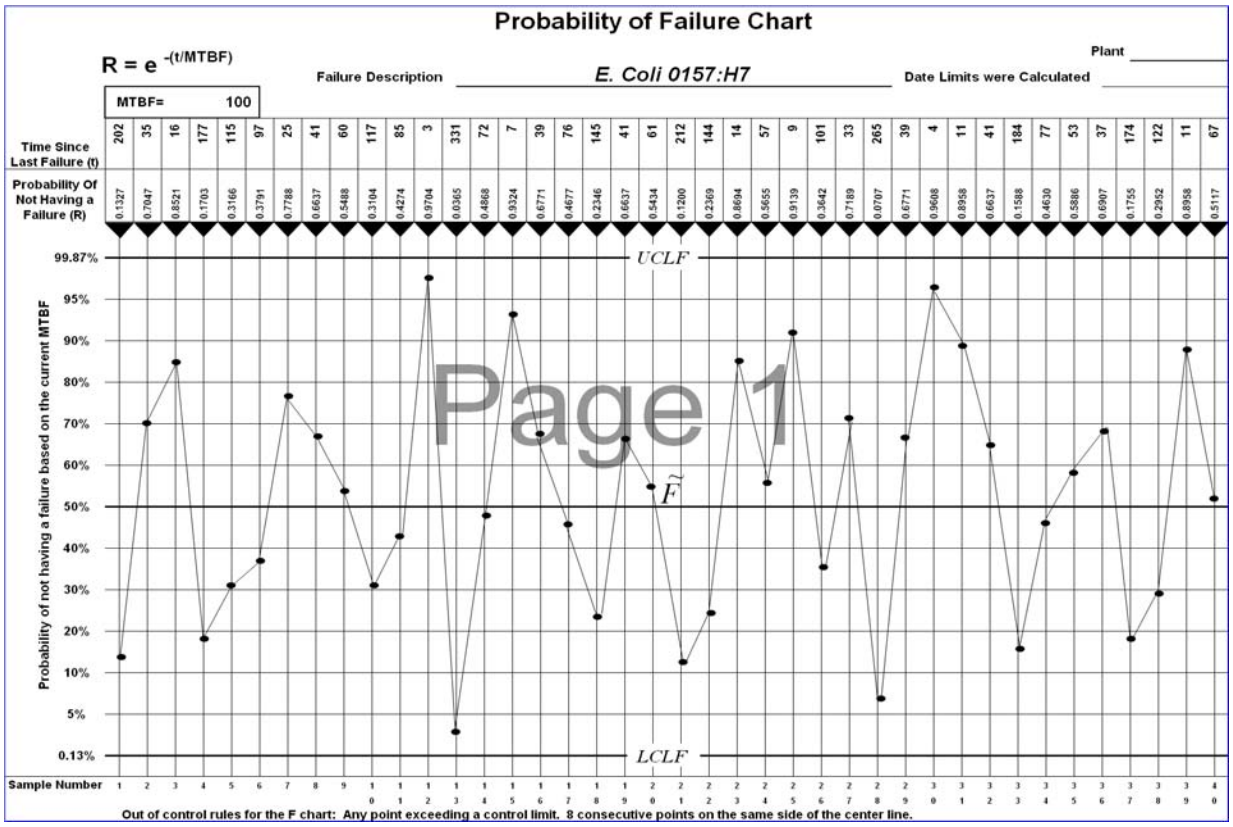



**Figure 1. An F Chart of *E. coli* 0157:H7 events for a process with a MTBF of 100 samples**

For this example, with a MTBF of 100 samples, the first event on this chart occurred after 202 samples (actually the 203rd sample). This corresponds to a probability of not having a failure of 0.1327 because

$R = e^{\frac{-t}{MTBF}} = e^{\frac{-202}{100}} = 0.1327.$

Stated another way the process had about an 87% chance of having an event before it did. In setting the chart up in this manner, it allows one to look at the graph of events in a similar fashion that one would look at a single attribute, such as, the NP chart discussed in Appendix 4. For example, eight consecutive points below $\tilde{F}$ (Figure 2) would indicate an increase in MTBF and eight consecutive points above $\tilde{F}$ (Figure 3) would indicate the MTBF is decreasing. Other indicators that the MTBF is increasing or decreasing would be a single point below the $LCL_F$ or a single point above the $UCL_F$, respectively. Figure 4 illustrates a process with a point below the $LCL_F$. (Note, as discussed above, that because the MTBF is only 100, for a point to exceed the UCLF requires two consecutive positive samples, or zero samples between failures.)

Of course, as with all SPC charts, an out of control signal should be investigated. An investigation of points above the $UCL_F$ or eight consecutive points above $\tilde{F}$ = 0.5 would help one identify processing conditions which raise the probability of an event. Removal of the conditions which raise the probability of an event could lower the probability of the event. This would in turn cause an increase in the MTBF, resulting in either a point below the $LCL_F$ or eight consecutive points below $\tilde{F}$.

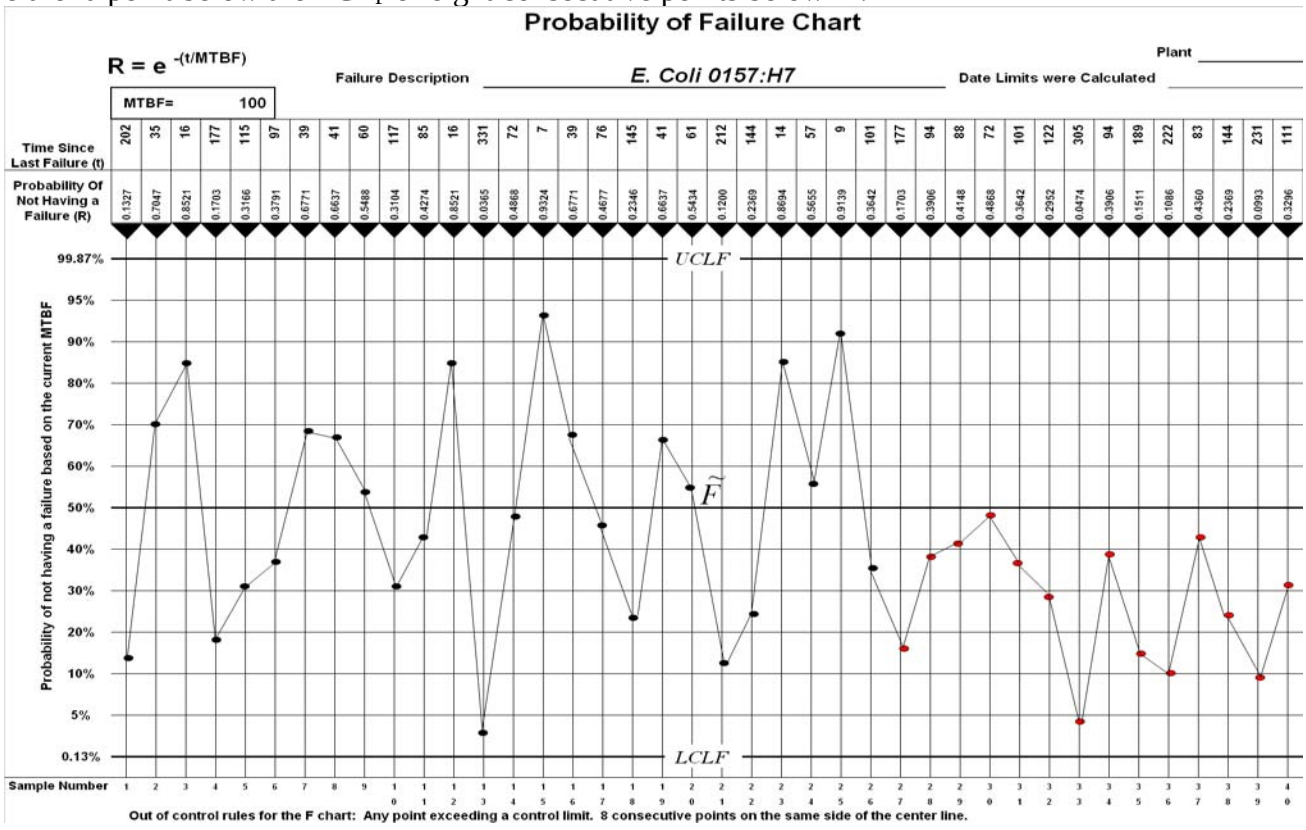



**Figure 2. An F Chart showing an increase in the MTBF**

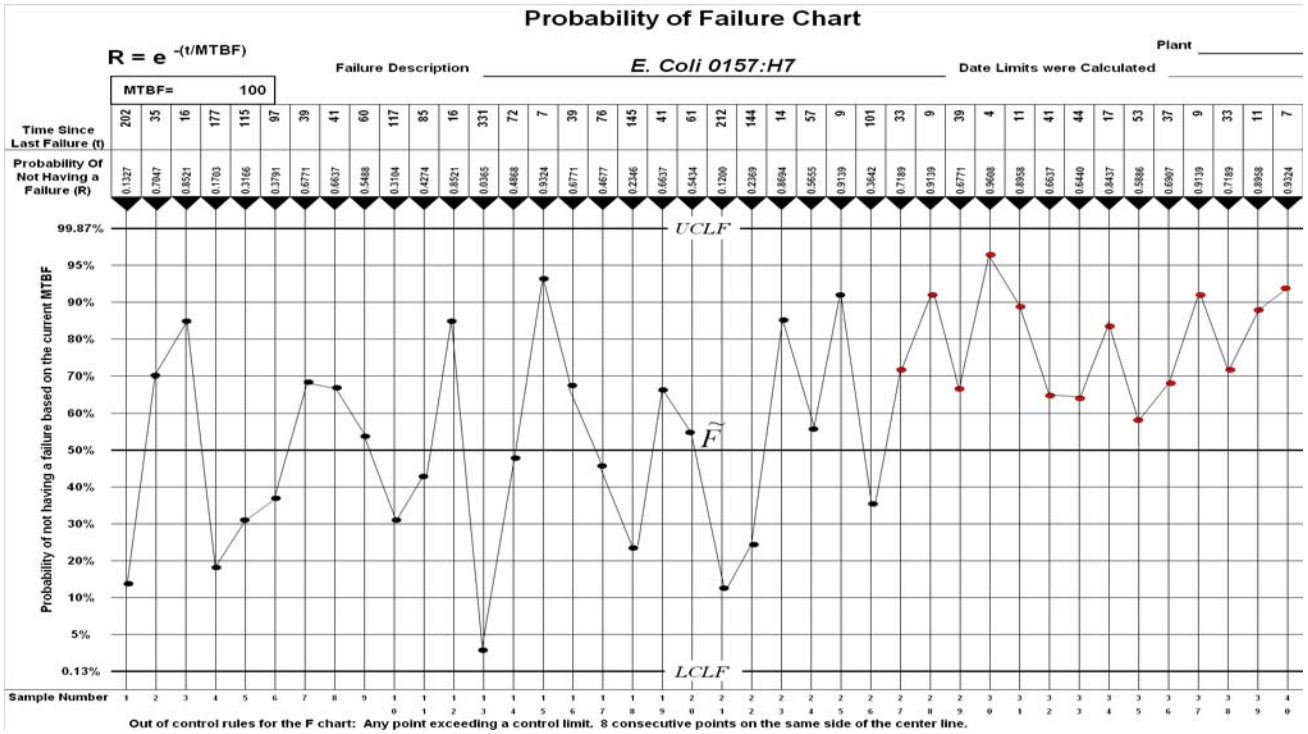



**Figure 3. An F Chart showing a decrease in the MTBF**

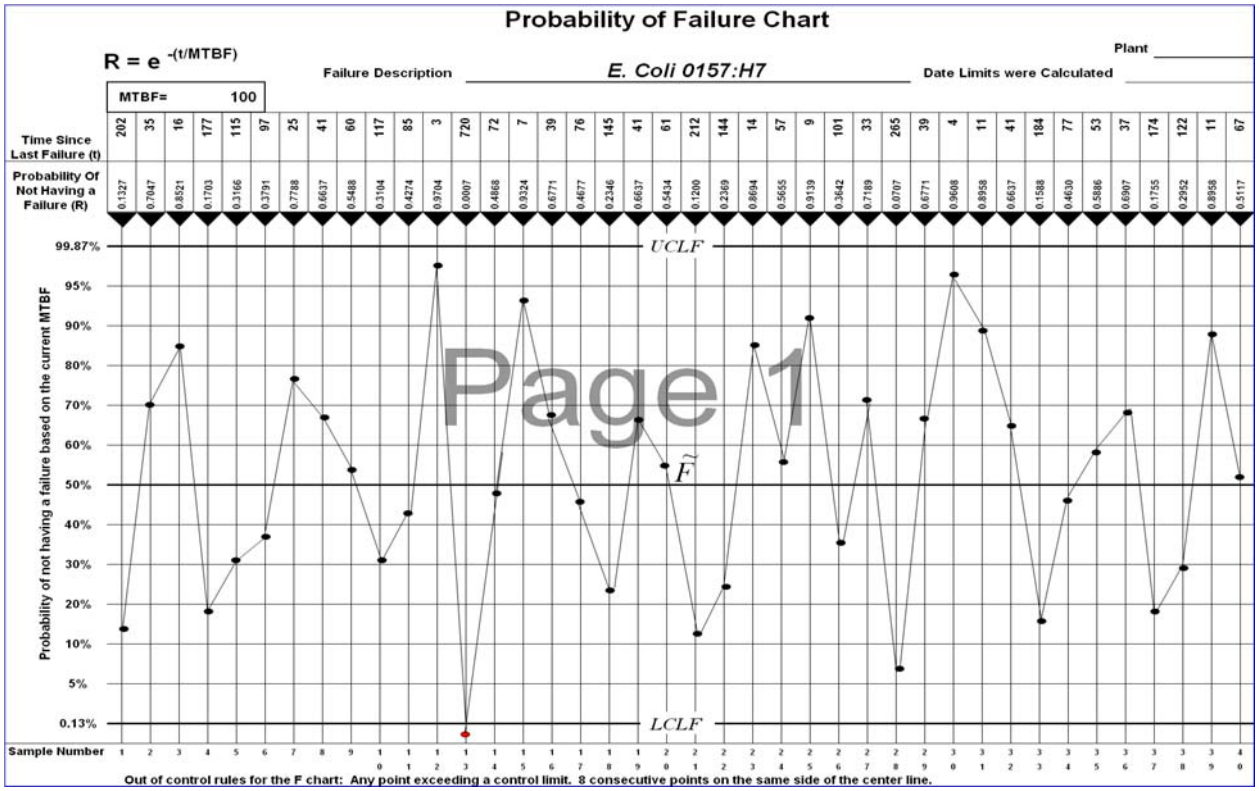



**Figure 4. An F Chart showing a point below the LCLF, indicating the MTBF may be increasing**

[1] Almost any microbiological data set of counts will have a distribution that is skewed so that a log transformation would make the distribution more symmetric. See for example the USDA Food Safety Inspection microbiological baseline surveys for counts of generic E. coli and other organisms. In fact this seems to be mostly true for any population data of living things. Part of the reason might be due to the inherent randomness associated with growth or cell division that "pure living" systems exhibit – namely an exponential growth.

[2] American Society for Testing Materials, from ASTM Manual on Quality Control of Materials, Philadelphia, January 1951, p. 115.

[3]Discussion of maximum likelihood estimation, and maximum likelihood ratio and chi-square statistics can be found, in: Kendall and Stuart's The advanced theory of statistics, vol. 2. Charles Grifffen and Company Limited, London.

[4] SAS, SAS Institute Inc. , Cary NC.

[5] Wheeler, Donald J. and Chambers, David S. 1992 "Understanding Statistical Process Control. 2nd edition." SPC Press, Inc.

# AOAC INTERNATIONAL
# Presidential Task Force on
# Best Practices for Microbiological Methodology
# US FDA Contract #223-01-2464, Modification #12

Executive Summary
Statistics Working Group

**Objective:** The objective of the STWG has been to review current practices and alternative approaches for validation of microbiological test methods (including growth-related and chemical tests), and make recommendations based on the past experiences and ongoing activities of the group members. This was to include statistical methods for analyzing data from validation studies. At the outset of this project it was acknowledged by the BPMM Steering Committee and by the STWG that the project provided neither the time nor the resources to fully validate all the recommendations of the group. Some recommendations, such as the use of $LOD_{50}$ for qualitative methods, may require further development, before being widely used.

**Determining method performance:** Performance standards should be based on criteria based on fitness for the intended use, including public health needs. In general, statistical methods should be used to assist in setting realistic performance standards. These methods should be based on control of Type I and Type II error, which implies the determination of levels of unsatisfactory performance that must be detected (with stated probability) and controlled. It also implies use of appropriately determined sample sizes to meet the stated goals relative to stated $\alpha$ and ß. This approach would be a change from current practices in which studies are accepted on the basis of standard designs for number of laboratories, materials, and replicates, and standard criteria for suitability of the summary statistics. The design specifications and resulting reliability estimates should form the basis of applicability statements for test and measurement methods. (Ref 1, 9, 10, 13, 14) (*Task 2: What are the scientific/statistical bases for developing performance standards against which the validation of methods should be based*?).

The committee supports the use of appropriate international consensus standards. For consensus standards that are currently under development, the STWG recommends active participation in the development and/or validation of the standards. In general, the STWG acknowledges the value of rigorous consensus processes and international harmonization of method validation procedures. Specific approved international consensus standards include the following:

a) ISO 16140 Microbiology of food and animal feeding stuffs – Protocol for the validation of alternative methods
b) ISO 5725 Series: Accuracy (trueness and precision) of measurement methods and results.
c) ISO 11843 Series: Capability of Detection
d) CLSI/NCCLS EP17-A: Limits of Detection and Limits and Quantitation for Quantitative Measurement Procedures.

Standards under development include ISO draft Technical Specification 19036: Microbiology of food and animal feeding stuffs – Guide on estimation of measurement uncertainty for quantitative determinations.

Statistical procedures recommended in ISO 16140 are appropriate for "alternative methods" where there is an accepted reference method, but many of the procedures can also be used where there is no reference method. This document recommends use of robust statistical procedures that do not necessarily assume a normal distribution and are not so severely affected by extremely large or small outlier results that can be misleading with more conventional procedures. It also recommends against the removal of outliers from collaborative studies, except for assignable causes. The STWG fully agrees with these recommendations.

The committee strongly urges caution in applying the concept of "false negative" and "false positive" results because of the difficulty of confirming all positives and negatives, and the likelihood of misinterpretation. Alternative confirmation procedures should be considered, such as nucleic acid testing. Any estimates of "sensitivity" for low level samples should be corrected using appropriate statistical methods, such as adjustments for expected true negatives predicted with the Poisson distribution. Protocols should continue to include the appropriate Chi Square test based on whether or not samples are paired. (Ref 1-5, 14) (*Task 10. What are the appropriate statistical tools to be used for interpretation of validation studies*?)

Predictor and response variables important to the study design as well as for validating methods must be discussed and accepted by all subcommittees and the Steering Committee after review of all reports. Initial considerations should include variables that have been identified in the reports from other task groups. (Ref 15) (*Task 11: What are the test variables (e.g., number of strains, foods, inoculum levels) that should be considered for each of the factors listed in Task 8?*)

**Estimating uncertainty**: Uncertainty in measurements using quantitative procedures is best estimated following an all-inclusive, or "top down" approach. This approach does not attempt to estimate all components of uncertainty separately and it does not require a detailed mathematical model of how those components are combined. This approach is in contrast to a "bottom up" approach, which provides an estimate of the uncertainty of the method rather than the measurement and requires estimation and combination of variances at all stages of an analysis. This cannot be done routinely, however, so standard, or assumed, variances are used which aligns the combined estimate to the basic method rather than the analytical result. The "bottom up" approach is likely to underestimate uncertainty due to sources of uncertainty that are not considered. By contrast, the "top down" approach makes no attempt to set generic estimates of uncertainty for specific test methods and rightly aligns the estimate of measurement uncertainty with a specific analysis (or set of analyses). The "top down" approach is consistent with the Guide to the Expression of Uncertainty in Measurement (GUM) principles that allow combination of sources of uncertainty that are difficult to estimate individually. Comprehensive estimates of uncertainty can be obtained from collaborative studies, from carefully designed validation studies, or in some cases from routine quality monitoring data.

For qualitative methods, measurement uncertainty for the result cannot be expressed directly – instead, the measurement uncertainty relates to the probability of reporting an incorrect result. This can be estimated with false negative and false positive rates, for those methods with confirmation procedures (Ref 7). For some measurement procedures, uncertainty can be expressed as the standard error of a limit value estimation e.g. the $LOD_{50}$, as estimated by the Spearman-Karber or some alternative method (Ref 11,15,16). This procedure estimates uncertainty where it is most important, which is at the border of the determination of "present" or "absent" (that is, in the area of the detection limit). The work of ISO Technical Committee 34, Subcommittee 9 is not yet completed, so the STWG recommends active participation in the efforts of this subcommittee. (Ref 5) (*Task 6. What are the effective means for articulating the uncertainty associated with microbiological methods?*)

**Limit of Detection:** The detection limit for qualitative tests is best described as the "$LOD_{50}$", or number of organisms per gram of sample at which 50% of the tests are positive. This is determined with a nonparametric (distribution free) version of probit analysis, and an experimental study using at least 4 dilutions in which at least two of the dilutions have "fractional positives" in order to better estimate the $LOD_{50}$ and perhaps allow for estimates of other percentiles, such as the $LOD_{90}$ (number of organisms per gram of sample where 90% of results are positive).This procedure also assumes that one dilution level has 0% positive results and one dilution level has nearly 100% positive results (allowing for measurement error in the test laboratories). (Ref 12,13, 17, 18).

For quantitative methods, the committee recommends use of the ISO 16140 procedure, which presents limits of detection and quantification as functions of the variability of blank (or very low) samples. The committee recognizes, however, that alternative procedures exist that should be investigated, such as the ISO 11843 Series on capability of detection, or the nonparametric analog of that procedure, as described in the CLSI document EP17-A on Limits of Detection and Quantitation. These procedures recognize the importance of Type I and Type II errors, and that variances of signals from truly negative and truly positive samples can be different (Ref 1, 3, 4). There are related strategies for designing experiments to use the ISO/CLSI approach (*Task 4: What are scientific/statistical bases for determining the lower limit of detection for microbiological methods? How is the lower limit of detection validated during the validation of a method? How is the relative performance of a method determined as the lower limit of detection is approached and what is the best way of characterizing this performance?*)

**Topics for further research**

In the course of this review, the STWG identified several areas where further research was needed, or a more comprehensive review of the documents developed for this study. The areas of further review include the following (Ref 19):

1. Further development of procedures for describing the Limit of Detection for quantitative methods.
2. Further development of recommendations for use of the generalized Spearman-Karber method for estimating the $LOD_{50}$ for qualitative methods.

3. Evaluation of alternative approaches to the Spearman-Karber method e.g. Logit, Probit and other statistical procedures currently under investigation by the ISO TC34/SC9/SWG.
4. Investigation of the effectiveness of current AOAC Official Methods for Single Laboratory Validation (SLV) procedures, Multiple laboratory Validation procedures (MLV) and harmonized Collaborative Validation studies (HCV), relative to the recommendations concerning the design of verification studies.
5. Use of existing AOAC study data to evaluate the alternative statistical methods proposed.
6. Use of existing AOAC data for assisting in design issues for future validation studies. This could include proper consideration of Type II error in addition to Type I error, and should develop a structured approach for making decisions based on the data.

## References

1. ISO 16140: Microbiology of food and animal feeding stuffs – Protocol for the validation of alternative methods

2. ISO 5725 Series (Parts 1-6): Accuracy (trueness and precision) of measurement methods and results.

3. ISO 11843 Series (Parts 1-4): Capability of Detection

4. NCCLS EP17-A: Protocols for the Determination of Limits of Detection and Limits of Quantitation. Clinical and Laboratory Standards Institute, Wayne, PA., 2004.

