ISO/ (to be submitted)

COMPARISON OF METHODS FOR DRINKING WATER BACTERIOLOGY – CULTURAL TECHNIQUES

4 January 2000

CONTENTS

1 INTRODUCTION

2 SCOPE OF THE DOCUMENT

2.1 Statistical approach and acceptance criteria

3 BASIC CONCEPTS AND DEFINITIONS

3.1 Microbiological cultural methods
3.2 Validation
3.3 Definitions

3.3.1 Confirmed counts
3.3.2 Overdispersion
3.3.3 Poisson distribution
3.3.4 Precision
3.3.5 Presumptive counts
3.3.6 Primary validation
3.3.7 Repeatability
3.3.8 Reproducibility
3.3.9 Robustness
3.3.10 Secondary validation

4 SOURCES OF VARIATION AND ERROR

4.1 Sample error
4.2 Natural (e.g. random) variation
4.3 Other sources of variation
4.4 Statistical detection of other sources of variation
4.5 Limitation of errors
4.