5. ISO TS19036: Microbiology of food and animal feeding stuffs – Guide on estimation of measurement uncertainty for quantitative determinations (draft)

6. Feldsine, Abeyta, Andrews: Journal of AOACI Vol 85, No.5, 2002: AOAC INTERNATIONAL Methods Committee Guidelines for Validation of Qualitative and Quantitative Food Microbiological Official Methods of Analysis

7. Wilrich, P-T (2005b) *The determination of precision of measurement methods with qualitative results by interlaboratory experiments.* Discussion paper for the ISO TC34 SC9 Statistics Group Meeting, Paris, April 2005

8. Linnet K, Kondratovich M. Partly Nonparametric Approach for Determining the Limit of Detection. *Clinical Chemistry* 50(4): 732-740; 2004

9. McClure F D, Lee J K. Sample Sizes Needed for Specified Margins of Relative Error in the Estimates of the Repeatability and Reproducibility Standard Deviations. *JAOACI*, in press, 2005.

10. McClure F D. Design and Analysis of Qualitative Collaborative Studies: Minimum Collaborative Program. *JAOACI*, 73(6): 953-960. 1990

11. Miller J, Ulrich R. On the analysis of psychometric functions: The Spearman-Karber method. Perception & Psychophysics 63(8): 1399-1420. 200

12. Paulson, D. S. (2003). *Applied Statistical Designs for the Research.* New York: Marcel-Dekker.

13. Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for Experiments: An Introduction to Design, Data Analysis, and Model Building*. New York: John Wiley & Sons.

14. STWG report: Task2&10.doc

15. STWG report: Study Variables.doc 022505

16. STWG report: Combined report BPMM SWG_Uncertainty consolidated_rev 1.pdf

17. STWG report: LOD50% 20050721.doc

18. STWG Excel worksheet: LOD50% Spearman-Karber.xls

19 STWG report: Recommendations for Future Research.doc

# Developing standards and validating performance: scientific/statistical bases for describing the validation of performance.

## I. Principles:

1. Performance standards should be based on appropriate statistics for describing method performance and expert/regulatory objectives for the intended use. Performance statistics might vary for different measurement technologies, although attempts should be made to harmonize these. Minimal performance criteria could be different for every performance statistic and for every use.

2. The performance characteristics should be estimated with experimental protocols to assure that confidence intervals for the statistics will be small enough for their intended use (this controls Type I error and Type II error).

3. All performance characteristics should be specific for a defined organism or strain of particular interest. That is, the "measurand" (or "analyte" in common usage) should be defined exactly; sometimes this might include a single strain, sometimes a class of strains, a genus, or a group of organisms. Similarly, the measurand might be a specific microbial toxin, a group of toxins, or some other parameter. With a carefully described measurand, the concepts of "inclusivity" and "exclusivity" are variants of "sensitivity" and "specificity". These statistics are useful for general descriptions of a method that is approved for a microorganism with many important strains.

4. Results from experiments in a single laboratory can be useful for the design of collaborative studies, but should not be used alone to establish claims for method performance (except for the particular laboratory's own purposes). Performance characteristics should be estimated, wherever possible, with experiments conducted in two or more laboratories that have demonstrated competence with this type of microbiological procedure and experimental protocol. At least one of the laboratories must be independent of the manufacturer/developer of the method. It is acceptable to validate performance of a method in a single laboratory and have a second laboratory verify the performance on a carefully selected subset of matrices, but this needs to be done with care, and following the recommendations from the BPMM Matrix Extension Recommendations or ISO 16140. Therefore determinations of bias between an alternative and reference method cannot be determined in a single laboratory, nor can inclusivity or exclusivity (sensitivity and specificity), unless the determinations are verified in at least one other laboratory.

5. It is essential to differentiate between the uncertainty (*sic* lack of precision) of a method and the uncertainty of an estimated value derived using that method (*sic* the measurement uncertainty). Estimates of measurement uncertainty should be derived from appropriate "top down" procedures using intra- or inter-laboratory randomized trials; in some instances, such estimates may be specific to each individual laboratory undertaking a specific test. By contrast, uncertainty estimates for the method are derived by "bottom up" procedures and must be used with care since they normally underestimate the true extent

of uncertainty of a measurement. Upper limits for uncertainty estimates may be used by those laboratories that can demonstrate competence with the method.

## II. Considerations for Statistics and Statistical Methods

1. The statistics used should possess the following qualities:
   - Unbiased, maximum likelihood estimates for the performance characteristics of interest.
   - Appropriate for the distribution of data from which they will be calculated.
   - Understandable and intuitive for microbiologists and regulators.
   - Sensitive to the most common sources of error or deviation from expected performance.

2. Criteria for the suitability of the performance statistics should be based on the following considerations:
   - Professional judgment on the performance level required for the method for the intended use. This should be based on considerations for public health or fitness for purpose, and technical knowledge of the method.
   - There should be definitions of performance that is not suitable for a prescribed purpose; that is, poor performance that should be detected with high probability.
   - Probability of improperly rejecting a method as unsuitable, when in fact it is suitable for the intended use (control of Type I error).
   - Probability of improperly accepting a method as suitable, when in fact it is not suitable for the intended use (control for Type II error).

3. The data used to characterize performance should be based on the following:
   - Data from more than one laboratory.
   - Statistics based on all results received from competent laboratories, all following the same well-defined instructions for the measurement procedure and reports (discard only those data outliers for which there is a known cause).
   - Data transformed to reasonable normality and analyzed using appropriate robust or nonparametric methods. Severe non-normality of the transformed data (many statistical outliers) or evidence of bimodality should be resolved prior to analysis of the data.

## III. Considerations for calculating performance statistics from collaborative studies.

Carefully designed collaborative studies are preferred for describing the performance capabilities of a measurement procedure. The guidelines for determining the numbers of laboratories, levels, and replicates are well established (see for instance McClure & Lee, 2005). Procedures for analysis of the data are less well established.

1. Before summary statistics are generated, it is important to look first for laboratories that seemed to have difficulties with more than one sample, or whose results are consistently high, low, or highly variable across levels. These are the laboratories that were possibly affected by ambiguous instructions, a missing step in the procedure, or other inherent weakness in the measurement procedure. These situations must be investigated before data

analysis proceeds. Any truly erroneous results must be eliminated - or in some cases they can be corrected (as in decimal point errors or switched samples). No results should ever be eliminated for purely statistical reasons. If reasons cannot be found, then the variability is assumed to be representative of the procedure. Obvious any bimodality in the data must also be resolved, possibly using 'bump hunting' procedures. Once the truly erroneous results are eliminated then the statistical processing can commence.

2. ISO 16140 recommends use of robust statistical procedures rather than conventional parametric statistical techniques, and the BPMM STWG agrees with this recommendation. However, whether extreme results are eliminated as outliers or have their impact limited with robust techniques is less important than the analyst's investigation of how such outlying results occurred.

3. For qualitative method comparison studies, the definition of the "Reference Method" is important for naming the performance statistic. If the Reference Method is definitive for confirming the presence and absence of an organism, then it is possible to use "false positive" and "false negative" as summary measures for an alternative method. Similarly, if definitive confirmation techniques are available then "false positive" and "false negative" for an alternative method may be reported. However if the Reference Method is not definitive, then performance measures are relative to the Reference Method and must be described as "relative sensitivity" and "relative specificity".

4. McNemar's Chi-Square test is appropriate for testing for significant disagreement between Reference and Alternative Methods, but is appropriate only when samples are truly pairs – that is when they share a common enrichment or pre-enrichment step. Artificially linked samples are not appropriate for McNemar's test.

5. The term "false negative" (FN) is a confusing concept, even when using only confirmed positives. When there are few organisms in the sample, the FN rate may be a combination of results where an organism was present but not detected and results from samples that truly contained none of the target organisms, due to inhomogeneous distribution of organisms in the larger sample. It is possible using the Poisson distribution (or if appropriate the Binomial or Negative Binomial distribution), to adjust the false negative rate to account for the estimated number of true negatives. Therefore, even when only confirmed positives are used, false negative rates should be adjusted for the theoretical likelihood of having a true negative.

6. The $LOD_{50}$ is an independent descriptor of performance, and is preferred to measures that are relative to the Reference method (such as false negative or false positive). .

7. If possible, all samples should have their positive/negative status confirmed by an independent methodology (except for those samples that are positive by both the reference and alternative methods). Samples that are negative by both methods should also be confirmed by independent methodology, if possible.

## IV. Recommended performance statistics

Qualitative Methods

1. Number of cfu per gram of matrix for 50% probability of a positive signal ($LOD_{50}$)
2. Number of cfu per gram of matrix for 90% probability of a positive signal ($LOD_{90}$)
3. Probability of a negative signal when the Reference Method indicates no organisms is present (relative specificity).
4. Probability of a negative signal when common contaminants, but not the target organism, are added to a sterile sample (specificity).
5. Probability of a positive signal when the Reference Method indicates organisms are present (relative sensitivity).
6. Proportion of replicates with same result (repeatability)
7. Proportion of results from different laboratories with the same correct result (reproducibility).
8. Standard Error of the $LOD_{50}$, for use in estimating the effect of measurement uncertainty on the probability of obtaining an incorrect result.

Quantitative Methods

1. Difference between replicate samples obtained under repeatability conditions (intra-laboratory repeatability).
2. Difference between replicates from the same material in the same laboratory, using changed conditions (intermediate reproducibility).
3. Difference between average results from different laboratories, testing the same material (reproducibility).
4. Average difference between the Alternative Method and the Reference Method pooled across multiple competent laboratories (relative method bias).
5. The extent to which the measurement signal is proportional to the number of organisms in the sample (linearity).
6. Range of quantification: the lowest and highest signals that can be detected with adequate uncertainty, obtained by dilution. For plate count methods, this is the range of counts per plate where results can be obtained with a stated degree of repeatability precision.
7. Lowest level where results can be obtained with a stated uncertainty that is fit for its purpose (limit of quantitation).

# STWG Objective #11 – Variables

1. Number of strains
   For qualitative studies with pathogens, it depends on the organism. It has been AOAC practice to require "inclusivity" testing of at least 50 target strains in pure culture and "exclusivity" of at least 30 non-target strains in pure culture. The collection of target strains should be representative of the breadth of the target group in terms of genetic or serological types. Non-target strains should represent those most closely related to the target group biochemically, serologically, or genetically, e.g., other members of the Enterobacteriaceae for the case of *Salmonella* spp. as a target. Target strains are to be cultured under the enrichment conditions specified for the test (selective enrichment if this is part of the procedure). Non-target strains are cultured under non-selective conditions to present a worst-case scenario.

   For inoculation of food samples, a different target strain is used for each food type. In the early years, there was an attempt to pair foods and strains based on historical illness outbreak or product recall data (e.g., *Salmonella* Enteritidis in eggs).

   These issues will be addressed by the BPMM Task Force as proposed guidelines are drafted.

2. Number of foods
   Historically, for *Salmonella*, normally 20 foods tested in a pre-collaborative study (or AOAC PTM study) plus 6 foods in a collaborative are viewed as sufficient to support claims for "all foods". Some within the AOAC review community are becoming uncomfortable with this, since it has been found subsequently that some methods approved for all foods have been shown to be ineffective for certain foods not tested in their validation studies. There has been talk of limiting approval claims to those foods actually tested. Restricted approvals are less helpful to the end user of the test, but obviously more accurately indicative of what is known about the test capabilities. It might be better to warn users that the procedure has been tested only in certain circumstances and users should be required to advise AOAC of genuine false negative or false positive findings obtained in subsequent investigations; this should then lead to the issue of an administrative warning concerning use of the procedure in such circumstances.

   The approach to *Listeria* has been a bit different. Normally 15 foods are tested, and claims are issued for product groups – meats and poultry, seafoods, dairy products, fruits and vegetables, environmental samples, etc.

   For *E. coli* O157:H7, the list of foods of interest is much narrower, generally limited to raw beef and perhaps sprouts and freshly pressed apple juice.

The BPMM task force does not prescribe a specific number of foods, but tends toward allowing the method developer to make a claim based only on those foods successfully validated. The Matrix Extension group has developed new food categorization schemes with associated rules for matrix extension (see Appendix B).

3. Inoculation levels
   The current AOAC requirement is for at least one level with 20 replicates where the results produce "fractional positives", i.e., less than 20 are positive by at least one of the methods (test or reference). This is taken as evidence that the majority of test samples contain not more than 1 cfu per 25 grams. In order to produce this result, analyst inoculates at ~ 1 cfu per 25 grams of product, sometimes higher if inoculum die-off is expected. The actual inoculation levels are estimated by an MPN determination. There is also a requirement for at least 5 uninoculated control samples. Normally, analysts prepare 2 levels of inoculation in the hope that at least one of the levels will produce fractional positives. The BPMM Statistics working group recommends that 4 levels of inoculation be used in estimating the LOD50: two levels with fractional recovery, one level all or nearly all positive and one level all or nearly all negative.

4. Choice of reference method
   Currently, in AOAC *Official Methods*SM pathogen test studies, the FDA BAM (*Bacteriological Analytical Manual*, available on-line), USDA MLG (Microbiological Laboratory Guidebook, available on-line) or appropriate AOAC *Official Method of Analysis*SM method is used as the reference procedure. In AOAC Research Institute *Performance Tested Methods*SM validations, other recognized official methods can be used, such as ISO or Health Canada. All of these are standard microbiological culture methods. Depending on the choice of reference method, the alternative method could compare differently. While the BPMM task force has not recommended any change in the choice of reference method, it does offer validation methods appropriate for cases in which no reference method is available. When a reference method is available, it will be included in the validation study and its performance will be evaluated alongside the alternative method.

5. Manufacturer of culture media
   Typically, the culture media used in a validation study is stated in the validation study report. However, the media source is generally not specified in terms of the method approval, i.e., there is a general doctrine of equivalence. This is a tough problem, because we know that there are differences in media performance from manufacturer-to-manufacturer and also lot-to-lot, which cannot be controlled. It is appropriate to note in the published AOAC method that medium from Manufacturer X was used in the validation procedure (which cover several batches of that medium) and that any laboratory which intends to use the procedure with media from other manufacturers must first check the alternative medium against that from Manufacturer X.

6. Manufacturer of reagents
   AOAC Official Methods of Analysis are supposed to be described in a generic way. So, in theory, someone could copy the method described for a commercial kit and claim that the copy is an AOAC method. This situation has not arisen, though. There usually is not enough information given about critical reagents such as DNA probes or antibodies to allow someone to easily replicate a method. AOAC Research Institute PTM approvals are specific to the individual commercial kit and annual reviews ensure that if reagent sources change, that a method modification study is performed to demonstrate equivalence.

7. Physical state of the cells
   In the ideal case, validation studies would only use naturally contaminated samples. However, these are rarely available, especially low-moisture foods contaminated with *Salmonella*. So, in the inoculated studies, the analyst stresses the inoculum and mixes it into the matrix so as to simulate conditions of natural contamination. For low-moisture foods, the model is to use a lyophilized cell pellet as the inoculum, mix it into the food, and allow the food to sit for 14 days to "stabilize". An MPN determination is then done to estimate the contamination level. For frozen foods, the food is thawed, inoculated with a culture dilution, and then re-frozen for 3 days before testing. In some cases, models of heat or preservative injury might be appropriate. For high-moisture refrigerated foods (e.g., raw or cooked chicken), the food is inoculated with a culture dilution and then refrigerated for 3 days before testing. There may be little or no injury in this case, but the thinking is that this realistically simulates the natural state of pathogen contamination of this type of product. No change to these procedures is currently recommended by the BPMM task force.

8. Phenotype vs. genotype
   Differences between molecular-based methods and immunological methods must be taken into consideration in method development and the design of validation studies. While molecular methods do not require expression of protein products, immunological methods must ensure that expression occurs at sufficient levels for detection to occur. Factors such as matrix, enrichment media, and enrichment temperature can potentially influence protein expression.

9. Sporulation
   No comments

# Uncertainty Associated with Microbiological Analysis

## 1. Introduction

1.1. There are only two absolute certainties in life: death and taxes! Whatever task we undertake, no matter how menial or how sophisticated, we are faced with a lack of certainty in the outcome! It is therefore essential to have a common understanding of what is meant by uncertainty in relation to our specific tasks in defining BPMM.

1.2. In microbiological laboratory practice, we can identify many causes of variability, for instance:

1.2.1. The ability of an isolate to give typical reactions on a diagnostic medium;

1.2.2. The use of the incorrect ingredients in a culture medium;

1.2.3. The consequence of changing brands of commercial media;

1.2.4. Use of non-standard conditions in the preparation, sterilisation and use of a culture medium;

1.2.5. Equipment and human errors in weighing, dispensing, pipetting and other laboratory activities;

1.2.6. The tolerance applied to the shelf life of test reagents;

1.2.7. The relative skill levels of different technicians;

1.2.8. The relative well-being of any technician who is undertaking analyses;

1.2.9. and so on, and so on …. *ad infinitum!*

1.3. These are but a few trite examples of biological, instrumental and personal bias that affect the accuracy, precision and hence the uncertainty of microbiological tests; a situation that constantly faces scientists involved in laboratory management.

1.4. To interpret properly the results obtained using any analytical procedure, whether physical, chemical or biological, requires careful consideration of the diverse sources of actual or potential error associated with the results obtained. Any analytical result is influenced by a complex of three major error groups:

1.4.1. *Random errors*, associated with the original sample matrix, the analytical (test) sample, the culture media, etc;

1.4.2. Inherent *systematic errors* associated with the analytical procedure; and

1.4.3. Modification of the systematic errors due to a particular laboratory's environment and equipment together with individual analysts' personal traits in carrying out the test procedure.

1.5. **Accuracy and Precision**

1.5.1. Accuracy is a qualitative concept (VIM, 1993). In simple terms, accuracy can be defined as the correctness of a result, relative to an expected outcome; whilst precision is a measure of the variability of test results.

1.5.2. Accuracy is defined (ISO3534-2:2003) as "*the closeness of agreement between a test result or a measurement result and the true value*." Accuracy is a combination of trueness and precision (a combination of random components and systematic error or bias components). This differs from the definition given by VIM (1993): "*the closeness of agreement between the result of a measurement and a true value of a measurand*".

1.5.3. "***Accuracy***" is essentially "absence of error"; the more accurate a result the lower the associated error of the test. It is important to note that the term "accuracy " applies only to results and can not be applied to methods, equipment, laboratories or other general matters.

1.5.4. "***Trueness***" is defined (ISO, 2003) as, "the closeness of agreement between the average value obtained from a large series of test results and an accepted reference value".

1.5.5. *Trueness* is equivalent to an absence of "***bias***", which is the difference between the expectation of the test results and an accepted reference value and is a measure of total systematic, but not random, error.

1.5.6. *Trueness*, unlike *accuracy*, may correctly be contrasted with *precision.*

1.6. "***Precision"*** is defined as the closeness of agreement between independent test results obtained under stipulated conditions.

1.6.1. *Precision* depends only on the distribution of random errors and does not relate to a true value or a specified value.

1.6.2. The *measure of precision* is expressed usually in terms of imprecision and computed as a standard deviation of the test results.

1.6.3. Lower precision is reflected by a larger standard deviation.

1.6.4. *Independent test results* means results obtained in a manner not influenced by any previous results on the same or similar test object.

1.6.5. *Quantitative measures of precision* depend critically on the stipulated conditions. Repeatability and reproducibility conditions are particular sets of extreme stipulated conditions (ISO 3534: 3.14).

1.7. Fig 1 illustrates schematically the relationships between trueness, accuracy, precision and uncertainty (AMC, 2003).

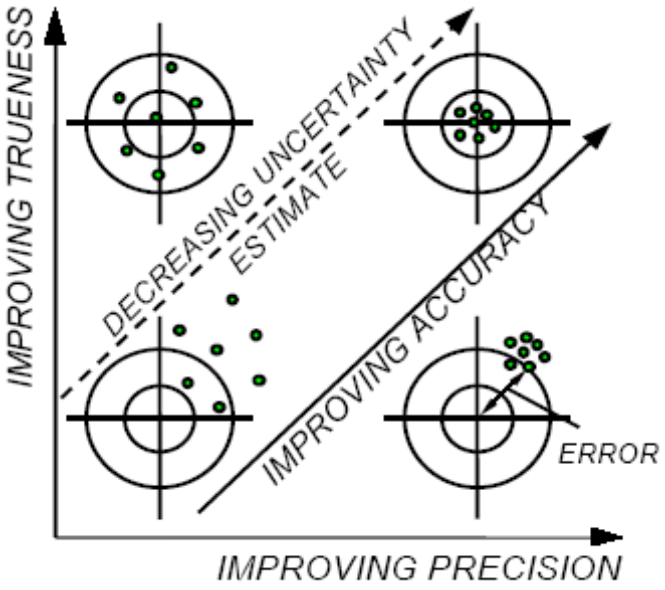



***Fig 1. Relationships between trueness, accuracy, precision and uncertainty in analytical results (AMC, 2003).*** (Reproduced by permission of the Royal Society of Chemistry, London)

1.8. The concepts of accuracy and trueness must take account of error and precision. Uncertainty estimates (*qv*) provide a simple way to quantify such needs. However, since in a real-life situation we never know what the "true" or "correct" answer is, trueness can be assessed only in a validation-type trial against an accepted reference value. This is much more complex in microbiology than it is in physics, and chemistry.

## 2. Uncertainty of Measurement

2.1. The ISO/Eurachem (2000) definition of **Uncertainty of a Measurement** is

2.2. "*A parameter associated with the result of a measurement that characterises the dispersion of the values that could reasonably be attributed to the measurand*". The term "measurand" is a bureaucratic way of saying "analyte".

2.3. Translated into simple English this definition can be rewritten, as "*Uncertainty is a measure of the likely range of values that is indicated by an analytical result*."

2.4. For quantitative data (e.g. colony counts, MPNs or $LOD_{50}$ values) a measure of uncertainty may be any appropriate statistical parameter associated with the test result. Such parameters include the standard deviation, the standard error of the mean or a confidence interval around that mean.

2.5. Measures of repeatability and reproducibility are the corner stones of estimation of analytical uncertainty. They are defined (ISO 2004) as:

2.5.1. *Repeatability* is "a measure of variability derived under specified ***repeatability conditions"***, i.e. independent test results are obtained with the same method on identical test items in the same laboratory by the same analyst using the same equipment, batch of culture media and diluents, and tested within short intervals of time.

2.5.2. *Reproducibility* is "a measure of precision derived under ***reproducibility conditions"*** i.e. test results are obtained with the same method on identical test items in different laboratories with different operators using different equipment. A valid statement of reproducibility requires specification of the conditions used.

2.5.3. **Intermediate Reproducibility** (ISO 5725-2:1994 ) is defined as "a measure of reproducibility derived under reproducibility conditions within a single laboratory".

2.5.4. **Standard Uncertainty** of a measurement ($u(y)$) is defined (GUM, 2000) as "the result obtained from the values of a number of other quantities, equal to the positive square root of a sum of terms, the terms being the variances or covariances of these other quantities weighted according to how the measurement result varies with changes in these quantities"

2.5.5. **Expanded Uncertainty** ($U$) is defined as "the quantity defining an interval about a result of a measurement expected to encompass a large fraction of the distribution of values that could reasonably be attributed to the measurand".

2.5.6. The "*Expanded Uncertainty"* values are derived by multiplying the SD's with a "*coverage factor*" to provide confidence intervals for repeatability and reproducibility around the mean value. Routinely, a coverage factor of 2 is used to give approximate 95% distribution limits (confidence interval) around the "normalised" mean value.

2.6. For qualitative data (e.g. presence or absence tests) uncertainty measures cannot be derived in the same way. However, other procedures e.g. use of the standard error associated with derived values for e.g. $LOD_{50}$ (*qv*) and by binomial analysis of the relative proportions of positive and negative results in a comparative evaluation of methods (see 3.4 below).

## 3. How is uncertainty estimated?

3.1. There are two totally different approaches to the estimation of uncertainty:

3.1.1. The "bottom up" approach in which the errors associated with **all** the relevant steps undertaken during an analysis are used to derive a value for the "*combined standard uncertainty*" associated with a method (Eurachem 2000; Niemelä, 2002). Essentially this approach provides a broad indication of the possible level of uncertainty associated with method rather than a measurement; ISO TC34 SC9 considers the approach always to under-estimate the extent of variation since it cannot take into account either matrix-associated errors or the actual day-to-day variation seen in a laboratory. For

these reasons, ISO has recommended that this approach is not appropriate for microbiological analyses.

3.1.2. The "top-down" approach is based on statistical analysis of data generated in intra- or inter-laboratory collaborative studies on the use of a method to analyze a diversity of matrixes. It therefore provides *an estimate of the uncertainty of a measurement* associated with the use of a specific method.

3.1.3. Statistical aspects of the procedures, together with worked examples, for both approaches are summarised in Annexes I & II.

3.1.4. A review of measurement uncertainty in quantitative microbiological analysis is currently in press (Corry et al, 2006).

3.2. ***Quantitative Tests.*** For quantitative data (e.g. colony counts and MPN estimates), measures of "repeatability" and "reproducibility" are derived as the standard deviations of repeatability ($s_r$) and reproducibility ($s_R$). However…

3.2.1. Microbiological data do not normally conform to a "normal" distribution, and usually require mathematical transformation prior to statistical analysis. For most purposes, a $\log_{10}$ transformation is used to "normalise" the data but in cases of significant over-dispersion the use of a negative-binomial transformation may be necessary (Jarvis, 1989; Niemelä, 2002). If there is reason to believe that data conform to a Poisson distribution, then a square root transformation is required, since the variance ($\sigma^2$) is numerically equal to the mean ($m$) value.

3.2.2. Statistical analyses of collaborative trial data are generally done by Analysis of Variance (ANOVA) after removing any outlying values, as described by Youden & Steiner (1975) and by Horwitz (1995). However, it has been argued (e.g. AMC 1989, 2001) that it is wrong to eliminate outlier data and that application of Robust Methods of analysis is preferable.

3.2.3. One approach to robust analysis is a "robusticised" ANOVA procedure based on Huber's H15 estimators for the robust mean and standard deviation of the data (AMC, 1989, AMC 2001, ISO 5725-5:1998).

3.2.4. An alternative approach is that of the Recursive Median (REMEDIAN) procedure (ISO 2000; Wilrich, 2005).

3.2.5. Worked examples of traditional and robust analyses are shown in Annexe III.

3.2.6. A major drawback to use of these robust techniques for inter-laboratory trials is that they do not permit the derivation of Components of Variance. A novel approach to overcome this disadvantage is by the use of stepwise robust analysis for "nested" trial data, as described by Hedges & Jarvis (2006).

3.3. ***Intermediate Reproducibility of Quantitative Tests.*** Similar procedures may be used to estimate intermediate (intra-laboratory) reproducibility associated with the use of an analytical procedure in a single laboratory. Even data obtained, for instance, in laboratory quality monitoring can be used to provide an estimate of intra-laboratory

reproducibility. ISO/PTDS 19036:2005 (Part 6) describes a statistical procedure for analysis of paired data. A worked example is shown in Annex IV.

3.4. ***Qualitative Tests***. Estimation of uncertainty associated with qualitative (e.g. presence or absence) methods has not been well documented and is currently the subject of discussion within ISO.

3.4.1. Many of the potential errors that affect quantitative methods also affect qualitative methods; but there are also some additional potential errors that are inherent in the analytical procedure. For example:

3.4.1.1. In taking a sample for analysis, it is of critical importance to have knowledge of the probable distribution of organisms in the test matrix, especially when testing for organisms at the limit of detection of a method. Whilst it *may* be possible to ensure reasonable conformity with a Poisson (random) distribution of index organisms in artificial test matrixes, such distribution should not be assumed to occur in natural matrixes and requires confirmation (e.g. using an Index of Dispersion Test such as that described by Fisher et al, 1922) before using such matrixes in collaborative studies. In real life testing, erroneous decisions can result from an assumption that all microorganisms are distributed randomly at low level – there are some well-documented examples where "over dispersion" of organisms (e.g. due for instance to clumping) has resulted in a significant level of genuine false negative surveillance data.

3.4.1.2. There is an intrinsic need to ensure effective growth of the index organism to critical levels during all pre-enrichment, enrichment and differential/diagnostic culture stages – so culture medium composition, incubation times & temperatures, etc are critical to the success of the test.

3.4.1.3. It is critical to ensure that the confirmatory stages of a test protocol do actually identify the index organism.

3.4.1.4. Knowledge of the potential effect of competitive organisms is of major importance for all cultural and confirmatory stages of a test protocol.

3.4.1.5. The decision on use of either true pairs or non-paired samples is of great importance in the interpretation of potential false negative or false positive results for method validation studies.

3.4.2. The output of qualitative tests is a series of positive and negative responses. One approach to seeking to quantify such data was the derivation of the Accordance and Concordance concept (Langton et al, 2002) that sought to provide measures "equivalent to the conceptual aspects of repeatability and reproducibility". However, it is now considered that this approach is not sufficiently robust to be used in the manner proposed and adds no value to the original data.

3.4.3. Provided that a sufficient number of parallel tests has been undertaken at each of several levels of potential contamination, then it is possible to quantify the

test responses in terms of an estimated Level of Detection for (e.g.) 50% positives [$LOD_{50}$](for details see Hitchins, 2005).

3.4.3.1. This statistical approach essentially estimates the Most Probable Number of organisms at each test level and then analyses the relative MPN values using the Spearman-Karber approach.

3.4.3.2. Alternative approaches including Probit and Logit analyses may also be appropriate in specific circumstances.

3.4.3.3. What these methods have in common is an ability to transform purely qualitative data into a quantitative format for which error values can be derived so permitting an estimate of the uncertainty of the test result.

3.4.3.4. An extrapolation of the approach would be to determine also the $LOD_0$ and $LOD_{90}$ values such that a dose-response curve can be derived. This may be of importance in differentiating between methods capable of detecting specific organisms at a similar $LOD_{50}$ level but for which the absolute limit of non-detection ($LOD_0$) and a selected higher limit of detection (e.g. $LOD_{90}$) differ.

3.4.3.5. An alternative approach is to estimate the uncertainty associated with the proportions of test samples giving a positive response, based on the binomial distribution.

3.4.4. Examples of the way in which such approaches to analysis of qualitative data can be used are illustrated in Annex V.

## 4. Reporting of Uncertainty

4.1. The expression of uncertainty is of some importance in interpretation of data. Assuming a mean aerobic colony count (ACC) = 5.00 ($log_{10}$) cfu/g and a reproducibility standard deviation of ± 0.25 ($log_{10}$) cfu/g, then the expanded uncertainty is given, for instance, by:

4.1.1. Aerobic colony count on product X is 5.00 ± 0.50 ($log_{10}$) cfu/g; or

4.1.2. Aerobic colony count on product X is 5.00 ($log_{10}$) cfu/g ± 10%

4.2. It is important not to refer to analytical methods as having a precision of e.g. ± 10% based on uncertainty estimates. Uncertainty is a measure of variability i.e. a measure of the lack of precision.

5. The use of uncertainty measures in assessing compliance of a test result with a defined criterion is of some importance and has been considered by the European Commission (Anon, 2003). Jarvis et al (2004) and Jarvis & van der Voet (2005) have discussed the interpretation of data in relation to microbiological criteria for foods.

For more information, please contact Basil Jarvis at basil.jarvis@btconnect.com.

## 6. References cited

Analytical Methods Committee (1989) Robust Statistics – How Not to Reject Outliers. Part 1: Basic concepts, *The Analyst* **114**, 1693 – 1697. Part 2: Inter-laboratory trials, *The Analyst* **114**, 1699 – 1702.

Analytical Methods Committee (2001) Robust statistics: a method of coping with outliers. AMC Brief No.6, Royal Society of Chemistry, London.

Analytical Methods Committee (2003) Terminology - the key to understanding analytical science. Part 1: Accuracy, precision and uncertainty. AMC Brief No.13, Royal Society of Chemistry, London.

Anon (2003) The relationship between analytical results, the measurement uncertainty, recovery factors and the provisions in EU food and feed legislation. Report to the EU Standing Committee on the Food Chain and Animal Health Working Group Draft, 5 June 2003

Corry, J, Jarvis, B, Passmore, S and Hedges, A (2006) A critical review of measurement uncertainty in the enumeration of food microorganisms. *Food Microbiology* (in press)

Eurachem (2000) *Quantifying Uncertainty in Analytical Measurement*. 2nd edition, Laboratory of the Government Chemist, London.

Fisher, R A , Thornton, H G & Mackenzie, W A (1922) The accuracy of the plating method of estimating the density of bacterial populations. *Annals Applied Biology,* **9**, 325 – 359.

Hedges, A & Jarvis, B (2006) Application of robust methods to the analysis of collaborative trial data using bacterial colony counts.. *J Microbiological Methods (in press).*

Hitchins, A J (2005) Proposed Use of a 50 % Limit of Detection Value in Defining Uncertainty Limits in the Validation of Presence-Absence Microbial Detection Methods. *BPMM Report for the Statistics WG* (010705).

Horwitz, W (1995) Protocol for the design, conduct and interpretation of method performance studies. *Pure & Applied Chemistry*, **67**, 331 – 343.

ISO 5725-2:1994: Accuracy (trueness and precision) of measurement methods and results – Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method. *International Standardisation Organisation, Geneva*

ISO 5725-5:1998: Accuracy (trueness and precision) of measurement methods and results – Part 5: Alternative methods for the determination of the precision of a standard measurement method. *International Standardisation Organisation, Geneva*

ISO FDIS 16140:2000 Microbiology *of food and* animal *feeding stuffs – Protocol for the validation of alternative methods. International Standardisation Organisation, Geneva.*

ISO 3534-1:2003 Statistics – Vocabulary and Symbols. *International Standardisation Organisation, Geneva.*

ISO 16140:2003 Microbiology of food and animal feeding stuffs — Protocol for the validation of alternative methods. *International Standardisation Organisation, Geneva.*

ISO TS 21748:2004 Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty estimation *International Standardisation Organisation, Geneva*

ISO PTDS 19036:2005 Microbiology of food and animal feeding stuffs –Guide on estimation of measurement uncertainty for quantitative determinations*. International Standardisation Organisation, Geneva*

Jarvis, B (*1989*) *Statistical analysis of the microbiological analysis of foods.* Progress in Industrial Microbiology Vol 21. Elsevier, Amsterdam.

Jarvis, B (2000) Sampling for Microbiological Analysis. In The Microbiological Safety and Quality of Food. Ed B M Lund, A C Baird-Parker & G W Gould. Vol II, pp.1691 – 1733. Aspen Pub Inc, Gaithersburg, MA.

Jarvis, B, Hedges, A, Corry, J E L & Wood, R (2004) Certainty or Uncertainty? The impact of uncertainty on the interpretation of colony count data in relation to microbiological criteria*. Poster presented at the Society for Applied Microbiology conference on "Food and Dairy Microbiology", Cork, July 2004.*

Jarvis, B & van der Voet, H (2005) Guidelines on the use of uncertainty measurements in the assessment of data for compliance with quantitative microbiological criteria for foods. Draft working paper for ISO TC34 SC9 Statistics WG.

LaBarre, D D, Zelenke, D & Flowers, R (2005) Intra-laboratory and Inter-laboratory Variability. *BPMM Sampling WP document* (draft 5 – 6/9/05)

Langton, S.D, Chevennement, R, Nagelkeke, N, Lombard, B. (2002). Analysing collaborative trials for qualitative microbiological methods: accordance and concordance. *International Journal of Food Microbiology* **79**, 171-181.

Niemelä, S.I, 2002 Uncertainty of quantitative determinations derived by cultivation of microorganisms. 2nd edition, Centre for Metrology and Accreditation, Advisory Commission for Metrology, Chemistry Section, Expert Group for Microbiology, Helsinki, Finland, Publication J3/2002

Niemelä, SI (2003) Measurement uncertainty of microbiological viable counts. *Accreditation and Quality Assurance*, 8: 559-563.

NMKL (2002). Measurement of uncertainty in microbiological examination of foods. NMKL *Procedure no 8, 2nd ed., Nordic Committee on Food Analysis*

Rousseeuw, P. J., and Croux, C. (1993): Alternatives to the median absolute deviation. *J. Am. Stat. Ass*. **88**, 1273-1283

SMITH, P.A. & KOKIC, P. (1996) Winsorisation in ONS business surveys. *Working paper no. 22 at the UN Data Editing Conference 1996, Voorburg.*

Van der Voet, H (2005) Alternative models for measurement uncertainty of microbiological count data. Paper presented to ISO TC34 SC9 Statistics WG meeting, Parma, April 2004.

VIM (1993) International vocabulary of basic and general terms in metrology. (ISO 1993)

Wilrich, P-T (2005a) Robust estimates of the theoretical standard deviation to be used in inter-laboratory precision experiments. Discussion paper for the ISO TC34 SC9 Statistics Group Meeting, Paris, April 2005.

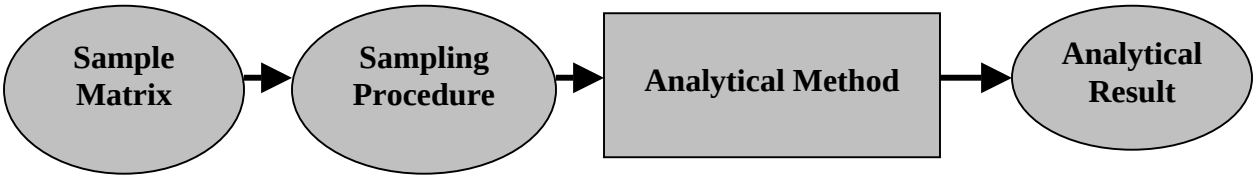
Wilrich, P-T (2005b) The determination of precision of measurement methods with qualitative results by interlaboratory experiments. Discussion paper for the ISO TC34 SC9 Statistics Group Meeting, Paris, April 2005.

Youden, W J and Steiner, E H (1975) *Statistical Manual of the AOAC*. AOAC, Washington.

# Annex I

## Top-Down Procedure For Estimation Of Uncertainty

1. The basis of the "top down" approach described by GUM (Eurachem 2000) is to identify and take account of all procedural stages of an analytical method. The variance associated with each individual stage is combined with the variances all the other stages and interactions that make up an analytical procedure in order to estimate a generic level of uncertainty for a method. This is illustrated diagrammatically in the schematic below.




2. Consider first the sample matrix: what are the likely errors that will affect the analytical result?

2.1. The largest potential error sources will be: the spatial distribution of the microorganisms (random, under- or over dispersion as exemplified by evidence of clumping); the condition of the microorganisms (viable and vital, sublethally damaged, non-cultivatable); the effects of competitive organisms on the recoverability of specific types; whether the organisms are located primarily on the surface of, or more generally distributed throughout, the matrix; etc.

2.2. However, the intrinsic nature of the matrix will also affect the results of an analysis.

3. How representative is an analytical sample taken from a matrix?

3.3. Should the analytical sample be totally representative of the whole matrix, or should it relate only to a specific part, e.g. the surface of a meat carcass? If the former should the matrix be homogenized prior to taking a sample; if the latter should the surface layer be excised, swabbed, rinsed or tested using a replica plating technique? What ever the method of sampling to what extent is the microflora in the analytical sample representative of both the number and types of microorganisms present in the original matrix.

3.4. If the matrix is a composite food, should the sample represent the whole or individual parts of the food matrix (e.g. in the case of a meat pie should the pastry and the meat be analysed separately)?

3.5. What size of sample should be tested? Increasing the size of an analytical sample results in a decrease in the standard error associated with the mean weight of sample taken. Similarly, increasing the weight of sample taken tends to increase the apparent colony count whilst reducing the overall variance of the mean count (Jarvis, 1989).

4. At its simplest, the analytical process consists of taking an analytical sample, suspending that sample in a defined volume of a suitable primary diluent, macerating the sample, preparing serial dilutions, plating measured volumes onto or into a culture medium, incubating the plates, counting and recording the numbers of colonies and deriving a final estimate of colony forming units (cfu) in the original matrix. At all stages throughout this process, errors will occur.

4.1. Some errors, e.g. those associated with the accuracy of weighing, the accuracy of pipette volumes, the accuracy of colony counting, etc, etc can be quantified and measures of the variance can be derived.

4.2. Some errors can be assessed, but not necessarily quantified; for instance, laboratory quality control procedures can be used to assess the extent to which a culture medium will support the growth of specific organisms. Such data may potentially provide a correction factor for the yield of organisms on a particular culture medium; whether or not the use of a correction factor should be employed in microbiological practice is a matter of debate!

4.3. However, other errors, such as those associated with individual technical performance on a day, cannot be quantified.

5. Some analytical errors associated with microbiological practices are possibly not significant when compared to other errors, but how do you know this if the errors cannot be quantified? To assess the uncertainty of an analytical microbiological procedure from the "top down" requires a full evaluation of **all** potential sources of error for each and every stage of an analytical procedure.

6. Estimation of the standard uncertainty of an analytical procedure, once a reliable schedule of quantifiable errors has been produced, is done simply by combining the errors:

$$s_R^2 = s_a^2 + s_b^2 + ..... + s_x^2 + s_y^2 + s_z^2$$

where $s_R^2$ = reproducibility variance of the method and $s_{a....z}^2$ = variance of any stage (a….z) within the overall method.

By definition, the reproducibility standard deviation ( $s_R$ ) is derived from the square root of the variance:

$$s_R = \sqrt{s_a^2 + s_b^2 + ..... + s_x^2 + s_y^2 + s_z^2}$$

7. The expanded uncertainty is derived by multiplying the standard uncertainty by a coverage factor *k*, which has a value from 2 to 3. A value of 2 is normally used to give approximate 95% confidence limits; hence

$$U = k.\ s_R = 2.\ s_R$$

8. Niemelä (2002, 2003) gives a more detailed explanation of the "top down" approach to assessment of measurement uncertainty in microbiological analysis.

# Annex II

## "Bottom-up" Approach to Estimation of Uncertainty

1. Traditionally, the parameters used to derive uncertainty measures are estimated from the pooled results of a "valid" inter-laboratory collaborative study, or in the case of intermediate reproducibility, from an intra-laboratory study. Appropriate procedures to ensure that the study design is valid have been described *inter alia* by Youden & Steiner (1975) and by ISO (1994, 1998).

2. The data from all participating laboratories are subjected to analysis of variance (ANOVA) after first checking for:

2.1. Conformance with a "normal distribution" either by plotting the data or by application of appropriate tests for "normality".

2.2. Identification and removal of "outliers" using the methods described by Youden and Steiner (1975) or Horwitz (1995), followed if necessary by repeating the tests for conformance with "normality".

3. Quantitative microbiological data (e.g. colony counts and MPNs) do not conform to a normal distribution and require transformation to "normalise" the data before analysis.

4. Transformations are done by converting each of the raw data values ($x_i$) into the $\log_{10}$ value ($y_i$) where $y_i = \log_{10} x_i$. Strictly, it is more correct to use the natural logarithmic transformation (i.e. $y_i = \ln x_i$) (van der Voet, 2004).

5. For low level counts (typically < 100 cfu/g) that conform to the Poisson distribution (mean value ($m$) = variance ($s^2$)), the data are transformed by taking the square root of each data value (i.e. $y_i = \sqrt{x_i}$).

6. However, because of problems of over-dispersion frequently associated with microbial contamination, it may be preferable to test for (or to assume) conformance with a negative binomial distribution. Some statistical packages (e.g. Genstat) include a facility to make this transformation (using the Maximum Likelihood Method programme RNEGBINOMIAL), but such procedures are not universally available and it can be very time-consuming to calculate manually (Jarvis, 1989; NMKL, 2002, Niemelä, 2003; van der Voet, 2004).

7. Assuming a fully "nested" experimental design (e.g. duplicate testing of duplicate samples by "*A*" analysts in each of "*L*" laboratories), the residual mean variance (i.e. the variance of the replicated analyses on each sample) of the ANOVA provides an estimate of repeatability variance ($s_r^2$). The estimate of reproducibility variance ($s_R^2$) first requires computation of the contributions to variance of the samples, analysts and laboratories. This is illustrated below.

8. The repeatability standard deviation ($s_r$) and the reproducibility standard deviation ($s_R$), being the square root values of the respective variances, are the measures of standard uncertainty from which the expanded uncertainty estimates are derived.

9. ***Statistical Procedure to Derive Component Variances from an ANOVA Analysis***

*Assume*: trial consists of (*p*) laboratories (*p=20*) in each of which 2 analysts test 2 replicate samples and make duplicate analyses of each sample. Hence, each laboratory carries out 8 replicate analyses and the total number of analyses = 8p = 160.

Each data value ($y_{pijk}$) is allocated to a cell in the data table in the sequence laboratory (*p*), analyst (*i*), sample (*j*) and replicate (*k*), as shown below, and are then analysed by multivariate analysis of variance.

| Laboratory (*p* = *1…20*) | Analyst (*i =1*) | | | | Analyst (*i = 2*) | | | |
|---|---|---|---|---|---|---|---|---|
| | Sample (*j = 1*) | | Sample (*j = 2*) | | Sample (*j = 1*) | | Sample (*j = 2*) | |
| | Replicate (*k = 1*) | Replicate (*k = 2*) | Replicate (*k = 1*) | Replicate (*k = 2*) | Replicate (*k = 1*) | Replicate (*k = 2*) | Replicate (*k = 1*) | Replicate (*k = 2*) |
| 1 | $y_{1111}$ | $y_{1112}$ | $y_{1121}$ | $y_{1122}$ | $y_{1211}$ | $y_{1212}$ | $y_{1221}$ | $y_{1222}$ |
| 2 | $y_{2111}$ | $y_{2112}$ | $y_{2121}$ | $y_{2122}$ | $y_{2211}$ | $y_{2212}$ | $y_{2221}$ | $y_{2222}$ |
| 3 | $y_{3111}$ | $y_{3112}$ | $y_{3121}$ | $y_{3122}$ | $y_{3211}$ | $y_{3212}$ | $y_{3221}$ | $y_{3222}$ |
| 4 | $y_{4111}$ | $y_{4112}$ | $y_{4121}$ | … | … | … | … | … |
| … | … | … | … | … | … | … | … | … |
| … | … | … | … | … | … | … | … | … |
| 20 | $y_{20111}$ | $y_{20112}$ | $y_{20121}$ | $y_{20122}$ | $y_{20211}$ | $y_{20212}$ | $y_{20221}$ | $y_{20222}$ |

## ANOVA table for a four-factor fully-nested experiment

| Source of Variation | Sum of Squares | Degrees of freedom | Mean Square | Expected Mean Square Components* |
|---|---|---|---|---|
| Laboratories | $SS_{lab}$ | p-1 = 19 | $SS_{lab}/19 = MS_{lab}$ | $\sigma_r^2 + 2\sigma_{sam}^2 + 4\sigma_{ana}^2 + 8\sigma_{lab}^2$ |
| Analysts | $SS_{ana}$ | p = 20 | $SS_{ana}/20 = MS_{ana}$ | $\sigma_{\boldsymbol{r}}^2 + 2\sigma_{\boldsymbol{sam}}^2 + 4\sigma_{\boldsymbol{ana}}^2$ |
| Samples | $SS_{sam}$ | 2p = 40 | $SS_{sam}/40 = MS_{sam}$ | $\sigma_r^2 + 2\sigma_{sam}^2$ |
| Residual | $SS_{res}$ | 4p = 80 | $SS_{res}/80 = MS_{res}$ | $\sigma_r^2$ |
| Total | Total SS | 8p –1=159 | | |

* The components are shown as population variances since this is an expectation table.

The residual mean square ($MS_{res} = s_r^2$) provides the repeatability variance between duplicate analyses done on the same replicate sample.

The variance due to l samples ( $s^2_{sam}$ )is given by $[MS_{sam} - s^2_r]/2$

The variance due to analysts ( $s^2_{ana}$ ) is given by $[MS_{ana} - 2s^2_{sam} - s^2_r]/4$

The variance due to is laboratories ( $s^2_{lab}$ ) given by $[MS_{lab} - 2s^2_{sam} - 4s^2_{ana} - s^2_r]/8$

The Reproducibility Variance ( $s^2_R$ ) is given by $[s^2_{sam} + s^2_{ana} + s^1_{lab} + s^2_r]$

The Reproducibility Standard Deviation is given by $\sqrt{s^2_{sam} + s^2_{ana} + s^1_{lab} + s^2_r}$

The Repeatability Standard Deviation is given by $\sqrt{s^2_r}$ .

**WORKED EXAMPLE** (10 Labs x 2 Analysts x 2 Samples x 2 Replicate analyses)

***Log transformed colony counts ($Log_{10}$ cfu/g)***

| Laboratory | Analyst (*i = 1*) | | | | Analyst (*i = 2*) | | | |
|---|---|---|---|---|---|---|---|---|
| | Sample (*j=1*) | | Sample (*j=2*) | | Sample (*j=1*) | | Sample (*j=2*) | |
| | Replicate (*k = 1*) | Replicate (*k = 2*) | Replicate (*k = 1*) | Replicate (*k = 2*) | Replicate (*k = 1*) | Replicate (*k = 2*) | Replicate (*k = 1*) | Replicate (*k = 2*) |
| 1 | 5.56 | 5.73 | 5.76 | 5.59 | 6.08 | 5.96 | 6.07 | 5.99 |
| 2 | 6.02 | 5.88 | 5.87 | 5.80 | 5.54 | 5.63 | 5.92 | 5.79 |
| 3 | 6.26 | 6.30 | 6.46 | 6.54 | 6.42 | 6.49 | 6.11 | 6.42 |
| 4 | 5.07 | 5.11 | 4.90 | 4.61 | 4.63 | 4.81 | 4.42 | 4.56 |
| 5 | 5.39 | 5.25 | 5.28 | 5.52 | 5.34 | 5.46 | 5.47 | 5.49 |
| 6 | 5.98 | 5.88 | 6.02 | 5.64 | 5.96 | 6.06 | 5.70 | 5.57 |
| 7 | 5.43 | 5.18 | 5.16 | 5.08 | 6.15 | 5.76 | 5.44 | 5.43 |
| 8 | 5.94 | 5.73 | 5.28 | 5.47 | 5.99 | 6.01 | 5.92 | 6.13 |
| 9 | 5.45 | 5.35 | 5.49 | 5.42 | 5.68 | 5.57 | 5.74 | 5.69 |
| 10 | 5.51 | 5.74 | 6.18 | 6.13 | 5.83 | 5.91 | 5.76 | 5.60 |

Tests for normality (e.g. Shapiro-Wilk, W = 0.9830, p= 0.0885) did not disprove the hypothesis that the $\log_{10}$ transformed data conform reasonably (although not perfectly) to a normal distribution. However, application of the Cochran Test (Horwitz, 1995) identified Laboratory 7 as an outlier; subsequently evaluation using the Grubbs test did not eliminate other laboratories although laboratories 3 & 4 appeared to be possible outliers.

**ANOVA table for the four-factor fully nested experiment (All data included)**

| Source of Variation | Sum of Squares | Degrees of freedom | Mean Square (rounded to 4 places) | Mean Square Components |
|---|---|---|---|---|
| Laboratories | 12.636 | 9 | 1.4040 | $s_r^2 + 2s_{sam}^2 + 4s_{ana}^2 + 8s_{lab}^2$ |
| Analysts | 1.4906 | 10 | 0.1491 | $s_r^2 + 2s_{sam}^2 + 4s_{ana}^2$ |
| Samples | 1.346 | 20 | 0.0673 | $s_r^2 + 2s_{sam}^2$ |
| Residual | 0.5554 | 40 | 0.0139 | $s_r^2$ |
| Total | 16.0272 | 79 | | |

The residual mean square ($MS_{res} = s_r^2 = 0.0139$) provides the repeatability variance between duplicate analyses done on the same replicate sample.

*Component* Variances

Sample variance ( $s_{sam}^2$ ) = [$MS_{sam}$- $s_r^2$ ]/2 = [0.0673 – 0.01389]/2 = 0.0267

Analyst variance ( $s_{ana}^2$ ) = $[MS_{ana} - 2s_{sam}^2 - s_r^2]/4$ = [0.1491 – 0.0673]/4 = 0.02045

Laboratory variance ( $s_{lab}^2$ )= $[MS_{lab} - 2s_{sam}^2 - 4s_{ana}^2 - s_r^2]/8$ = [1.4040 – 0.1491]/8 = 0.1548

Hence, Reproducibility Variance ( $s_R^2$ ) = $[s_{sam}^2 + s_{ana}^2 + s_{lab}^1 + s_r^2]$ =

[0.0139 +0.0267 +0.02045 +0.15686] = 0.2179

Reproducibility Standard Deviation = $s_R = \sqrt{s_{sam}^2 + s_{ana}^2 + s_{lab}^1 + s_r^2}$ = √0.2179 = ±0.4668

Repeatability Standard Deviation = $s_r = \sqrt{s_r^2}$ = √0.01389 = ±0.1178

The mean colony count = 5.6682 ≈ 5.67 ($\log_{10}$) cfu/g

Hence, Relative Standard Deviation of Reproducibility ($RSD_R$)

= 100 x 0.4668/5.6682 = 8.24%

and, Relative Standard Deviation of Repeatability ($RSD_r$)

= 100 x 0.1178/5.6682 = 2.08%

From these values the 95% expanded uncertainty of reproducibility is given by:

$U = 2\ s_R$ = 2 x 0.4668 = ±0.9336. ≈ ±0.93 ($\log_{10}$) cfu/g

The upper and lower limits of the 95% Confidence Interval on the mean colony count are:

$U_L = 5.67 + 0.93 = 6.60$ ($\log_{10}$) cfu/g

$L_L = 5.67 - 0.93 = 4.74$ ($\log_{10}$) cfu/g

***Repeat analyses for 9 laboratories( after elimination of data for laboratory 7)***

| Source of Variation | Sum of Squares | Degrees of freedom | Mean Square (rounded to 4 places) | Mean Square Components |
|---|---|---|---|---|
| Laboratories | 12.227 | 8 | 1.5284 | $s_r^2 + 2s_{sam}^2 + 4s_{ana}^2 + 8s_{lab}^2$ |
| Analysts | 1.0249 | 9 | 0.1139 | $s_r^2 + 2s_{sam}^2 + 4s_{ana}^2$ |
| Samples | 1.0405 | 18 | 0.0578 | $s_r^2 + 2s_{sam}^2$ |
| Residual | 0.4449 | 36 | 0.0124 | $s_r^2$ |
| Total | 16.0272 | 71 | | |

The component variances were derived as:

Repeatability variance ($s_r^2$) = 0.0124 Sample variance ($s_{sam}^2$) = 0.0227

Analyst variance ($s_{ana}^2$) = 0.0140 Laboratory variance ($s_{lab}^2$) = 0.1168

Hence, Reproducibility Variance ($s_R^2$) = 0.2279

Reproducibility Standard Deviation = $s_R = \sqrt{0.2279} = \pm 0.4753$

Repeatability Standard Deviation = $s_r = \sqrt{0.0124} = \pm 0.1112$

The mean colony count = 5.6921 ≈ 5.69 ($\log_{10}$) cfu/g

Hence, Relative Standard Deviation of Reproducibility ($RSD_R$) = 8.35%

and, Relative Standard Deviation of Repeatability ($RSD_r$) = 1.95%

From these values the 95% expanded uncertainty of reproducibility is given by:

$U = 2\ s_R = 2 \times 0.4753 = \pm 0.9506. \approx \pm 0.95$ ($\log_{10}$) cfu/g

The upper and lower limits of the 95% Confidence Interval on the mean colony count are:

$U_L = 5.69 + 0.95 = 6.64$($\log_{10}$) cfu/g

$L_L = 5.67 - 0.95 = 4.72$ ($\log_{10}$) cfu/g

## Comparison of ANOVAs with and without removal of outlier laboratory

The table below shows that removal of one set of data (from the outlier laboratory) marginally increased the mean colony count and reduced the component variances for repeatability, samples, analysts and laboratories. However the overall effect, in this specific example, was marginal in relation to the derived values for repeatability and reproducibility; and hence there was little effect on the level of expanded uncertainty.

| Parameter | 10 Laboratories | 9 Laboratories |
|---|---|---|
| Mean Colony Count ($log_{10}$ cfu/g) | 5.6682 | 5.6921 |
| Repeatability Variance | 0.0139 | 0.0124 |
| Sample Variance | 0.0267 | 0.0227 |
| Analyst Variance | 0.0205 | 0.0140 |
| Laboratory Variance | 0.1548 | 0.1168 |
| SD repeatability ($SD_r$) | ±0.1178 | ±0.1112 |
| Relative $SD_r$ | 2.08% | 1.95% |
| SD reproducibility ($SD_R$) | ±0.4668 | ±0.4753 |
| Relative $SD_R$ | 8.24% | 8.35% |
| Expanded Uncertainty ($U$) | ±0.93 | ±0.95 |
| Upper Limit of 95% CI ($log_{10}$ cfu/g) | 6.60 | 6.64 |
| Lower Limit of 95% CI ($log_{10}$ cfu/g) | 4.74 | 4.72 |

## Annex III

# Estimation of Intermediate Reproducibility based on Routine Monitoring Data

1. Intra-laboratory uncertainty estimates can be made either by carrying out a full internal collaborative trial, with different analysts testing the same samples over a number of days or, for instance, using different batches or even different brands of commercial culture media. In such a case the statistical procedure of choice is that described in Annex II.

2. However, if a laboratory undertakes routine quality monitoring tests, it is possible to estimate reproducibility from these test data. One approach is to use a 1-way ANOVA and to take the mean residual square as the estimate of reproducibility. A preferred, and simpler procedure, is described fully in ISO19036: 2005; this determines the variance for each set of transformed replicate data values.

3. The reproducibility standard deviation is derived from the square root of the sum of the duplicate variances divided by the number of data sets. The equation is:

$$S_R = \sqrt{\sum_{i=1}^{n} \frac{(y_{i1} - y_{i2})^2 / 2}{n}}$$

where $y_{i1}$ and $y_{i2}$ are the log transformed values of the original duplicate counts ($x_1$ and $x_2$) and $n$ is the number of pairs of counts.

4. A worked example (based on $\log_{10}$ transformation) is presented below.

5. Confusion sometimes arises between repeatability and intermediate reproducibility. It must always be remembered that repeatability requires all stages of the replicated tests to be done **only** by a single analyst, carrying out repeat determinations on a single sample in a single laboratory, using identical culture media, diluents, etc within a short time period e.g. a few hours. If more than one analyst undertakes the analyses and/or tests are done on different samples and/or on different days then the calculation derives a measure of intermediate reproducibility. The procedure can be used to determined average repeatability estimates for individual analysts provided all the repeatability criteria are met..

6. Internal laboratory quality management is aided by the use of statistical process control (SPC). The estimates of intermediate reproducibility provide a source of data that is amenable to SPC.

**Worked Example (modified from ISO 19036:2005)**

The data below were derived from enumeration of aerobic mesophilic flora in mixed poultry meat samples. The duplicate data values ($x_{iA}$ and $x_{iB}$) are log transformed to give $y_{iA}$ and $y_{iB}$, respectively. The mean $\log_{10}$ counts ($\bar{y}$) are derived from $(y_{iA} + y_{iB})/2$; the variances ($S_{Ri}^2$) are derived from $(y_{iA} - y_{iB})/2$; and the RSD values from $100*\sqrt{S_{Ri}^2}/\bar{y}$.

| Test(*i*) | *Colony Count A (cfu/g)* | *Colony Count B (cfu/g)* | *Log count A* | *Log count B* | *Mean log Count* | *Absolute Difference in log count* | Variance | Relative Standard Deviation (%) |
|---|---|---|---|---|---|---|---|---|
| | $x_{iA}$ | $x_{iB}$ | $y_{iA}=\log_{10}(x_{iA})$ | $y_{iB}=\log_{10}(x_{iB})$ | $\bar{y}$ | $y_{iA} - y_{iB}$ | $S_{Ri}^2$ | $RSD_{Ri}$ |
| i=1 | 6.70E+04 | 8.70E+04 | 4.83 | 4.94 | 4.88 | 0.11 | 0.00643 | 1.64% |
| i=2 | 7.10E+06 | 6.20E+06 | 6.85 | 6.79 | 6.82 | 0.06 | 0.00173 | 0.61% |
| i=3 | 3.50E+05 | 4.40E+05 | 5.54 | 5.64 | 5.59 | 0.10 | 0.00494 | 1.26% |
| i=4 | 1.00E+07 | 4.30E+06 | 7.00 | 6.63 | 6.82 | 0.37 | 0.06717 | 3.80% |
| i=5 | 1.90E+07 | 1.70E+07 | 7.28 | 7.23 | 7.25 | 0.05 | 0.00117 | 0.47% |
| i=6 | 2.30E+05 | 1.50E+05 | 5.36 | 5.18 | 5.27 | 0.19 | 0.01723 | 2.49% |
| i=7 | 5.30E+08 | 4.10E+08 | 8.72 | 8.61 | 8.67 | 0.11 | 0.00622 | 0.91% |
| i=8 | 1.00E+04 | 1.20E+04 | 4.00 | 4.08 | 4.04 | 0.08 | 0.00313 | 1.39% |
| i=9 | 3.00E+04 | 1.30E+04 | 4.48 | 4.11 | 4.30 | 0.36 | 0.06595 | 5.98% |
| i=10 | 1.10E+08 | 2.20E+08 | 8.04 | 8.34 | 8.19 | 0.30 | 0.04531 | 2.60% |
| | | | | | | | | |
| **Σ** | | | | | | | **0.2193** | |
| **Average** | | | | | **6.18** | | **0.0219** | |

Using the $\log_{10}$-transformed data ($y_{ij}$), the reproducibility standard deviation is derived from:

$$S_R = \sqrt{\frac{\sum_{i=1}^{n}(y_{i1} - y_{i2})^2 / 2}{n}} = \sqrt{\frac{0.00643 + 0.00173 + ... + 0.04531}{10}} = \sqrt{0,02193} \approx \pm 0,15 \ \log_{10} \text{ cfu/g}$$

Average % Relative Standard Deviation ($RSD_{av}$)= $100*(S_R/\bar{y})$=100*(0.15/6.18)=2.39%

Individual tests ($i$ = 1….10) gave RSD values ranging from 0.47% to 5.98%, with an overall value of 2.39%.

***Note: it is incorrect to take the average of the individual RSD values.***

Once sufficient data are available, a moving $RSD_{av}$ can be determined and used in a statistical process control system.

# Annex IV

## Application of Robust Methods of Statistical Analysis

1. Because of the problems with the occurrence of outlier data, several alternative approaches to the Analysis of Variance have been developed, based on Robust Methods of Statistical Analysis.

2. Rather than relying on identification and removal of outlying data (which values could actually be valid results, albeit considerably different from most of the data) and then estimating the variance around the mean, alternative robust procedures rely on estimation of the variation around the median value.

3. A mean value will be affected significantly by one or more high (outlier) values within a data set, whereas the median value is not affected. Consider the following examples:

A. 1, 4, 3, 6, 3, 5, 6, 3, 4, 5 $n = 10$, $\sum = 40$, Mean = 4.0 Median = 4.0

B. 1, 4, 3, 6, 3, 5, **26**, 3, 4, 5 $n = 10$, $\sum = 60$, Mean = 6.0, Median = 4.0

C. 1, 4, 3, 6, 3, 5, **26**, 3, 4, **15** $n = 10$, $\sum = 70$, Mean = 7.0, Median = 4.0

D. 1, 4, 3, 6, 3, 5, **126**, 3, 4, **15** $n = 10$, $\sum = 170$, Mean = 17.0,Median = 4.0

E. 1, 4, 3, 6, 3, 5, 3, 4, $n = 8$, $\sum = 29$, Mean = 3.6, Median = 3.5

4. The presence of one or more high values (Examples B, C, D) has a significant effect on the mean value but no effect on the median value. Removal of the high outliers (E) reduces both the mean and the median values.

5. A similar effect would be seen with low value outliers. Of course, occurrence of both high and low outliers could balance out the effect on the mean.

6. There are two primary alternative techniques of robust analysis currently in use:

6.6.1. The Analytical Methods Committee of the Royal Society Chemistry (AMC 1989, 2001) describes one approach. The procedure calculates the median absolute difference (MAD) between the results and their median value and then applies Hüber's H15 method of winsorisation. Winsorisation is a technique for reducing the effect of outlying observations on data sets (for detail see Smith & Kokic, 1996). The procedure can be used with data that conform approximately to a normal distribution but with heavy tails and/or outliers. An example is shown below. The procedure is not suitable for multimodal or heavily skewed data sets. The AMC website[1] provides downloadable software for use either in Minitab or Excel (97 or later version).

6.1.1. An alternative approach, known as the Recursive Median is based on extrapolation of the work of Rousseeuw & Croux (1993). One version of this approach

[1] www.rsc.org/lap/rsccom/amc/amc_software.htm#robustmean

(described fully in ISO 16140:2003) uses Rouseeuw's recursive median $S_n$. However, Wilrich (2005a) recommends a modified approach to this procedure also based on Rousseeuw's $S_n$ computation.

**Worked Example - Analysis of data set containing outliers**

| **Duplicate Series of Colony Counts (as $Log_{10}$ cfu/g) done by 1 Analyst in each of 10 Laboratories** | | |
|---|---|---|
| Laboratory | A | B |
| 1 | 4.83 | 4.94 |
| **2** | **4.05** | **3.99** |
| **3** | **6.84** | **6.92** |
| 4 | 4.90 | 4.93 |
| 5 | 5.28 | 5.23 |
| 6 | 4.86 | 4.72 |
| 7 | 5.62 | 5.51 |
| 8 | 4.50 | 4.68 |
| 9 | 5.48 | 5.11 |
| 10 | 5.04 | 5.34 |

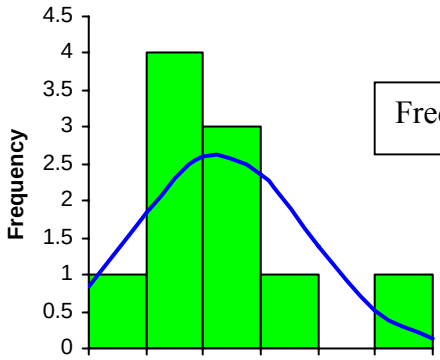
Laboratory 2 data look to be slightly low and laboratory 3 data to be high when compared with the other data.

## Graphical and Descriptive Analysis of the Data

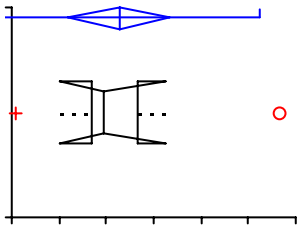
| | |
|---|---|
| **n** | 10 |
| **Mean** | 5.140 |
| **95% CI** | 4.602 to 5.678 |
| **Variance** | 0.5662 |
| **SD** | 0.7524 |
| **SE** | 0.2379 |
| **CV** | 15% |
| **Median** | 4.970 |
| **97.9% CI** | 4.500 to 5.620 |
| **Range** | 2.79 |
| **IQR** | 0.49 |

| **Percentile** | |
|---|---|
| **2.5th** | - |
| **25th** | 4.838 |
| **50th** | 4.970 |
| **75th** | 5.330 |
| **97.5th** | - |

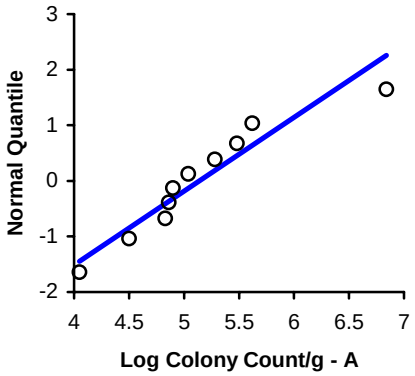
| | **Coefficient** | **p** |
|---|---|---|
| **Shapiro-Wilk** | 0.9204 | 0.3605 |
| **Skewness** | 1.1222 | 0.1009 |
| **Kurtosis** | 2.4815 | - |




Frequency Plot



Box Plot showing outliers




Although there is evidence of kurtosis and positive skewness, the log-transformed data conform fairly well to a "normal" distribution. The Box plot shows the presence of a potential low-level outlier (+) and a significant high-level outlier (○).

## One-way Analysis of Variance (ANOVA) without removal of outliers

| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
|---|---|---|---|---|---|---|
| Between Laboratories | 10.086 | 9 | 1.1207 | 70.816 | 7E-08 | 3.0204 |
| Within Laboratories | 0.15825 | 10 | 0.0158 | | | |
| Total | 10.2443 | 19 | | | | |

Repeatability SD = $\sqrt{0.0158} = 0.1258$
Reproducibility SD = $\sqrt{(1.1207 + 0.0158)} = \sqrt{1.1365} = 1.0661$

## One-way Analysis of Variance (ANOVA) after removal of high outlier (lab 3)

| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
|---|---|---|---|---|---|---|
| Between Laboratories | 3.3464 | 8 | 0.4183 | 24.281 | 3E-05 | 3.2296 |
| Within Laboratories | 0.15505 | 9 | 0.0172 | | | |
| Total | 3.50145 | 17 | | | | |

Repeatability SD = √0.0172 = 0.1311
Reproducibility SD = √ (0.4183+0.0172) = √0.4355 = 0.6599

## One-way Analysis of Variance (ANOVA) after removal of both low and high outliers (labs 2 & 3)

| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
|---|---|---|---|---|---|---|
| Between Laboratories | 1.42124 | 7 | 0.203 | 10.599 | 0.0017 | 3.5005 |
| Within Laboratories | 0.15325 | 8 | 0.0192 | | | |
| Total | 1.57449 | 15 | | | | |

Repeatability SD = √0.0192 = 0.1311
Reproducibility SD = √ (0.203+0.0192) = √0.2222 = 0.4714

## Analysis of Variance using the AMC Method

Software for this analysis, compatible with Microsoft Excel, can be downloaded from Royal Society of Chemistry statistical software. A version for use in Minitab is also available.

***ROBUST ESTIMATES***

| Parameter | Value |
|---|---|
| *Grand Mean* | 5.060622 |
| *Within-laboratory/repeatability SD* | 0.116772 |
| *Between-laboratory SD* | 0.476556 |
| *Reproducibility SD* | 0.490654 |

c=1.5: Convcrit=0.0001
Repeatability SD = 0.1168
Reproducibility SD = 0.4907

**Comparison of data analyses by ANOVA, without and with removal of the high (*) outlier and both the high and low outliers (**), by Robust ANOVA (AMC 1989, 2001) and by Recursive Median (ISO 16140:2003)**

| Parameter* | ANOVA | ANOVA* | ANOVA** | ROBUST | RECMED |
|---|---|---|---|---|---|
| **Mean** | 5.14 | 4.95 | 5.06 | 5.06 | |
| **Median** | | | | | 5.05 |
| **$SD_r$** | 0.126 | 0.131 | 0.138 | 0.117 | 0.115 |
| **$RSD_r$** | 2.45% | 2.65% | 2.73% | 2.31% | 2.28% |
| **$SD_R$** | 1.066 | 0.660 | 0.471 | 0.491 | 0.5590 |
| **$RSD_R$** | 20.45% | 13.33% | 9.42% | 9.70% | 11.07% |

* $SD_r$ = Standard Deviation of repeatability; $SD_R$ = Standard Deviation of reproducibility
$RSD_r$ = % relative standard deviation of repeatability
$RSD_R$ = % relative standard deviation of reproducibility

The effect of the outlier values on the Standard Deviation of reproducibility is clear from the above data. Removal of the high outlier (*) reduces both the mean and the $SD_R$; removal of both the high and low outliers (**) reduces both the mean value and the $SD_R$ to a level similar to than that seen in the Robust ANOVA. The Recursive Median technique (working data not shown) produces a similar value for $SD_r$ but a somewhat higher $SD_R$ value than does the Robust Method.

# ANNEX V

## Uncertainty Associated with Qualitative Methods

1. By definition, a non-quantitative method merely provides an empirical answer to a question regarding the presence or absence of a specific index organism or a group of related organisms in a given quantity of a representative sample.

2. Provided that multiple samples are analysed, and on the assumption that the test method is "perfect", then the number of tests giving a positive response provides an indication of the incidence of defective samples within a "lot".

2.1. For instance, if a test on 10 parallel samples found 4 positive and 6 negative samples then the perceived incidence of defectives would be 40% (*sic* of the samples analysed).

2.2. However, if no positive samples were found the apparent incidence of defectives in the "lot" would be zero. However, it is not possible to say that the "lot" is not contaminated because the true incidence of defective samples will be greater than zero.

3. Sampling theory for occurrence of defectives is based on the binomial distribution, in which the probability of an event occurring (p) or of not occurring (q) can be derived and an error estimate can be made based on a realistic number of samples analysed. Unfortunately, in laboratory practice it is not usually possible to analyse a realistic number of samples for the presence of specific microorganisms.

3.1. Table 1 below shows the statistical probability of occurrence of 0, 1, or 2 defective units in 10 sample units from "lots" containing from 0.1 to 30% true defectives. For a lot having only 0.1% defective units, the probability of detecting one or more defective (*sic* positive) samples is only 1 in 100 whilst for a lot having 5% true defectives there is still only a 40% probability of obtaining a positive result; even with 20% true defectives there is still a 20% chance of ***not*** finding defective units when testing 10 sample units.

3.2. Table 2 shows the probability of detecting 0, 1 or 2 defective units with increasing numbers of sample units tested when the true incidence of defectives is 10%. The probability of finding no defective samples is 59% if only 5 samples are tested, 35% with 10 samples and 12% with 20 samples.

3.3. These examples illustrate a basic characteristic of undertaking qualitative tests for specific organisms: unless the likelihood of contamination of the matrix is high, *and* the number of sample units tested is considerable *and* the analytical test itself is perfect, then the probability of detecting positive samples in food matrix is very low.

**Table 1. Binomial Probability of detecting 0, 1 or 2 defective units in 10 sample units tested with increasing incidence of true defectives (mod from Jarvis, 2000)**

| True Incidence (%)of Defective Units in a lot | Probability (p) of detecting defective units | | |
|---|---|---|---|
| | **0** | **1** | **2** |
| 0.1% | 0.99 | 0.01 | <0.001 |
| 1% | 0.90 | 0.09 | <0.01 |
| 5% | 0.60 | 0.32 | 0.08 |
| 10% | 0.35 | 0.39 | 0.19 |
| 20% | 0.20 | 0.35 | 0.28 |
| 30% | 0.03 | 0.12 | 0.23 |

**Table 2. Binomial probability of detecting defective units with increasing sample units from a lot having 10% true defectives (mod from Jarvis, 2000)**

| Number of Sample units (*n*) tested | Probability of detecting the following number of defective units | | |
|---|---|---|---|
| | **0** | **1** | **2** |
| 5 | 0.59 | 0.33 | 0.07 |
| 10 | 0.35 | 0.39 | 0.19 |
| 20 | 0.12 | 0.27 | 0.29 |
| 50 | <0.01 | 0.03 | 0.08 |

3.4. ***Maximum Incidence and Level of Contamination***. Even when all test results are negative, use of the binomial distribution concept permits the derivation of a probable maximum contamination limit for a test lot.

3.4.1. Assuming that results on all (*n*) sample units are negative, then for a given probability (*p*) the maximum incidence (*d*) of defective units is given by:

$$d = 100(1 - \sqrt[n]{(1-p)}$$

Hence, if $n = 10$ and $p = 0.95$, then

$$d = 100(1 - \sqrt[10]{(1-0.95)} = 100(1 - \sqrt[n]{0.05}) = 100(1-0.741) = 25.88\% .$$

3.4.2. Knowing the maximum incidence of defective sample units and the size of the sample units we can derive a Maximum Contamination level (C) from:

$$C = (d/100)(1/W) \text{ organisms per g,}$$

where $W$ is the weight of the sample unit tested. For the example given above and assuming that each of the 10 samples weighed 25g, then the maximum contamination level would be given by

$$C = (25.88/100)(1/25) = 0.0104 \text{ organisms/g} \equiv 10.4 \text{ organisms/Kg}$$

3.4.3. In other words, the failure to detect a positive in 10 parallel tests merely indicates, at a 95% probability, that the index organism would be present in not more than 26% of similar samples throughout the lot; and that the maximum contamination level would be 11 organisms/Kg of product.

3.4.4. It might be thought that such a level of product security is insufficient, in which case it would be necessary to analyse a greater number of sample units and ideally to test larger quantities of sample. It is essential also to recognise that this presupposes that the test method is "perfect".

3.5. ***Multiple Test Most Probable Number Estimates***. If some test results are positive, then we can derive an estimate of population density (the basis for derivation of a Most Probable Number) for multiple tests even at a single dilution level.

3.5.1. The following equation provides the derivation of the MPN:

$M = -\frac{1}{V}.ln\left(\frac{s}{n}\right)$, where $M$ = Most probable number, $V$ = quantity of sample, $s$ = number of sterile tests out of $n$ tests inoculated.

3.5.2. Assume 10 tests are set up on replicate 25g samples of product, 3 tests are positive and 7 are negative. Then the MPN of contaminating organisms is:

$$M = -\frac{1000}{25}.ln\left(\frac{7}{10}\right) = -40. - 0.3567 = 14.27 \text{ organisms/Kg} \approx 14 \text{ organisms/Kg}$$

3.5.3. Unfortunately, it is not possible to derive an estimate of the error of the MPN when tests are done at a single dilution level.

3.6. ***Level of Detection Estimates.*** The equation used in 3.5 is also the basis for deriving MPN values for use in the Spearman-Karber procedure to estimate the $LOD_{50}$ for a test. This is the level of organisms that will give 50% positive results when tested by an appropriate protocol. Details of the procedure together with worked examples are given in the report by Hitchins (2005). This method of quantification has the benefit that it is possible to derive a value for the standard error of the mean (*sic* LOD) estimate. The procedure can be used to compare performance of two or more

methods where both have been evaluated under identical conditions in two or more laboratories.

3.7. ***Estimation of repeatability and reproducibility for qualitative tests.*** In a paper produced for ISO SC9 TC34, Wilrich (2005b) proposed the estimation of repeatability and reproducibility estimates for qualitative test procedures based on the binomial probability of detection of positive and negative results in different laboratories operating either at equal or at dissimilar sensitivity levels. A set of simulation studies is presented, together with analyses of a set of practical interlaboratory assessments, which support the proposal but the method has yet to be evaluated in detail.

**4. Estimation of Error based on test performance.** One of the traditional problems associated with presence or absence tests relates to the likelihood that a method may give either a false negative (Type I error) or a false positive (Type II error) result. A false negative result fails to detect the occurrence of a known index organism in a sample. A false positive result indicates the presence of a specific index organism even though it is not present in the test matrix. Such errors create specific problems for interpretation of test results.

4.1. In a real life situation, where tests are done on natural matrixes, it is impossible to estimate the likelihood of detecting such false results. It is essential therefore to ensure during development, evaluation and use of any method that the likelihood of such errors occurring is at an absolute minimum. An efficient laboratory proficiency scheme provides a way to monitor the efficiency of a test procedure in any individual laboratory.

4.2. But to be sure that false results do not occur requires the use of reference materials that can be relied upon to contain the index organism at a given level. For high-level contamination that is not a major problem; the issue arises primarily where the level of detection is intended to be close to the minimum level of detection. For instance, to detect 1 cfu of a specific organism in (say) 25g of sample implies that the organism is evenly distributed throughout a lot of test material such that each 25g sample unit is likely to contain the organism. Only if it were possible to add a single test organism to each individual 25g sample could the probability that each sample would contain that organism be achieved and even then there is the real possibility that the organism would not survive the preparation and storage process.

4.3. If larger quantities of test organism are added to a large batch of test matrix, which is then thoroughly mixed, the distribution of organisms throughout the lot would at best be random but could possibly be over dispersed due to the presence of clumps of organisms.

4.4. Table 3 shows the probability of occurrence of 0 or ≥1 organisms in a 25g sample for different levels of inoculation. If the target inoculum level is only 1 organism per 25g, there is a 37% chance that less than 1 organism will not be present in the sample; to have a 99% probability that 1 or more organisms occur in a perfectly distributed sample matrix requires inoculation at a level of at least 5 organisms/25g. Even then

one has to assume that the original inoculum contains the test organism at the relevant level – it must not be forgotten that the organisms in a well-mixed inoculum will themselves be distributed in accordance with Poisson. It is therefore perhaps not surprising that at low inoculum levels, negative results may be found frequently. It is for such reasons that we recommend the $LOD_{50}$ approach of Hitchins (2005) for comparison of two or more methods of analysis.

## 5. The effect of competitor organisms and other factors on the recovery of organisms to critical levels.

5.1. A further potential cause of a false negative result is that during the multistage test protocol, the index organisms must be able to grow to a critical level to ensure effective transfer between different stages of the test. The ability of an organism to grow is dependent not only on the physiological condition of the index organism in the sample matrix, but also on the micro-environmental conditions within the test system, the presence or otherwise of competitive organisms that may affect the growth of the index organism and the time/temperature factors used in the protocol.

5.2. In their (BPMM) paper on Inter-laboratory Variability, LaBarre, Zelenka and Flowers (2005) have reviewed in detail the effects of competitive growth, problems associated with test media and other practical considerations in relation to the critical level of organisms for use in confirmatory tests following enrichment procedures. The paper describes also the statistical aspects of optimisation of test protocols.

**Table 3. Probability of occurrence of 0 or at least 1 organism in a 25g sample assuming perfect random distribution of test organisms in a matrix.**

| Target inoculation level (cfu/25g) | Probability ($p_x$) of Occurrence in 25g sample units | |
|---|---|---|
| | <1 organism | 1 or more organisms |
| 1 | 0.3679 | 0.6321 |
| 2 | 0.1353 | 0.8647 |
| 3 | 0.0498 | 0.9502 |
| 4 | 0.0183 | 0.9817 |
| 5 | 0.0067 | 0.9933 |
| 10 | <0.0001 | >0.9999 |

# Proposed Use of a 50 % Limit of Detection Value in Defining Uncertainty Limits in the Validation of Presence-Absence Microbial Detection Methods

## Background

A 50 % endpoint Limit of Detection ($LOD_{50}$) procedure can be used to calculate the absolute performance efficacies, and their associated uncertainties, of presence/absence methods for microbial detection in foods (3, *5*, *8*, *10*).

Validation of methods for microbial detection in foods or other matrices involves determining microbe recoveries. Recoveries are expressed qualitatively as presence-absence data, which are obtained from quantitative spiking experiments. Replicate samples of foods are spiked with the microbe of interest, generally at several concentration levels. Usually three different levels are used. However, only the data from the level which gives partial recovery are considered relevant. Such data are most reflective of a method's detection endpoint but a limit of detection is rarely estimated. The calculation discussed here maximizes the use of the data from such trials by using data from more than one spiking level to calculate an $LOD_{50}$.

Performance efficacies of new microbial detection methods are usually determined by comparison to recognized standard methods. This comparison is only strictly valid when common samples are used for the new and the standard methods. Then the methods are being compared at an equal microbial concentration. The situation is more complicated when comparisons involve a non-paired sample experimental design. Nevertheless, to a first approximation, comparative method validations always have the advantage of not really needing to determine the exact spiking concentration and thus virtually side-steps the fundamental problem of microbial enumeration variability at low concentrations. However, it is difficult to compare the results from different trials because the variability of the proportions of positive recoveries can be at least partly due to the technical difficulty of standardizing the spike levels from trial to trial. Also, sometimes only a single method may be validated so no intra-study comparison is possible.

The $LOD_{50}$ method *normalizes* the results of such studies by estimating the spiking concentration (cfu/analytical portion size), which would correspond to 50 % recovery. Importantly, it also provides a measure of the uncertainties in terms of confidence intervals (at the 95% level) of the estimated $LOD_{50}$. A 50 % endpoint is used because the low concentration region of the recovery-concentration relationship is theoretically a sigmoid curve, it being governed by the Poisson distribution. In the case of *Listeria* methods, at least, recovery-concentration curves are clearly describable by the Poisson relationship (6). The confidence intervals of asymptotic region estimates are somewhat narrower than those of estimates in the mid-region. Nevertheless, the concentration corresponding to the midpoint of a sigmoid recovery curve can be more precisely determined than for a point in the one of the asymptotic regions tending toward either 0 or 100% recovery.

## Methods for calculating an $LOD_{50}$

The $LOD_{50}$ calculation could potentially employ one of several mathematical tools (Table 1). These are used to calculate the dose corresponding to a 50% response value ($ID_{50}$ and $LD_{50}$) from the log-normal dose-response curve observed in an animal infection and mortality study. Thus in the $LOD_{50}$ determination, the proportion of replicates at a given spiking level that is culture negative (nominally uninfected) is treated just as would be the proportion of uninfected or surviving animals at a given challenge dose. Conversely, a test culture positive result is analogous to an animal infection or death. These calculation methods have been reviewed (*1*, *4*, *7*). They have various limitations and advantages (Table 1). The calculations are often laborious but this is not a major factor given the appropriate

computer application software. The methods differ statistically but appear to give endpoint estimates that differ only by a few percent. This variation is insignificant relative to the imprecision of spiking level estimation (7).

**Table 1. Estimation methods for $LD_{50}$ and $ID_{50}$ values**

| Name | Characteristics |
|---|---|
| Probit analysis | High efficiency; reiterated interpolation; replicates/spike & spiking intervals can vary |
| Reed & Muench | Lowest efficiency |
| Spearman-Kärber | Symmetrical doses; 0 and 100 % response values needed |
| Moving Average | Simple interpolation; curve shape not presumed |

The lack of clear statistical superiority of the other calculation methods to the Spearmann- Kärber method along with its previous application to $LOD_{50}$ calculations in studies of foods spiked with pathogens (5, *8, 9, and 10*) is the reason for its use in the present proposal. Also, an Excel version of the generalized Spearman-Kärber $LOD_{50}$ calculation for 3, 4, and 5-level spiking protocols (*2;* Anthony.Hitchins@cfsan.fda.gov), now makes it more easily circulated and PC-user friendly. The accompanying Excel file provides a 3-level spike example, a trial worksheet, a back-up copy, a revealed code version, and the generalized Spearman-Kärber formula.

The $LOD_{50}$ Determination

Foods are quantitatively spiked in replicate (*at least* in triplicate) with the test microbe at several inoculum levels (*at least* three). The proportion of replicates in which the microbe is detected at each spiking level is used to calculate the $LOD_{50}$ by the generalized version of the Spearman-Kärber method . The confidence interval of the estimate narrows with increasing replication. The spiking level enumerations have their own confidence limits, which can be quite broad as in a 3-replicate MPN, but the overriding effect of any one MPN value is more or less ameliorated by the use of 3 or more enumeration levels in estimating the $LOD_{50}$ value. Furthermore, the number of replicates can be increased to reduce the confidence interval of the MPN.

When there is comparison with a standard method the spiking level can be determined from the standard method result, since the MPN enumeration would be done with the standard method anyway. Thus the proportion of negative culture at a given spiking level yields, by the Poisson equation, the mean spiking level. In this method, the number of replicates should be preferably 10 or more. Replication values of 40 or more are easily achievable in multilaboratory experiments.

Incidentally, in multilaboratory experiments the LOD 50% can be calculated from the pooled data or it can be estimated as the mean of the individual laboratory LOD 50% values. In the latter case an estimate of interlaboratory uncertainty can be made.

Table 2 shows a simulated $LOD_{50}$ experiment.

**Table 2. Example of an $LOD_{50}$ experiment using hypothetical data for a 4-level spike**

| Spiking Level (cfu/25 g) | Microbe Recovery | | | |
|---|---|---|---|---|
| | No. replicates | No. positive | No. negative | $LOD_{50}$ (CI)[a] |
| 0[b] | 10 | 0 | 10 | . |
| 1 | 10 | 5 | 5 | 1.26 (0.53 – 3.03) cfu/25-g |
| 10 | 10 | 9 | 1 | . |
| 100 | 10 | 10 | 0 | . |

[a] Calculated by the Spearman-Kärber method. CI = 95 % confidence interval.
[b] A value of 0.1 was assumed for the calculation.

An $LOD_{90}$ value can be calculated from the $LOD_{50}$ value in Table 2: it is 2.87 cfu/25-g test portion. This calculation assumes that the LOD endpoint curve is described by the Poisson equation even when the observed $LOD_{50}$ value is different from the theoretical Poisson-based minimum $LOD_{50}$ value of 0.307 cfu/25-g test portion. This assumption is reasonable for the majority of published *Listeria* method validation studies (6).

Typically collaborative qualitative microbiology method validations involve 3 spiking levels and 5 replicate determinations per level for each of 10 or more laboratories. This provides 150 or more data points (10 laboratories x 3 levels x 5 replicates). Intuitively, the $LOD_{50}$ estimate by mathematical interpolation will be more accurate the greater the number of data points comprising the curve in the zone around the $LOD_{50}$ point. Increasing the number of concentration levels does not require maintaining the same level of replication in order to sustain a given confidence level interval with a constant number of laboratories. This is illustrated in Table 3.

Table 3. Confidence Intervals for Two Spike-Level:Replicate Trade-off Scenarios with Similar $LOD_{50}$ Results

| 10 lab x 7 level x 3 rep - 20[a] = 190 data points | | | 10 lab x 3 level x 7 rep - 20[a] = 190 data points | | |
|---|---|---|---|---|---|
| Mean Level[b] (cfu/25g) | Replicates per level | Positive Replicates[c] | Mean Level[b] (cfu/25g) | Replicates per level | Positive Replicates[c] |
| 4.6 | 30 | 30 | 4.6 | 90 | 90 |
| 2.3 | 30 | 27 | . | . | . |
| 1.15 | 30 | 21 | . | . | . |
| 0.625 | 30 | 13 | 0.625 | 90 | 39 |
| 0.313 | 30 | 6 | . | . | . |
| 0.157 | 30 | 1 | . | . | . |
| 0(<0.075) | 10[c] | 0 | 0(<0.075) | 10[c] | 0 |
| | | | | | |
| $LOD_{50}$ = 0.760 cfu /25-g analytical portion | | | $LOD_{50}$ = 0.700cfu /25-g analytical portion | | |
| 95% confidence interval = 0.575– 0.875 | | | 95% confidence interval = 0.550-.875 | | |

[a] The number of replicates at the zero level can be less than at the other levels, say 1 per laboratory, since their purpose here is to provide a zero positives data point as well as the usual assurance of a negligible natural contamination rate.

[b] Level intervals based on 1:2 dilutions as in R. Flowers's dilution to extinction method. Not all levels used in the 3-level scenario. More levels in the LOD concentration zone could be set-up with a lower dilution rate, e.g. 1in 1.5.

[c] Common levels of the two scenarios have equal proportions of positive replicates.

Discussion and Recommendations

The method is broadly applicable (3) to all published AOAC collaborative studies except that in a proportion of the results it has been necessary to resort to dummy values for the required 100% positive response data points. The dummy concentration value for 100% positive response is currently over conservatively set at 10x the experimental concentration that yielded the highest proportion of positives. In a planned revision of the Spearman-Kärber $LOD_{50}$ program, the 100% positive dummy concentration will be calculated by multiplying the highest concentration giving positives by the reciprocal of the proportion positive at that concentration. Of course, this necessity for a 100% positive dummy concentration can be largely avoided by increasing the number of concentration levels studied from the usual 3 levels to 4 or more concentration levels. The process of preparing concentrations that give partial positives is somewhat chancy and so it is likely that analysts are preparing levels giving 100% positive responses but are not presenting them since the current study protocol does not require them. So increase of the number of levels is unlikely to be onerous especially since the number of replicates per level can be correspondingly reduced (i.e. the product of the number of replicates per level and the number of levels need not be changed).

The generalized Spearman-Kärber method also requires a data point giving the concentration corresponding to zero positives. MPN limits of detection vary from <0.003 to <3 MPN/g depending on the maximum MPN sample size in the range from 100g down to 0.1g. There is no precise non-zero spike concentration (zero is not compatible with the logarithmic Spearman-Kärber calculation) corresponding to the controls used in AOAC studies. A value of 0.004 per g has been chosen as the concentration corresponding to the negative controls. This value is close to the minimum MPN likely to be encountered in spiking studies but more importantly is the extrapolation to zero% positive point of the midpoint region of the response curve, which is approximately linear and, which contains the $LOD_{50}$ point of interest. While one can interpolate the $LOD_{50}$ value from the experimental data, using the Spearman-Kärber method to obtain the $LOD_{50}$ also provides the confidence limits.

A proportion of published AOAC study results were not readily amenable to the LOD calculation. The use of 3-tube MPN sometimes gives sequential concentrations that are equal even though they should be different and even gives sequential values that are different but appear as if they have been inadvertently reversed. These problems can be solved by using a better MPN enumeration with more tubes per level or by using the standard method positive responses in a one level multi-tube MPN calculation (if a standard method is available) or by using the method suggested by R. Flowers. Nevertheless, retention of the conventional statistical tests used currently would be advisable for rare instances were the $LOD_{50}$ cannot be calculated.

Conclusions

The calculation of an $LOD_{50}$ value by the generalized Spearman-Kärber method provides a convenient way to condense virtually *all* of the raw data from a multi-level food spiking trial into one readily comprehensible *absolute* value of performance efficacy. In addition, it provides the estimate's *uncertainty*, given as the 95% confidence interval. The breadth of the confidence interval will depend inversely on the number of replicates at each level. The replication at each level need not be constant in this generalized version of the Spearman-Kärber calculation. More sophisticated calculation methods may become available in the future but meanwhile the generalized Spearman-Kärber method is already available to do the job of calculating detection limits and moreover it has the advantage of not requiring complex computations. The problem with modeling an empirical response curve from all available study data is that each data point from study to study involves so many variables and the plot of % positive versus concentration is highly scattered.

In study designs where a new method and the standard method are compared, $LOD_{50}$ values do not just augment the conventional relative performance parameters with absolute performance parameters; in addition, they also provide estimates of the uncertainties of the method's performances. $LOD_{50}$ values for one-method study designs can be compared with previously published values for that and other methods and also with the theoretically expected minimal recovery value for a particular analytical portion size.

It is clear that the generalized Spearman-Kärber method will be most useful if the AOAC collaborative study design is adjusted appropriately by innovations such as Russ Flowers's dilution to extinction method.

REFERENCES

1. Armitage, P. and Allen, I. 1950. Methods of estimating the $LD_{50}$ in quantal response data. J. Hyg., Camb. **48**:298.
2. Encyclopedia of Statistical Sciences. 1982. **1**:355. John Wiley & Sons, New York.
3. Burney, A. A., and Hitchins, A. D. 2003. Determination of the limits of detection of AOAC validated qualitative microbiology methods. Manuscript in preparation.
4. Finney, D. J. 1952. *Probit Analysis*, 2nd ed. Cambridge University Press, London and New York.
5. Hitchins, A. D. 1989. Quantitative comparison of two enrichment methods for isolating *Listeria monocytogenes* from inoculated ice cream. J. Food Protect **52**:898.
6. Hitchins A. D. 1998. Retrospective interpretation of qualitative collaborative study results: Listeria methods. AOAC Intl. Annual Meeting **112**: 102, Abstract J-710.
7. Meynell, G. G., and Meynell, E. 1965 *Theory and Practice in Experimental Bacteriology*, pp 179-182. Cambridge University Press, London and New York.
8. Thunberg, R.L., Tran, T.T., and Walderhaugh, M. O. 2000. Detection of thermophilic *Campylobacter* spp. in blood-free enriched samples of inoculated foods by the polymerase chain reaction. J. Food Protect. **63**:299.
9. Tran, T. T., Stephenson, P. and Hitchins, A. D. 1990. The effect of aerobic mesophilic microfloral levels on the isolation of inoculated *Listeria monocytogenes* strain LM82 from selected foods. J. Food Safety **10**: 267.
10. Twedt, R. M., Hitchins, A. D., and Prentice, G. 1994. Determination of the presence of Listeria monocytogenes in milk and dairy products: IDF collaborative study. J. AOAC INTERNATIONAL **77**:395.

Version 09/03/2002

# Limit of Detection Program for Qualitative Microbiology Methods

**NOTE There is code only for 3-, 4-, and 5-level spiking protocols**

< Anthony.Hitchins@cfsan.fda.gov >

## PURPOSE

This programmed non-parametric statistical procedure (Spearman-Karber 50% Endpoint) will calculate the microbial analyte concentration (and confidence limits) in a given food matrix that corresponds to a 50 % probability of a positive result with the test method used. The microbe may be spiked or incurred.

## REQUIREMENTS

1. A minimum of three different concentrations is needed but more are preferable even at the expense of the degree of replication.
2. At least one spiking level should give a partially positive response otherwise no confidence limits can be calculated.
3. One of the concentration (spiking) levels should give a 100% response.
4. One of the concentration levels should give a 0% response (= negative control).
5. Three, and preferably at least 5 or 6, replicates per concentration (spiking) level are needed. (the confidence level window narrows with increased replication)
6. A *constant* analytical portion size.
7. Equal spacing of log spiking (concentration) levels and equal numbers of replicates per spike level are preferable but not obligatory.

## ACCOMMODATIONS

1. Sometimes it is possible to have only 2 replicates per concentration level.
2. An uninoculated control set of replicates is often used. In this case a concentration of 0.004 MPN/g(mL) or cfu/g(mL) can be allocated to such a set of data (since the calculations take the log of the concentration levels, 0 cannot be used, and a concentration of 0,004 is assumed to give always a negative result).
3. If no spike level has a 100% replicate growth response, use a dummy set of data. The spike level should be 10 times the uppermost level giving a partial response.

**PROCEDURE**

**Analytical portion size (g or mL)**

25

**Step 1.** Insert analytical portion size in the green-filled cell.

**Step 2a.** Insert spike sizes *either* on an analytical portion basis *or* on a per g/mL basis in the appropriate green-filled cells. Spike size must increase in the downward direction.

**Note:** If values are entered on a analytical portion basis they will be processed to a per g/mL basis

**Step 2b.** Enter number of replicates at each spiking level. (The number of spiking levels must include one with all replicates not grown and one with all replicates grown.)

[Read the guidelines in the hidden comments. They may be revised in the future.]

**Caution:** Do not copy from cells without color fill. They may contain hidden code and, if so, only the code will be copied.
Similarly, do not paste to such cells as any code present will be obliterated.

**Step 3.** Read cream color -filled Results panel.

| Enter spike size cfu or MPN/portion | Enter spike size cfu or MPN / g or mL | Enter number of replicates per spiking level | Enter number of replicates grown |
|---|---|---|---|
| | | | 0 |
| | Total degrees of freedom = | -1 | |
| From Table enter t value for 95% confidence level | | 2.03 | |

**RESULT**

| LOD 50 % (cfu or MPN per g/mL) | lower limit | upper limit |
|---|---|---|
| | | |

**TABLE t-values**

| df | t-value (2-tail) |
|---|---|
| 3 | 3.182 |
| 6 | 2.447 |
| 9 | 2.262 |
| 12 | 2.179 |
| 15 | 2.131 |
| 20 | 2.086 |
| 25 | 2.06 |
| 30 | 2.042 |
| 40 | 2.021 |
| 60 | 2 |
| infinite | 1.96 |

Jan-00

# Limit of Detection Program for Qualitative Microbiology Methods

< Anthony.Hitchins@cfsan.fda.gov >

## PURPOSE

This programmed non-parametric statistical procedure (Spearman-Karber 50% Endpoint) will calculate the microbial analyte concentration (and confidence limits) in a given food matrix that corresponds to a 50 % probability of a positive result with the test method used. The microbe may be spiked or incurred.

## REQUIREMENTS

1. A minimum of three different concentrations is needed but more are preferable even at the expense of the degree of replication.
2. At least one spiking level should give a partially positive response otherwise no confidence limits can be calculated.
3. One of the concentration (spiking) levels should give a 100% response.
4. One of the concentration levels should give a 0% response (= negative control).
5. Three, and preferably at least 5 or 6, replicates per concentration (spiking) level are needed. (the confidence level window narrows with increased replication)
6. A *constant* analytical portion size.
7. Equal spacing of log spiking (concentration) levels and equal numbers of replicates per spike level are preferable but not obligatory.

## ACCOMMODATIONS

1. Sometimes it is possible to have only 2 replicates per concentration level.
2. An uninoculated control set of replicates is often used. In this case a concentration of 0.004 MPN/g(mL) or cfu/g(mL) can be allocated to such a set of data (since the calculations take the log of the concentration levels, 0 cannot be used, and a concentration of 0,004 is assumed to give always a negative result).
3. If no spike level has a 100% replicate growth response, use a dummy set of data. The spike level should be 10 times the uppermost level giving a partial response.

**PROCEDURE**

| Analytical portion size (g or mL) |
|---|
| 25 |

**Step 1.** Insert analytical portion size in the green-filled cell.
**Step 2a.** Insert spike sizes *either* on an analytical portion basis *or* on a per g/mL basis in the appropriate green-filled cells. Spike size must increase in the downward direction.
**Note:** If values are entered on a analytical portion basis they will be processed to a per g/mL basis
**Step 2b.** Enter number of replicates at each spiking level. (The number of spiking levels must include one with all replicates not grown and one with all replicates grown.)
[Read the guidelines in the hidden comments. They may be revised in the future.]
**Caution:** Do not copy from cells without color fill. They may contain hidden code and, if so, only the code will be copied.
Similarly, do not paste to such cells as any code present will be obliterated.
**Step 3.** Read cream color -filled Results panel.

| Enter spike size cfu or MPN/portion | Enter spike size cfu or MPN / g or mL | Enter number of replicates per spiking level | Enter number of replicates grown |
|---|---|---|---|
| | 0.001 | 10 | 0 |
| | 0.01 | 10 | 1 |
| | 0.1 | 10 | 9 |
| | | | |
| | | | |
| | | | |
| | Total degrees of freedom = | 27 | |
| From Table enter t value for 95% confidence level | | 2.03 | |

| df | | Spike size cfu or MPN / g or mL |
|---|---|---|
| 9 | 1 | 0 |
| 9 | 1 | 0.01 |
| 9 | 1 | 0.1 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| | 3 | |

**RESULT**

| LOD 50 % (cfu or MPN per g/mL) | lower limit | upper limit |
|---|---|---|
| 0.025 | 0.016 | 0.04 |

**TABLE t-values**

| df | t-value (2-tail) |
|---|---|
| 3 | 3.182 |
| 6 | 2.447 |
| 9 | 2.262 |
| 12 | 2.179 |
| 15 | 2.131 |
| 20 | 2.086 |
| 25 | 2.06 |
| 30 | 2.042 |
| 40 | 2.021 |
| 60 | 2 |
| infinite | 1.96 |

## 3-SPIKING LEVEL DESIGN

| | spike levels MPN or cfu /g-n | number of replicates per spiking level | r of replicates grown | log spike | proportion grown |
|---|---|---|---|---|---|
| Lowest (or uninoculated) | 0.001 | 10 | 0 | -3 | 0 |
| | 0.01 | 10 | 1 | -2 | 0.1 |
| Highest level | 0.1 | 10 | 10 | -1 | 1 |

| | a | b | a*b | A | B | A*B | |
|---|---|---|---|---|---|---|---|
| | -2.5 | 0.1 | -0.25 | 0.01 | 1 | 0.01 | |
| | -1.5 | 0.9 | -1.35 | 0.01 | 1 | 0.01 | |
| log median | -1.6 | | | | | var | sqrt var |
| | | | | | | 0.01 | 0.1 |

| Log 50% endpoint | Log lower limit | upper limit |
|---|---|---|
| -3.6848 | -4.152309 | -3.22 |

**RESULT**

| 50 % endpoint | lower limit | upper limit |
|---|---|---|
| 0.025 | 0.016 | 0.04 |

## 4-SPIKING LEVEL DESIGN

| | spike levels MPN or cfu /g-n | number of replicates per spiking level | r of replicates grown | log spike | proportion grown |
|---|---|---|---|---|---|
| Lowest (or uninoculated) | 0.001 | 10 | 0 | -3 | 0 |
| | 0.01 | 10 | 1 | -2 | 0.1 |
| | 0.1 | 10 | 9 | -1 | 0.9 |
| Highest level | 0 | 0 | 0 | #NUM! | #DIV/0! |

| | a | b | a*b | A | B | A*B | |
|---|---|---|---|---|---|---|---|
| | -2.5 | 0.1 | -0.25 | 0.01 | 1 | 0.01 | |
| | -1.5 | 0.8 | -1.2 | 0.01 | ##### | ##### | |
| | #### | #### | #NUM! | | | var | sqrt var |
| log median | #NUM! | | | | | ##### | #NUM! |

| Log 50% endpoint | Log lower limit | upper limit |
|---|---|---|
| #NUM! | #NUM! | #### |

**RESULT**

| 50 % endpoint | lower limit | upper limit |
|---|---|---|
| #NUM! | #NUM! | #NUM! |

## 5-SPIKING LEVEL DESIGN

| | spike levels MPN or cfu /g-n | number of replicates per spiking level | r of replicates grown | log spike | proportion grown |
|---|---|---|---|---|---|
| Lowest (or uninoculated) | 0.001 | 10 | 0 | -3 | 0 |
| | 0.01 | 10 | 1 | -2 | 0.1 |
| | 0.1 | 10 | 9 | -1 | 0.9 |
| | 0 | 0 | 0 | #NUM! | #DIV/0! |
| Highest level | 0 | 0 | 0 | #NUM! | #DIV/0! |

| | a | b | a*b | A | B | A*B | |
|---|---|---|---|---|---|---|---|
| | -2.5 | 0.1 | -0.25 | 0.01 | 1 | 0.01 | |
| | -1.5 | 0.8 | -1.2 | 0.01 | ##### | ##### | |
| | #### | #### | #NUM! | #DIV/0! | ##### | ##### | |
| | #### | #### | #NUM! | | | var | sqrt var |
| log median | #NUM! | | | | | ##### | #NUM! |

| Log 50% endpoint | Log lower limit | upper limit |
|---|---|---|
| #NUM! | #NUM! | #### |

**RESULT**

| 50 % endpoint | lower limit | upper limit |
|---|---|---|
| #NUM! | #NUM! | #NUM! |

## 6 OR MORE SPIKING LEVEL DESIGNS

Not typically needed.

## Generalized Spearmann-Kärber Formula (see "Karber Method" entry on p. 354ff of Encyclopedia of Statistics, volume 4 )

$$\tilde{\mu} = \sum_{i=1}^{k-1} (p_{i+1} - p_i)\,(\mathbf{x}_i + \mathbf{x}_{i+1}) / 2$$

$\tilde{\mu}$ = estimator of the mean, $\boldsymbol{\mu}$

$\mathbf{x_i}$ 's are the log spiking concentrations with $\mathbf{x_i} < \ldots < \mathbf{x_k}$

**k** is the number of spiking levels

$\mathbf{p_i}$ 's are observed proportions of positive replicates in an experiment where $\mathbf{n_i}$ replicates are tested independently at spiking level $\mathbf{x_i}$ yielding $\mathbf{r_i}$ positive replicates , so $\mathbf{p_i} = \mathbf{r_i} / \mathbf{n_i}$, where $\mathbf{i} = 1, \ldots, \mathbf{k}$.
It is assumed that $\mathbf{p_i} = 0$ and $\mathbf{p_k} = 1$.

$$\text{var}(\tilde{\mu}) = \sum_{i=2}^{k-1} [p_i / q_i / (n_i - 1)]\,[x_{i+1} - x_{i-1})/2]^2$$

provided that $\mathbf{n_i} \geq 2$, and $\mathbf{i} = 1\ldots$ , $\mathbf{k}$ , and where $\mathbf{q_i} = 1 - \mathbf{p_i}$.

Version 09/03/2002

# Limit of Detection Program for Qualitative Microbiology Methods

< Anthony.Hitchins@cfsan.fda.gov >

## PURPOSE

This programmed non-parametric statistical procedure (Spearman-Karber 50% Endpoint) will calculate the microbial analyte concentration (and confidence limits) in a given food matrix that corresponds to a 50 % probability of a positive result with the test method used. The microbe may be spiked or incurred.

## REQUIREMENTS

1. A minimum of three different concentrations is needed but more are preferable even at the expense of the degree of replication.
2. At least one spiking level should give a partially positive response otherwise no confidence limits can be calculated.
3. One of the concentration (spiking) levels should give a 100% response.
4. One of the concentration levels should give a 0% response (= negative control).
5. Three, and preferably at least 5 or 6, replicates per concentration (spiking) level are needed. (the confidence level window narrows with increased replication)
6. A *constant* analytical portion size.
7. Equal spacing of log spiking (concentration) levels and equal numbers of replicates per spike level are preferable but not obligatory.

## ACCOMMODATIONS

1. Sometimes it is possible to have only 2 replicates per concentration level.
2. An uninoculated control set of replicates is often used. In this case a concentration of 0.004 MPN/g(mL) or cfu/g(mL) can be allocated to such a set of data (since the calculations take the log of the concentration levels, 0 cannot be used, and a concentration of 0,004 is assumed to give always a negative result).
3. If no spike level has a 100% replicate growth response, use a dummy set of data. The spike level should be 10 times the uppermost level giving a partial response.

**PROCEDURE**

| Analytical portion size (g or mL) |
|---|
| 25 |

**Step 1.** Insert analytical portion size in the green-filled cell.
**Step 2a.** Insert spike sizes *either* on an analytical portion basis *or* on a per g/mL basis in the appropriate green-filled cells. Spike size must increase in the downward direction.
**Note:** If values are entered on a analytical portion basis they will be processed to a per g/mL basis
**Step 2b.** Enter number of replicates at each spiking level. (The number of spiking levels must include one with all replicates not grown and one with all replicates grown.)
[Read the guidelines in the hidden comments. They may be revised in the future.]
**Caution:** Do not copy from cells without color fill. They may contain hidden code and, if so, only the code will be copied.
Similarly, do not paste to such cells as any code present will be obliterated.
**Step 3.** Read cream color -filled Results panel.

| Enter spike size cfu or MPN/portion | Enter spike size cfu or MPN / g or mL | Enter number of replicates per spiking level | Enter number of replicates grown |
|---|---|---|---|
| | 0.001 | 10 | 0 |
| | 0.01 | 10 | 1 |
| | 0.1 | 10 | 9 |
| | Total degrees of freedom = | 27 | |
| From Table enter t value for 95% confidence level | | 2.03 | |

**RESULT**

| LOD 50 % (cfu or MPN per g/mL) | lower limit | upper limit |
|---|---|---|
| 0.025 | 0.016 | 0.04 |

**TABLE t-values**

| df | t-value (2-tail) |
|---|---|
| 3 | 3.182 |
| 6 | 2.447 |
| 9 | 2.262 |
| 12 | 2.179 |
| 15 | 2.131 |
| 20 | 2.086 |
| 25 | 2.06 |
| 30 | 2.042 |
| 40 | 2.021 |
| 60 | 2 |
| infinite | 1.96 |

# Fit for Purpose Validation
# Classification Matrix

The following are proposed guidelines listing general categories (Purpose) and corresponding Minimum Validation Requirements. The level of method validation used should be based on several factors including risk, application, industry standards or regulatory requirements. When choosing a method and desired outcome, also, consider other factors that contribute to the result including sample size, sampling plan, laboratory/technician proficiency and measurement uncertainty.

| Purpose | Examples | Minimum Method Validation Requirements |
|---|---|---|
| Process Monitoring Product Monitoring | Raw Material Tests<br>In-process Tests<br>Indicator Test (Quality) | **SLV (Single Lab Validation) Methods:** Methods validated through single laboratory studies including inclusivity, exclusivity, ruggedness, stability and lot-to-lot variation. For qualitative methods, method performance is determined by LOD50 and compared to a reference method, if available. For quantitative methods, method performance is determined by LOD, LOQ, RSDr and linearity in comparison to a reference method. |
| Process Verification | Routine Sample Tests<br>HACCP Verification Tests<br>Supplier Verification Tests | **MLV (Mult-Lab Validation) Methods:** Methods have been validated by two or more laboratories. Inclusivity, exclusivity, ruggedness, stability and lot-to-lot variation studies are performed in one lab. Method performance studies (see SLV) are conducted in two or more labs following identical protocols using the same matrix/strain combinations. |
| Process Validation | New Process Validation Tests<br>Equipment Validation Tests | **MLV (Mult-Lab Validation) Methods:** Methods have been validated by two or more laboratories. Inclusivity, exclusivity, ruggedness, stability and lot-to-lot variation studies are performed in one lab. Method performance studies (see SLV) are conducted in two or more labs following identical protocols using the same matrix/strain combinations. |
| Regulatory Screening, Commercial Screening | Finished Product Release Tests<br>Routine/Scheduled Audit Tests<br>Routine Import Tests | **HCV (Harmonized Collaborative Validation) Methods:** Methods that have been validated by full collaborative study. The collaborative study must report valid data for method performance (see SLV) using robust statistcs without removal of outliers, except for assignable causes. The HCV must be preceded by a succssful SLV or MLV. |
| Regulatory Confirmation Testing | "Official Samples"<br>Tests in response to complaints or previous positives | **HCV (Harmonized Collaborative Validation) Methods:** Methods that have been validated by full collaborative study. The collaborative study must report valid data for method performance (see SLV) using robust statistcs without removal of outliers, except for assignable causes. The HCV must be preceded by a succssful SLV or MLV. |
| Forensic Testing | Lab Confirmation tests for BioTerrorism Agents | **HCV (Harmonized Collaborative Validation) Methods:** Methods that have been validated by full collaborative study. The collaborative study must report valid data for method performance (see SLV) using robust statistcs without removal of outliers, except for assignable causes. The HCV must be preceded by a succssful SLV or MLV. |
| Crisis Management | Emerging Pathogens | Use Best Available Method (dependent on critical time and risk)<br>**SLV (Single Lab Validation) Methods**: Methods validated through single laboratory studies including inclusivity, exclusivity, ruggedness, stability and lot-to-lot variation. For qualitative methods, method performance is determined by LOD50 and compared to a reference method, if available. For quantitative methods, method performance is determined by LOD, LOQ, RSDr and linearity in comparison to a reference method |
| | Emerging Disease Outbreaks | **MLV (Mult-Lab Validation) Methods:** Methods have been validated by two or more laboratories. Inclusivity, exclusivity, ruggedness, stability and lot-to-lot variation studies are performed in one lab. Method performance studies (see SLV) are conducted in two or more labs following identical protocols using the same matrix/strain combinations. |

# Recommendations for Future Research

In the course of this project, the working groups identified various areas where further research was needed, or a more comprehensive review of the documents developed for this project. The areas of further research include the following:

Validation Study Design and Statistical Analysis

1. Evaluate statistical approaches for qualitative and quantitative methods including: (1) further development of procedures for describing the Limit of Detection for quantitative methods; (2) further development of recommendations for use of the generalized Spearman-Karber method for estimating the $LOD_{50}$ for qualitative methods; and (3) evaluation of alternative approaches to the Spearman-Karber method e.g. Logit, Probit and other statistical procedures currently under investigation by the ISO TC34/SC9/SWG. Active participation in the ISO committee discussions is encouraged. A comparison of the generalized Spearman-Karber method to logit and probit analyses will be undertaken. It is important to determine what issues are important for an appropriate statistical method. The most appropriate method will depend on the study design and the assumptions of the statistical method. The consensus opinion of the task force is that more than two levels of contamination are needed for an $LOD_{50}$ analysis.

2. Use of existing AOAC data for assisting in design issues and choice of statistical methodology for future validation studies. This could include proper consideration of Type II error in addition to Type I error, and should develop a structured approach for making decisions based on the data. Non-AOAC data (e.g., clinical data) should be considered as well since AOAC data has design limitations. Effort should be made to identify individuals outside of AOAC (e.g., through FDA's Center for Devices and Radiological Health) that may be doing innovative work in this area.

3. There is a concern about the statistical comparisons used that are usually weighted to the prevention of Type I error (stating a difference exists when one does not) over Type II error (stating no difference exists when one does). The statistical hypothesis in testing evaluates if there is a different between two groups (two-tail test), one group is larger than another (upper-tail test), or one group is less than another (lower-tail test). There is no test for equivalence in significance-testing, yet that is often the major focus of AOAC testing. That is, that Lab A and Lab B results are not different. Perhaps a remedy as simple as increasing Type I error levels ($\alpha > 0.05$) and reducing Type II error levels ($\beta > 0.20$) would be useful.

4. The project timeline did not allow full discussion of the differences between Single Laboratory Validation (SLV), Multi-Laboratory Validation (MLV) and Collaborative Validation (HCV) in terms of the statistical confidence related to method performance and the effects of changes in number of samples, levels, analysts, labs, etc. The task force recommends that further work be done to elaborate these differences. In addition, the task force recommends investigation of the effectiveness of current AOAC Official Methods for Single Laboratory Validation (SLV) procedures, Multi-Laboratory Validation procedures (MLV) and harmonized Collaborative Validation studies (HCV), relative to the recommendations concerning the design of verification studies. Develop general guidelines for method validation protocols relative to

different applications (fit for purpose) and how these might be modified depending on the level of confidence required (how much uncertainty can be tolerated). Ideally, the guidelines would be flexible to allow for practical considerations, such as allowing an increased number of samples per lab to compensate for fewer labs in a study.

Confirmation of Results

5. As new innovative technologies are exploited for food pathogen detection, the gap between the $LOD_{50}$ of the alternative method and the $LOD_{50}$ of the reference cultural method ("gold standard") is expected to widen. This can result in presumptive positive results for the alternative method that cannot be confirmed culturally. In addition, as new pathogens emerge, gold standard methods may not exist. Finally, we must consider validation of methods to detect organisms that are viable but not culturable or not easily culturable, such as mycobacteria and viruses. In these cases, it is necessary to develop new approaches. Several approaches to be evaluated include:

a. Quantification of the confidence in the presumptive positive results in the method validation study. One proposal is to determine the incidence of positive results for a given uninoculated food matrix. Assuming the incidence is low, some statistical confidence is gained that presumptive positive results obtained in a validation study of the inoculated food matrix reflect the presence of the target analyte. The task force recommends that this concept, a modification of the clinical positive predictive values and negative predictive values, should be further discussed and developed.
b. Confirmation using methods based on technology distinct from the alternate method being validated. For example, a PCR test may be used to validate an immunoassay result. RNA targets could be used to ensure detection of live cells.
c. Confirmation based on detection of multiple analyte markers. In the absence of a suitable confirmatory test (high sensitivity and specificity), multiple tests could be used for confirmation and the level of agreement between these tests specified in order to achieve a true result.
d. Comparison of fractionally positive results to the theoretical Poisson and/or other distributions.

Preparation of Inoculated Samples

6. Test the dilution to extinction method for preparing samples for validation studies. Dilution to extinction is essentially an MPN method based on probability with an assumed distribution. The method should be further developed and tested to determine the number of levels and number of samples per level that should be tested and what level of recovery and statistical confidence can be achieved. Further, experimentally determine if these techniques can reliably calculate the level of target organism at the limit of detection.

Method Verification

7. The task force recommends that a laboratory intending to adopt a method that has been validated, verify the performance of that method in their laboratory. Future work is required to

develop procedures for verification of all validated methods, so that the method description will include a minimal verification procedure.

Ruggedness Testing

8. Ideally, every method validation would be initiated with a single lab validation to assess a variety of method performance parameters. If a method is intended to be validated through a multi-lab or collaborative study, this would occur after successful completion of the single lab validation. The task force recommends that some ruggedness testing of the method be performed as part of the single lab validation study. Critical parameters to be tested in the SLV ruggedness studies depend on the type of method under consideration. Future work will include development of guidelines for choosing parameters and designing ruggedness studies.

Note: Ultimately, the goal of the BPMM is to produce new proposed AOAC guidelines for validation, verification, modification and extension of microbiological methods.

**1. Purpose: This document lists and explains the terms used by the BPMM Task Force of AOAC International.**

**2. Scope: The definitions included in this Appendix are common terms used in the BPMM Task Force for AOAC International report to the contractor.**

**3. References:**

**FSIS Lab Quality Manual – Appendix A: Glossary, Rev. 02**
**Analytical Terminology for Codex Use, 2002.**
**AOAC INTERNATIONAL Guidelines for Laboratories Performing Microbiological and Chemical Analyses of Food and Pharmaceuticals, An Aid to Interpretation of ISO/IEC 17025, 2004.**
**AOAC INTERNATIONAL Methods Committee Guidelines for Validation of Qualitative and Quantitative Food Microbiological Official Methods of Analysis, 2002.**
**Official Methods of Analysis of AOAC INTERNATIONAL, 17th edition.**
**ISO Guide 2, 30, 9000.**
**State of Massachusetts Environmental Protection Agency Glossary for Quality Assurance Terminology.**
**21 CFR Part II.**

**4. Definitions:**

**Accuracy of a Measured Value:** A measure of the expected "closeness of agreement" between a measured value and the accepted, "true," or reference value. It is the expected value of the absolute value of the difference between the measured value and the true or accepted reference value.

**Accuracy of an Attribute Test:** The percentage of correct responses.

**Accuracy Index:** The square root of the sum of the bias squared and the variance for individual results, used as a measure of test accuracy within and among laboratories.

**Alpha α-probability**: The probability of a Type I error.

**Analyte:** The microorganism, substance or chemical constituent that is analyzed.

**ANOVA:** An acronym for a statistical procedure entitled Analysis Of Variance.

**Assignable Cause(s):** The reason, (root cause(s)) that a Shewhart Chart produces an "Out of Control Signal." Assignable causes may not ever be identified; in fact they may not exist.

**Attribute (k-class) test** – A test for which the measurement procedure yields k possible answers; applied usually when k is equal to 2 (e.g., pass/fail), or 3(e.g., acceptable, marginal, unacceptable).

**Average** ($\overline{X}$) **Chart:** A Shewhart Chart that plots the average of results from units considered as one sample versus sample number.

**Beta β-probability**: The probability of a Type II error.

**Bias:** The difference between the expected value of test results and an accepted reference value: Note: Bias is the total systematic error as contrasted to random error. There may be one or more systematic error components contributing to the bias. A larger systematic difference from the accepted reference value is reflected by a larger bias value.

**Binomial Probability Distribution:** A probability distribution characterized by situations having two possible outcomes, such as, coin toss or in microbiological situations the presence or absence of a pathogen: e.g. if p is the probability of a head, then on N independent tosses of the coin, the number of heads is distributed as a binomial distribution with parameter vales N and p. The expected value of the number of heads, considered as a random variable, is Np and the variance is Np(1-p).

**Blank:** A substance that contains none of the analytes of interest subjected to the usual analytical or measurement process to establish a baseline or background value. There are several types of blanks, each with a specific purpose including:

**Reagent Blank** - A blank containing no matrix elements that are carried through the analytical method to detect any contamination occurring during sample analysis.

**Method (Tissue) Blank** - A blank prepared to represent the sample matrix as closely as possible and treated like a sample through one or more phases of sample preparation and analysis. It serves to provide an estimate of all contamination occurring during all the processing/analysis steps to which it is subjected, as well as any endogenous matrix interferences.

**Calibration:** The set of operations which establish, under specified conditions, the relationship between values of quantities by a measuring instrument or measuring system, or values represented by a material measure or a reference material, and the corresponding values realized by standards.

**Calibration Laboratory:** A laboratory that performs calibration (as a service).

**Calibration Method:** A specified technical procedure for performing a calibration.

**Certified Reference Material (CRM):** A reference material, characterized by a metrologically valid procedure for one or more specified properties, accompanied by a certificate that provides the value of the specified property, its associated uncertainty, and a statement of metrological traceability reference material, accompanied by a certificate, one or more of whose property values are certified by a procedure which establishes traceability to an accurate realization of the unit in which the property values are expressed, and for which each certified value is accompanied by an uncertainty at a stated level of confidence

NOTE 1: The concept of value includes qualitative attributes such as identity or sequence. Uncertainties for such attributes may be expressed as probabilities

NOTE 2: Metrologically valid procedures for the production and certification of reference materials are given in, among others, ISO Guides 34 and 35
NOTE 3: ISO Guide 31 gives guidance on the contents of certificates

**Certified Reference Culture (CRC):** Microbiological; a reference culture certified by technically valid procedure, accompanied by or traceable to a certificate or other documentation which is issued by a certifying body; e.g., cultures used for verifying test systems, validation of methods. Cultures used for QC tests of media must include strains traceable to a type culture collection, where feasible.

**Coefficient of Variation (CV%):** 100 times the ratio of the standard deviation to the mean, expressed as a percentage, e.g., a CV of 20% means that the standard deviation is 0.2 times the mean. The CV is sometimes referred to as the Relative Standard Deviation.

**Common Cause Variation:** See inherent variation.

**Consensus Distribution:** The test sampling distribution used by a laboratory for evaluating laboratory performance within a quality assurance program.

**Consensus Standard:** A reference standard for a test agreed to by a group of laboratories as representing a value that can be used for proficiency testing.

**Confidence Interval:** A possible range of values for a parameter of interest (e.g., analyte concentration in a test sample), constructed from the observed result, based on the sampling procedure and method of measurement, so that this range has a specified probability of including the true value of the parameter (over identically repeated sampling, given all things being equal except for specified random variation). The specified probability (e.g., 95%) is called the confidence level, and the end points of the confidence interval are called the confidence limits or bounds.

NOTE: Confidence intervals reflect only the effects of random errors. They do not take systematic errors (bias) into account.

**Confirmation:** The unambiguous determination of an analyte's presence.

**Controlled Document:** A document subjected to controls that will ensure that the same version of the document is held by or is available to all personnel to whom the document is applicable.

**Control Chart:** See Shewhart Chart.

**Control Limit:** A line placed on a Shewhart Chart that is three standard deviations above, (Upper Control Limit, UCL), or below, (Lower Control Limit, LCL), the process average or process target value.

**Control Measure (CM):** An action or activity that is used to assure that a Performance Criterion (PC) or a Performance Standard (PS) is met.

**Controlled Variation:** Variation that is both expected and predictable over time. It is indicative of points falling randomly between the control limits on a Shewhart Chart (also see Inherent Variation).

**Count (C) Chart:** A Shewhart Chart that plots obtained counts on samples versus sample number.

**Counts per Unit (U) Chart:** A Shewhart Chart that is plots obtained the ratio of count versus sample size on samples versus sample number.

**Coverage Factor:** A numerical factor used as a multiplier of the standard deviation to determine a confidence interval.

**Covariance:** A measure of strength of association of two variables, x and y, calculated as: the expected value of the product of the deviations of the two variables for the same units from their respective expected values:

**Correlation:** A standardized measure of association of two variables equal to the ratio of the covariance divided by the product of the two standard deviations.

**CUSUM Chart:** A cumulative sum chart that cumulates, over successive samples, deviations from some target value. For charting purposes, the lower bound for the CUSUM value (for detecting positive bias) and an upper bound for the CUSUM value (for detecting negative bias) are imposed. This chart is especially useful if one wishes to detect moderate biases (relative to the standard deviation).

**Discrete Test** – A test for which the measurement procedure yields possible answers that can be mapped into the set of whole numbers.

**Empirical Method:** A method that determines a value that can only be arrived at in terms of the method per se and serves by definition as the only method for establishing the accepted value of the item measured.

**Measurement Error:** The difference between an individual test result and the true value of the measurand.

**Expected value**: The expected value of a quantity is the weighted average of all possible values of that quantity within some defined population of units that are assigned values of the quantity, where the weight for an individual value is equal to the probability of obtaining that value. The designation of the expected value is: E(x), where x is the variable of interest – considered as a random variable - and E is the expected value operator.

**False Negative:** A test result that wrongly identifies an analyte as absent, when in fact it is present.

**False Negative Rate:** The probability of a false negative (given that the analyte is present).

**False Positive:** A test result that wrongly identifies an analyte as present, when in fact it is absent.

**False Positive Rate:** The probability of a false positive (given that the analyte is absent).

**Fitness for Purpose:** The degree to which data produced by a measurement process enables a user to make technically and administratively correct decisions for a stated purpose.

**Food Safety Objective (FSO):** The maximum frequency and/or concentration of a hazard in a food at the time of consumption that provides an acceptable level of risk for the designated population.

**Harmonized Collaborative Validation (HCV):** HCV provides the highest measure of ruggedness in analytical methods. Method performance is characterized in a specified number of laboratories. (See also Interlaboratory Study.)

**Individual (Xi) Chart:** A Shewhart Chart that plots the results versus sample number.

**Inherent Variation:** Variation that is due to many random, unknown, events that affect the true results associated with individual units of some (homogeneous) population to differ from the expected value,

**Interlaboratory Study:** A study in which several laboratories measure a specified quantity in one or more "identical" portions of sufficiently homogeneous, stable materials under documented conditions, the results of which are compiled into a single document.

NOTE: The larger the number of participating laboratories, the greater the confidence that can be placed in the resulting estimates of the statistical parameters. The IUPAC-1987 protocol (Pure & Appl. Chem., 66, 1903-1911(1994)) requires a minimum of eight laboratories for method-performance studies.

**Interlaboratory Comparisons:** The organization, performance and evaluation of tests on the same or similar test items by two or more laboratories in accordance with predetermined conditions.
[ISO 13528:2005]

**Known Value (see reference material and recovery):** A value of some measurand that has measurement uncertainty which is considered insignificant to the extent that any value within the confidence interval associated with the measured value would not affect the evaluation of the true value.

**Laboratory:** A body that calibrates and/or tests.

**Limit of Detection (LOD):** The lowest concentration of analyte or level of measurand that can be reliably (with specified degree of confidence, e.g., 97.5% or 2 standard deviations above the

mean blank value) be observed or found in the sample matrix by the method used, when compared to the reagent blank or method tissue blank.

**$LOD_{50}$:** The concentration of analyte or level of measurand at which 50% of replicate samples are positive (e.g., exceed 2 standard deviations above the mean blank value) and 50% of replicate samples are negative.

**$LOD_{90}$:** The concentration of analyte at which 90% of replicate samples are positive and 10% of replicate samples are negative.

**Limit of Quantitation (LOQ):** The smallest measured amount of analyte in a sample that can be reliably quantified with a specified degree of precision.

**Linear Range:** The range of analyte concentrations over which instrumental or method responses are directly proportional to concentration.

**Lower Control Limit (LCL):** The value that is three standard deviations below a process average or target, or at some specified (low) percentile of a presumed distribution. On a Shewhart chart, the value is depicted as a line, for which an out of control signal occurs when a plotted point is below the line.

NOTE: An out of control signal in the case when monitoring microbiology or chemical characteristics would be interpreted here as a process improvement.

**Matrix:** The material or compound in which an analyte is retained.

**Mean or Sample Mean:** The sum of the individual sample values in a set divided by the number of values.

**Measurand:** The particular quantity subject to measurement.

**NOTE 1:** For example, vapor pressure of a given sample of water at 20 °C.

**NOTE 2**: The specification of a measurand may require statements about quantities such as time, temperature and pressure.

**Measurement Uncertainty:** A parameter, associated with the result of a measurement, which characterizes the dispersion of values that could reasonably be attributed to the measurand. (VIM)

NOTE: A measure of the reliability of an analytical result arising from random variation of measuring a measurand by some (specified) procedure. For a single quantity, for purposes of the AOAC, the measure is typically expressed as a confidence interval with finite confidence limits, symmetrical about a “central” estimate of the measurand. An exception would be in the situation where all sample results are non-detect (negative), or below a certain threshold, for which a confidence interval for the percentage positive or above the threshold would range from 0% to some upper confidence bound. Unless specified otherwise, the confidence level of the measurement of uncertainty is 95%.

**Method Detection Limit:** See **Limit of Detection.**

**Method:** A series of steps for performing an activity (e.g. sampling, analysis, quantification), systematically presented in the order they are to be executed.

**Moving Range (MR) Chart:** A Shewhart Chart plots the range of consecutive sample results versus the higher of the two sample numbers.

**Multi-Laboratory Validation (MLV):** MLV is a collaborative study of an analytical method for which repeatability and reproducibility are measured in at least two laboratories.

**Non-conformity:** The non-fulfillment of a requirement to a standard or guideline.

**Normal Distribution:** The probability distribution commonly referred to as the bell-shaped curve/ distribution. Sixty-eight (68) percent of results are expected to fall within one standard deviation (SD) of the mean and 95% within 2 SD of the mean. It is the distribution most often observed when measurement values are from a population for which the deviations from the expected value are due to inherent variation (see above). Under controlled situations, (processes or laboratory methods) the distribution of measured values can be estimated by assuming a normal distribution. The average of sufficiently many results can be often assumed to be normal distributed.

**Number (NP) Chart:** A Shewhart Chart that plots the obtained number of units, considered as one sample, which have the characteristic of interest versus the sample number.

**Outliers:** Specific value(s) from a set of values obtained from samples, considered not to belong to the same distribution of the other sample values, based on a statistical test, typically with specified $\alpha$-probability (e.g., 5% or less).

**Out of Control (of a process)**: The situation in which factors, not expected within normal operating conditions, are affecting the process output results.

**Out of Control Signal (for a process):** Results from a quality control sampling plan for which there is a low probability of occurrence, when the process is assumed to not to be out of control (See Statistical Process Control).

**Performance Characteristics:** Measurable outcomes of a method's behaviour derived from sample analysis, e.g. Accuracy, precision, recovery, specificity (selectivity), sensitivity (limits of detection), inclusivity, exclusivity linearity, range, scope of application,

**Performance Criterion (PC):** An output quantity and criterion that the output quantity must satisfy in order to provide or contribute to meeting a Performance or Food Safety Objective.

**Performance Objective (PO):** The maximum frequency and/or concentration of a hazard in a food at a specified step in the food chain before the time of consumption that provides or contributes to a Food Safety Objective or Appropriate Level of Protection, as applicable.

**Performance Standard (PS):** "Objectively measurable" quantities and a set of criteria associated with a process that is used to control some hazard, and is (often) imposed by government regulations. The PS statement envelopes the PC, PO or FSO and control measures (CM) and often is indistinguishable from one of these. The requirement of "objectively measurable" implies that the measures that are used must be easy to obtain, transparent, and reproducible. A consequence of this requirement is that the risk-effects of actions that are needed to comply with a performance standard may not be directly measurable or determinable.

**Performance Verification:** See Verification.

**Poisson Probability Distribution:** A distribution of the set of non-negative integers (e.g., counts) typically used when sampling from a medium for which the concentration, density or level (per gram or ml) is uniformly distributed throughout the medium. The distribution is characterized by one parameter, which is the expected value of the counts.

**Precision:** A measure of the expected closeness of agreement between independent test results under stipulated conditions; the square root of the expected value of the square of the difference of two "independent" results, given the stipulated conditions.

NOTE: Precision may also refer to the defined quantity divided by the square root of 2, which would provide an estimate of the standard deviation of individual results.

**Probability (of a value):** A number between zero and 1, inclusive, which is coupled with the value that is assigned to units of some population (of units). The quantity is equal to the proportion of times that the value is obtained when, either all the units with their corresponding values are listed, or when the units are randomly selected, an arbitrary large number of times (or infinite number of times), such that each unit has the same "chance" of being selected.

NOTE: The latter part of this definition is somewhat circular since the definition includes the notions of randomness and equal chance which are probabilistic notions. The definition is based on a frequentist view point – the implication is that probabilities and thus statistical characterizations of method and process performance cannot be made unless data are collected and the probabilities are estimated using statistical procedures. For a further discussion of probability, see von Mises (1951, "Probability, Statistics, and Truth, 2nd revised English ed, prepared by H. Geiringer, The MacMillan Co, NY, 1957).

**Procedure:** A series of detailed processes that impact on an analytical outcome.

**Process:** A step or steps that transform inputs (materials, labor, energy, methods and machines) into measurable outputs.

**Process Control Testing** - Sampling product of a process, and making measurements on the samples to determine whether the process is in control or not.

**Proportion (P) Chart:** A Shewhart Chart that plots the proportion of individual units (results from the units recorded together) having the characteristic of interest.

**Proficiency Testing:** Determination of laboratory calibration or testing performance by means of interlaboratory comparisons or comparison to assigned value of analyte.

**Proficiency Test Sample:** Test material with microorganisms or chemical analytes that is tested periodically by a number of locations to determine the proficiency of recovery, using statistical analysis where appropriate.

**Qualitative Method:** A method that identifies analytes based on chemical, biological, or physical properties and that gives a result in the form of presence or absence in a certain size of test portion.

**Quality Assurance (QA):** Those systematic activities, defined by management, that are done outside of the actual analysis to provide confidence that the analysis will satisfy given requirements for quality.
Note: Examples of these activities include training, audit and review.

**Quality Control (QC):**

1) Those activities that are performed during the analysis to fulfill the requirements for assuring quality. Examples include control charting, blank determinations, spiked samples, repeat determinations and blind samples.
2) Activities performed by an establishment to assure that process controls are not "out of control." Sampling of product and plotting results on a Shewhart control chart is an example of a QC activity.

**Quality Control Sample (QCS):** A test portion sample with known contents of analytes to carry through the entire method to verify and monitor laboratory performance.

**Quantitative Method:** A method that identifies analytes and provides an estimate of the amount present in the test sample, expressed as a numerical value in appropriate units, with trueness and precision fit for the purpose.

**Range:**

1. The range of an analytical procedure is the interval between the upper and lower concentration (amounts) of analyte in the sample (including these concentrations) for which it has been demonstrated that the analytical procedure has a suitable level of uncertainty.
2. When used to measure and control variation as with the Range (R) Shewhart Chart the range is the largest value in a subgroup of data minus the smallest value in the same subgroup.

**Range (R) Chart:** A Shewhart Chart that plots the range of results from units considered as one sample, versus sample number.

**Recovery:** The fraction of analyte quantified by the analytical method, expressed as a percentage of the amount "known" to be present in the sample.

**Reference Culture (RC):** A traceable culture with characteristics sufficiently well established to be used to calibrate/verify test systems, test media and validate methods.

**Reference Distribution:** The test sampling distribution used by a laboratory for evaluating its performance (within a quality assurance program).

**Reference Material:** Material characterized by a metrologically valid procedure for one or more specified properties, accompanied by a certificate that provides the value of the specified property, its associated uncertainty, and a statement of metrological traceability

NOTE 1: The concept of value includes qualitative attributes such as identity or sequence. Uncertainties for such attributes may be expressed as probabilities.

NOTE 2: Metrologically valid procedures for the production and certification of reference materials are given in, among others, ISO Guides 34 and 35.

NOTE 3: ISO Guide 31 gives guidance on the contents of certificates

**Relative Percent Difference:** Difference between two values divided by the average of the two, expressed as a percentage.

**Reference Standard:** A standard, generally having the highest metrological quality available at a given location in a given organization, from which measurements are derived.

NOTE: Generally, this refers to recognized national or international traceable standards such as National Institute of Standards and Technology (NIST) thermometers and weights. Other standards may not be traceable to a national standard such as filters for setting wavelengths.

**Relative Standard Deviation (RSD%):** See coefficient of variation.

**Repeatability:** (of results of measurements): The standard deviation of results of measurements of the same measurand carried out under the same conditions of measurement.

NOTES:

These conditions are called repeatability conditions which include:

- the same measurement procedure
- the same observer
- the same measuring instrument, used under the same conditions
- the same location
- repetition over a short period of time

**Repeatability Limit:** The value of which the absolute difference between two test results obtained under repeatability conditions may be expected to be less than or equal to, with a probability of 95%, (also called the critical difference for groups of test results).

**Replicate Test:** An analysis performed more than once. The result of each individual analysis is a replicate test result.

**Reproducibility (of results of measurements):** The standard deviation of results of measurements of the same measurand carried out under changed conditions of measurement, in the broadest sense. Basically it is assumed that the results arise from a population designated by all possible laboratories, analysts, environmental conditions, measuring instrument, reference standard and other pertinent factors that could affect the results. A valid statement of reproducibility requires specification of the conditions changed.

**Reproducibility Limit:** The value of which the absolute difference between two test results obtained under reproducibility conditions may be expected to be less than or equal to, with a probability of 95%.

**Robustness:** A measure of an analytical method capacity to remain unaffected by small but deliberate variations in method parameters and provides an indication of its reliability during normal usage.

**Ruggedness:** The ability of an analytical procedure to resist changes in results when subjected to minor changes in environmental and procedural variables, laboratories, personnel, etc.

**Sample:** Any material brought into the laboratory for analysis.

**Sample Handling:** The manipulation to which samples are exposed during the sampling process, from the selection from the original material through to the disposal of all samples and test portions.

**Sample Preparation:** The process of obtaining a representative test portion from the sample which includes selecting a subsample(s) and in-laboratory processing (i.e. mixing, reducing, coring, quartering, blending, and grinding).

**Sampling:** A procedure whereby a part of a substance, material or product is taken from a well-defined collection of substances, materials, or product, to be used for characterizing, testing or calibrating features of the population.
Two major types of sampling can be identified:

1. Probability or statistical sampling, where the collected material is considered as a "representative" of the whole – that is, the selected units are collected with known probability of selection, which enables deductively statistical based inferences to be made regarding the whole population.
2. Convenience, such as forensic analysis, where the sample is not "representative" of the population, but is determined by availability or convenience, quota (judgment), and for which inferences to a population must be made with the aid of judgment.

NOTE 1: Sampling procedures should describe the sampling plan, selection, withdrawal, and preparation of a sample. The resulting sample "represents" a larger quantity such as a lot or batch.

NOTE 2: The laboratory staff is often not involved in the sampling process, but analysts may be consulted concerning proper sample size (the amount of the sample, such as 25 grams, 5 one pound packages, etc) or the use of appropriate preservatives, and they may be asked to provide suitably prepared containers. ISO 17025 requires that,

where relevant, a statement to the effect that the results relate only to the items tested shall be made.

NOTE 3: Often the probabilities of selection are not known but can be approximated and assumed to be equal to some values without any significant introduction of bias.

**Scope of Application:** The range of matrices to which a method may be applied; usually based on method validation studies.

**Screening Method:** A method designed to detect the presence of an analyte in a sample at or above some specified concentration (target level). .

**Segregate:** Set apart; can represent setting apart by space and time. An example would be the separation (segregation) of samples and standards to avoid cross-contamination.

**Selectivity (Specificity):** The extent to which the analytical method can detect/determine particular analyte(s) in a complex mixture without interference from the other components in the mixture.

**Exclusivity:** The probability that the method will classify a test sample as negative, given that a test sample is a known negative.

**Sensitivity:** The difference in analyte concentration corresponding to the smallest difference in the response of the method that can be detected. It is represented by the slope of the calibration curve.

**Inclusivity:** The probability that the method will classify a test sample as positive given that a test sample is a known positive.

**Shewhart Chart:** A series of charts developed by Dr. Walter Shewhart in the 1920s to provide labors and management with a system for identifying when processes are operating in a steady state or when processes are not stable. They consist of plotting specified sampled results versus sample number, and could include horizontal lines depicting target values and out of control limits.

**Single Laboratory Validation (SLV**): SLV is a single laboratory study of an analytical method which determines performance characteristics other than reproducibility

**Special Cause Variation:** Variation that is unexpected and unpredictable over time. It is a type of variation that is responsible for causing “Out of Control Signals.”

**Standard Deviation:** The square root of the variance.

**Standard Deviation (s) Chart:** A Shewhart Chart that plots the standard deviation (described above) of the results of units considered as one sample versus sample number.

**Statistical Process Control (SPC):** A system and philosophy about maintaining control in a manufacturing environment that requires one to measure characteristics of process output when the process is presumably in control, using statistical methods, developing criteria based on these results using probability theory, and plotting results using, for example, a Shewhart Chart or CUSUSM chart, and then reacting to situations when the criteria are not being met, which could indicate the process is out of control.

**Statistical Process Control Chart:** See Shewhart Chart.

**Standard Uncertainty:** Uncertainty of the result of a measurement expressed as a standard deviation [GUM].

**System Suitability:** The fitness of instruments for the purpose at hand based on manufacturer specifications, instrumental Standard Operating Procedures, or specific requirements of the method.

**Test:** A technical operation that consists of the determination of one or more characteristics or the performance of a given product, material, equipment, organism, physical phenomenon, process or service according to a specified procedure.

NOTE: The result of a test is normally recorded in a document sometimes called a test report or a test certificate.

**Testing Laboratory:** A laboratory that performs tests.

**Test Method: A** specified technical procedure for performing a test.

**Test Portion:** The actual material weighed or measured for the analysis.

**Test Sample:** Material prepared from the laboratory sample and from which test portions will be taken.

**Traceability:** The property of the result of a measurement or the value of a standard whereby it can be related to stated references, usually national or international standards, through an unbroken chain of comparisons all having stated uncertainties.

**Trueness:** See bias.

**Type I Error:** The error of classifying test results as not belonging to an assumed distribution (often called the "null' hypothesis) when it actually does belong. (See alpha $\alpha$ – probability).

**Type II Error:** The error of classifying test results as belonging to an assumed distribution when it actually does not belong. (See beta $\beta$ – probability).

**Uncertainty:** See Measurement Uncertainty.

**Uncontrolled Variation:** Variation presumed to exists when a process experiences an "Out of Control Signal."

**Upper Control Limit (UCL):** The value that is three standard deviations above a process average or target, or at some specified (high) percentile of a presumed distribution. On a Shewhart Chart, the UCL is depicted as a line, for which an out of control signal occurs when a plotted values is above the line.

**Validation:** Establishment, by systematic laboratory studies, of the performance characteristics of an analytical method when applied to specific matrices and/or analytes. Validation may be performed in a single laboratory (SLV), Multi-laboratory (MLV) or by collaborative study (HCV).

**Variables Test** – A test that measures a quantitative value for which possible answers can be approximated by an interval of real numbers (possibly of infinite length), e.g., CFU/g of a microorganism in a food.

**Variance:** The expected value of the squared difference of individual values and the population mean; $E(x-\mu)^2$, where x is a value of the random variable from some distribution over which expected values are taken and $\mu$ is the expected value of x.

**Verification:** Confirmation, through the provision of objective evidence, that the performance characteristics of the method meet the specifications related to the intended use of the analytical results.

**Vulnerability:** A flaw or weakness in a business process (system procedures, design, implementation, or internal controls) that could be exercised and result in a disruption to the process.

**List of symbols used**

**FP** False Positive Test that corresponds to a "positive" result on a sample that is truly negative.

**FN** False Negative Test that corresponds to a "negative" result on a sample that is truly positive.

**NPV** Negative Predictive Value defined as the ratio or percentage of TN/(TN+FN)

**PPV** Positive Predictive Value defined as the ratio or percentage of TP/(TP+FP)

**Prev** Prevalence defined as the ratio or percentage of (TP+FN)/N where N equals TP+FP+TN+FN

**RSD** relative standard deviation equal to the ratio of the sample standard deviation divided by the sample mean which is usually converted to percent

**RSDr** relative standard deviation for a test within a laboratory (intra-laboratory)

**$RSD_R$** relative standard deviation for a test between laboratories (inter-laboratory)

**SDr** repeatability standard deviation for a test.

**$SD_R$** reproducibility standard deviation for a test between.

**Sn** Sensitivity is the probability of correctly detecting the presence of some analyte. This can be expressed as a function of the actual analyte level in the sample; e.g., the test has a sensitivity of 95% at the 0.1 ppb level.

**Sp** Specificity is the probability of correctly not-detecting the presence of some analyte.

**$S_{is}$** standard deviation of initial suspension for an intra-laboratory test equal to the sum of squared differences of identical samples and test protocols differing only in initial suspension conditions

**$S_R$** standard deviation of reproducibility

**Sr** standard deviation of replication and random error for an intra-laboratory test otherwise equal to the sum of squared differences of identical samples and test protocols

**$S_{cond}$** standard deviation of laboratory conditions for an intra-laboratory test equal to the sum of squared differences of identical samples and test protocols but including at least technician and time variability

**TN** Total Negative Tests (excluding FN tests)

**TP** Total Positive Tests (excluding FP tests)

$\overline{X}$ Arithmetic average

**Xi** Individual measure

**z-score** is the standard normal deviate for an observation from a normal sampling distribution which is the difference between the observed value and the expected value divided by the standard deviation.

# Presidential Task Force on Best Practices for Microbiological Methodology (BPMM) Final Roster of Participants

## Steering Committee

Dr. Russell Flowers, Co-Chair
CEO Silliker Group Corporation
900 Maple Road
Homewood, IL 60423
russ.flowers@silliker.com
Tel.: 1+(708) 957-7878
Fax: +1(708)957-8449
Cell: +1(708)259-0470

Robert E. Koeritzer, Co-Chair
Technical Manager
3M Microbiology Products
3M Center, Building 260-06-B-01
St. Paul, MN 55144-1000
rekoeritzer@mmm.com
Tel.: 651-736-6115
Fax: 651-733-1804

Michael H. Brodsky
President
Brodsky Consultants
73 Donnamora Crescent
Thornhill, ON L3T-4K6
CANADA
mhbrodsky@rogers.com
Tel.: (416) 816-9837
Fax: (905) 889-2276
Cell: (416)816-9837

Darrell W. Donahue, Ph.D.
Associate Professor and Coordinator of
Biological Engineering Department of
Chemical
And Biological Engineering
University of Maine
5737 Jenness Hall, Room 309
Orono, ME 04469-5737
ddonahue@umche.maine.edu
Tel: 207-581-2728
Fax: 207-581-2323

Bertrand Lombard, Ph.D.
Agence Francaise De Securite Sanitaire Des
Aliments (AFSSA)
23 Avenue Du General De Gaulle
F94 706 Maisons-Alfort
FRANCE
b.lombard@afssa.fr
Tel.: +33 1 49 77 26 96
Fax: +33 1 43 68 97 62

Daniel W. Tholen, M.S.
Dan Tholen Statistical Consulting
823 Webster Street
Traverse City, MI 49686
tholen@traverse.com
Tel.: 231-929-1721
Cell: 231-631-3591
Fax: 231-941-9713

## Detection Limits WG Chair

Philip T. Feldsine
President and CEO
BioControl Systems, Inc.
12822 SE 32nd Street
Bellevue, WA 98005-4318
ptf@biocontrolsys.com
Tel.: 425-603-1123
Fax: 425-603-0070

# Presidential Task Force on Best Practices for Microbiological Methodology (BPMM) Final Roster of Participants

## Guest Participants on the Steering Committee

Dr. Mark Coleman
Chair, Official Methods Board
Senior Research Scientist
Eli Lilly & Co.
Elanco Animal Health (GL10)
2001 W. Main Street
Greenfield, IN 46140-0708
Coleman_mark_r@lilly.com
Tel.: 317-277-4613
Fax: 317-277-4167

Dr. James R. Agin
Chair, Methods Committee on Microbiology
Microbiology Supervisor
Ohio Department of Agriculture
Building 3, Consumer Analytical Lab
8995 E. Main Street
Reynoldsburg, OH 43068
agin@odant.agri.state.oh.us
Tel.: 614-728-0198
Fax: 614-728-6322

# Presidential Task Force on Best Practices for Microbiological Methodology (BPMM) Final Roster of Participants

## Detection Limits Working Group

Philip T. Feldsine, Chair
President and CEO
BioControl Systems, Inc.
12822 SE 32nd Street
Bellevue, WA 98005-4318
ptf@biocontrolsys.com
Tel.: 425-603-1123
Fax: 425-603-0070

Robert E. Koeritzer
Technical Manager
3M Microbiology Products
3M Center, Building 260-06-B-01
St. Paul, MN 55144-1000
rekoeritzer@mmm.com
Tel.: 651-736-6115
Fax: 651-733-1804

Dr. Nick Cirino
Wadsworth Center
120 New Scotland Avenue
Albany, NY 12208
ncirino@wadsworth.org
Tel.: (518) 474-1838
Fax: (518) 222-5160

Dr. Graham Vesey
Chief Officer for BioBall
BTF Pty Ltd.
P.O. Box 599
North Ryde BC NSW 1670
AUSTRALIA
Graham.vesey@btfbiocom
Tel: 011 61 2 8877 9110 (direct)
Or 011 61 2 8877 9150 (switchboard)

Donald C. Singer
Head, Microbiological Services
GlaxoSmithKline
1250 S. Collegeville Road
UP 9200
Collegeville, PA 19426
Donald.c.singer@gsk.com
Tel.: 610-917-5751
Fax: 610-917-4183

Dr. Vivian Chi-Hua Wu
Assistant Professor
University of Maine
Department of Food Science & Human Nutrition
5735 Hitchner Hall
Orono, ME 04469-5735
+1 (207) 581-3101
+1 (207) 581-1636 (fax)
vivian.wu@umit.maine.edu

Mary L. Tortorello
National Center for Food Safety and Technology
6502 S. Archer Road
Summit-Argo, IL 60501-1957
Business: +1 (708) 728-4146
Business Fax: +1 (708) 728-4177
E-mail: mary.tortorello@fda.hhs.gov
Web Page: http://www.ncfst.iit.edu
Fax: 301-924-7089

# Presidential Task Force on Best Practices for Microbiological Methodology (BPMM) Final Roster of Participants

**Matrix Extensions Working Group**

Michael H. Brodsky, Chair
President
Brodsky Consultants
73 Donnamora Crescent
Thornhill, ON L3T-4K6
CANADA
mhbrodsky@rogers.com
Tel.: (416) 816-9837
Fax: (905) 889-2276
Cell: (416)816-9837

Bertrand Lombard, Ph.D.
Agence Francaise De Securite Sanitaire Des Aliments (AFSSA)
23 Avenue Du General De Gaulle
F94 706 Maisons-Alfort
FRANCE
b.lombard@afssa.fr
Tel.: +33 1 49 77 26 96
Fax: +33 1 43 68 97 62

Axel Colling
D.V.M., Dr. med. vet.
Veterinary Diagnostic Scientist
Diagnosis, Surveillance and Response Unit
Australian Animal Health Laboratory (AAHL)
CSIRO Livestock Industries
5 Portarlington Road
Private Bag 24
Geelong, Victoria 3220
AUSTRALIA
axel.colling@csiro.au
Tel. 0061-3-5227-5255
Fax 0061-3-5227-5555

Christina Egan, Ph.D.
Wadsworth Center
NYSDOH
120 New Scotland Ave.
Albany, NY 12208
eganc@wadsworth.org
518-474-4177
Fax 518-486-7971

Arvind Bhagwat
Research Microbiologist
USDa-ARS
Produce Quality and Safety Lab
BARC – West, Building 002, Room 117
Beltsville, MD 20705-2350
bhagwata@ba.ars.usda.gov
Tel.: 301-504-5106
Fax: 301-504-0632

William C. Cray, Jr. Ph.D.
Chief, Microbiology Branch
USDA/FSIS/OPHS/Eastern Laboratory
950 College Station Road
Athens, GA 30605
William.cray@fsis.usda.gov
Tel.: 706-546-3120
Fax: 706-546-3589

Lee Ann Jaykus
North Carolina State University
Food Science Department
Campus Box 7624
318 Schaub Hall
Raleigh, NC 27695
Leeann_jaykus@ncsu.edu
Tel.: 919-513-2074
Fax: 519-515-7124

Dr. Forest D. Reichel
Director of Marketing And Sales
Microbiologics, Inc
217 Osseo Avenue North
Saint Cloud, MN 56303

Dr. Karl Friedrich Eckner
Research Manager
Norsk Matanalyse -Oslo
Nils Hansens V 4
Postboks 6166 Etterstadd
Oslo, 0602
NORWAY

# Presidential Task Force on Best Practices for Microbiological Methodology (BPMM) Final Roster of Participants

## Matrix Extensions Working Group Advisors

Wallace Andrews
FDA-CFSAN
Div Microbiological Studies HFS 516
Room 3E-023
5100 Paint Branch Parkway
College Park, MD 20740
Wallace.Andrews@cfsan.fda.gov
Tel.: 301-436-2008
Fax: 301-436-2644

Reginald Bennett
Supervisory Microbiologist
DHHS/FDA/CFSAN/OO/OPDF/DMS/
MMRB
301-436-2009
reginald.bennett@fda.hhs.gov

Carl Custer
USDA FSIS OPHS
STOP 3777(Room 344 The Aerospace Center)
1400 Independence Avenue, SW
Washington, DC 20250
Carl.Custer@FSIS.usda.gov
Tel.: (202) 690-6645
Fax: (202) 690-6364

Thomas Hammack
FDA-CFSAN
Div of Microbiological Studies
5100 Paint Branch Pkwy
College Park, MD 20740
Thomas.hammock@cfsan.fda.gov
Tel.: 301-436-2010
Fax: 301-436-2644

Lihan Huang
USDA Agricultural Research Service
Eastern Regional Research Center
600 E. Mermaid Lane
Wyndmore, PA 19038
lhuang@errc.ars.usda.gov
Tel.: 215-233-6552
Fax: 215-233-6406

James Hungerford
USFDA
Seafood Product Research Center
22201 23rd Dr SE
Bothell, WA 98021
James.Hungerford@fda.gov
Tel.: 425-483-4894
Fax: 425-483-4996

Vijay Juneja
USDA-ARS-ERRC
600 E. Mermaid Lane
Wyndmoor, PA 19038
vjuneja@errc.ars.usda.gov
Tel.: 215-233-6500
Fax: 215-233-6697

Dr. John Marugg
QS / Microbiological Safety
Nestlé Research Center,
Lausanne, France

Ynes R. Ortega, PhD, MPH
Center for Food Safety
University of Georgia
1109 Experiment St.
Griffin, GA 30223
Phone and fax: 770-233-5587
E-Mail: ortega@uga.edu

Joan Pinkas
McCormick & Co., Inc.
Corporate Research & Development
Hunt Valley,
MD 21053 USA
joan_pinkas@mccormick.com
Tel: 410-436-7811
Fax: (410)527-8022

Katherine J. Swanson, PhD
Vice President, Food Safety,
Ecolab, Inc.
St. Paul, Minnesota

# Presidential Task Force on Best Practices for Microbiological Methodology (BPMM) Final Roster of Participants

Sterling Thompson
Senior Manager of Microbiological Research & Services
Hershey's Chocolate
sthompson@hersheys.com
Tel: 717-534-5207

Mary W. Trucksess, Ph.D.
Research Chemist
FDA-CFSAN
Division Of HFS-840
5100 Paint Branch Parkway, Rm. 4EL18
College Park, MD 20740-3835
mtruckse@cfsan.fda.gov
Tel.: (301) 436-1957
Fax: (301) 436-2665

David D. Wagner, Ph.D.
Research Animal Scientist
FDA-CVM
8401 Muirkirk Road
HFV-520
Laurel, MD 20708
301-210-4350
david.wagner@fda.gov

Michael H. Brodsky, Chair
President
Brodsky Consultants
73 Donnamora Crescent
Thornhill, ON L3T-4K6
CANADA
mhbrodsky@rogers.com
Tel.: (416) 816-9837
Fax: (905) 889-2276
Cell: (416)816-9837

Bertrand Lombard, Ph.D.
Agence Francaise De Securite Sanitaire Des Aliments (AFSSA)
23 Avenue Du General De Gaulle
F94 706 Maisons-Alfort
FRANCE
b.lombard@afssa.fr
Tel.: +33 1 49 77 26 96
Fax: +33 1 43 68 97 62

Axel Colling
D.V.M., Dr. med. vet.
Veterinary Diagnostic Scientist
Diagnosis, Surveillance and Response Unit
Australian Animal Health Laboratory (AAHL)
CSIRO Livestock Industries
5 Portarlington Road
Private Bag 24
Geelong, Victoria 3220
AUSTRALIA
axel.colling@csiro.au
Tel. 0061-3-5227-5255
Fax 0061-3-5227-5555

Christina Egan, Ph.D.
Wadsworth Center
NYSDOH
120 New Scotland Ave.
Albany, NY 12208
eganc@wadsworth.org
518-474-4177
Fax 518-486-7971

# Presidential Task Force on Best Practices for Microbiological Methodology (BPMM) Final Roster of Participants

## Sampling Working Group

Darrell W. Donahue, Ph.D., Chair
Associate Professor and Coordinator of Biological Engineering Department of Chemical And Biological Engineering
University of Maine
5737 Jenness Hall, Room 309
Orono, ME 04469-5737
ddonahue@umche.maine.edu
Tel: 207-581-2728
Fax: 207-581-2323

Dr. Russell Flowers
CEO Silliker Group Corporation
900 Maple Road
Homewood, IL 60423
russ.flowers@silliker.com
Tel.: +1(708) 957-7878
Fax: +1(708)957-8449
Cell: +1(708)259-0470

Richard (Dick) Whiting
U.S. Food & Drug Administration
Center for Food Safety and Nutrition
5100 Paint Branch Parkway, HFS-302
College Park, MD 20740-3835
rwhiting@cfsan.fda.gov
Tel.: 301-436-1925
Fax: 301-436-2632

Michael S. Curiale
Nestle
201 Housatonic Ave.
New Milford, CT 06776
Tel.: 860 355-6261
michael.curiale@rdct.nestle.com

Ian Jenson
Food Safety Program Manager
Meat & Livestock Australia
Locked Bag 991
North Sydney NSW 2059
AUSTRALIA
ijenson@mla.com.au
Tel.: (02) 9463 9264
Fax: (02) 9463 9182
m: 0408 602 903

Daniel Zelenka
Director of Statistics
Tyson Foods, Inc.
2210 Oaklawn Drive
Springdale, AR 72762
+1 (479) 290-2970
dan.zelenka@tyson.com

Dr. Adriaan M.H. van der Veen
NMi Van Swinden Laboratorium
Department of Energy
(Visiting address:)
Schoemakerstraat 97
2628 VK Delft (NL)
(Mailing address:)
Postbus 654
2600 AR Delft (NL)
Tel.: +31 15 2691 733
Fax.: +31 15 261 29 71
avdveen@nmi.nl

David D. LaBarre DVM
USDA/FSIS/OPHS/RAD/TAEB
333 Aerospace Center Rm 389
1400 Independence Ave. S.W.
Washington, DC 20520-3700
202 690 6424 (voice)
202 690 6337 (fax)
David.LaBarre@fsis.usda.gov

# Presidential Task Force on
# Best Practices for Microbiological Methodology (BPMM)
# Final Roster of Participants

Nelson Clinch

USDA/FSIS
Room 202 Annex
300 12th St., SW
Washington, DC 20250
Phone: 202-720-7471 OR 202-720-3219
FAX: 202-690-0824
E-mail: nelson.clinch@fsis.usda.gov

Harry Marks
USAD/FSIS
14th and Independence Ave. SW
Cotton Annex Rm. 112
Washington DC 20250
Telephone (202) 720-5857
e-mail harry.marks@fsis.usda.gov

Cathy Pentz
Chief, Microbiology Quality Assurance Branch
USDA, FSIS, OPHS, Laboratory QA/QC Division
950 College Station Rd.
Athens, GA 30629 cathy.pentz@fsis.usda.gov
Tel.: 706-546-3570

# Presidential Task Force on Best Practices for Microbiological Methodology (BPMM) Final Roster of Participants

## Statistics Working Group

Daniel W. Tholen, M.S., Chair
Dan Tholen Statistical Consulting
823 Webster Street
Traverse City, MI 49686
tholen@traverse.com
Tel.: 231-929-1721
Cell: 231-631-3591
Fax: 231-941-9713

Prof. Basil Jarvis
Ross Biosciences, Ltd.
Daubies Farm
Upton Bishop
Ross-on-Wye
Herefordshire HR9 7UR
UNITED KINGDOM
01989 720698
basil.jarvis@btconnect.com
Tel: +44-(0)1989-720-698
Cell: +44-(0)7778-23-33-65
Fax: +44-(0)1989-720-154

Bertrand Lombard, Ph.D.
Agence Francaise De Securite Sanitaire Des Aliments (AFSSA)
23 Avenue Du General De Gaulle
F94 706 Maisons-Alfort
FRANCE
b.lombard@afssa.fr
Tel.: +33 1 49 77 26 96
Fax: +33 1 43 68 97 62

Mark A. Mozola
V.P. of Research & Development
Neogen Corp.
620 Lesher Place
Lansing, MI 48912
mmozola@neogen.com
Tel.: (517) 372-9200
Fax: (517) 372-0108

Dr. Daryl S. Paulson
BioScience Laboratories, Inc.
300 North Wilson, Suite 1
P.O. Box 190
Bozeman, MT 59771-0190
dpaulson@biosciencelabs.com
Tel.: 1-877-858-2754 or 1-406-587-5735
tanderson@biosciencelabs.com
Fax: 406-586-7930

Dawn M. Mettler
Rockbridge Laboratory Services
Food Safety Consultant
24671 Woltz Road
Rockbridge, OH 43149
dmettler@hocking.net
Tel.: 740-380-3012
Fax: 740-385-7207

Kenneth Newton
National Measurement Institute Headquarters
PO Box 264
Lindfield, NSW 2070
AUSTRALIA
Tel.: (61 2) 8467 3600
Fax: (61 2) 84 67 3610
Ken.Newton@measurement.gov.au

Dr. Anthony D. Hitchins
Research Microbiologist
U.S. Food & Drug Administration
HFS 516, Room 3E-024
5100 Paint Branch Parkway
College Park, MD 20740-3835
Anthony.hitchins@cfsan.fda.gov
Tel.: 301-436-1649
Fax: 301-436-2644

# Presidential Task Force on Best Practices for Microbiological Methodology (BPMM) Final Roster of Participants

**AOAC Staff Liaisons to the BPMM Task Force**

Sharon L. Brunelle, Ph.D.
Lead Microbiologist
AOAC INTERNATIONAL
Woodinville, WA
sbrunelle@aoac.org
Tel.: 425-922-1607

E. James Bradford, Ph.D.
Executive Director
AOAC INTERNATIONAL
481 N. Frederick Avenue, Suite 500
Gaithersburg, MD 20877
jbradford@aoac.org
Tel.: 301-924-7077 x 102
Fax: 301-924-7089

Arlene Fox
Sr. Director, Proficiency Testing
AOAC INTERNATIONAL
481 N. Frederick Avenue, Suite 500
Gaithersburg, MD 20877
afox@aoac.org
Tel.: 240-924-7077x143
Fax: 301-924-7089

Diana Hopkins
Director, Governance & Executive Affairs
AOAC INTERNATIONAL
481 N. Frederick Avenue, Suite 500
Gaithersburg, MD 20877
dhopkins@aoac.org
Tel.: 301-924-7077 x 101
Fax: 301-924-7089

Deborah McKenzie
Manager, Technical Programs, Research Institute
481 N. Frederick Avenue, Suite 500
Gaithersburg, MD 20877
afox@aoac.org
Tel.: 240-924-7077x157
Fax: 301-924-7089

Anita Mishra
Principal Scientific Liaison, Government and Industry
AOAC INTERNATIONAL
481 N. Frederick Avenue, Suite 500
Gaithersburg, MD 20877
amishra@aoac.org
Tel.: 240-912-1456
Fax: 301-924-7089

Scott Coates
Managing Director
AOAC Research Institute
481 North Frederick Avenue
Gaithersburg, MD 20877
scoates@aoac.org
Tel.: 301-924-7090
Cell: 240-506-4346
Fax: 301-924-7089

Liz Cribbin
Contract Administrator
AOAC INTERNATIONAL
481 N. Frederick Avenue, Suite 500
Gaithersburg, MD 20877
lcribbin@aoac.org
Tel.: 240-912-1471
Fax: 301-924-7089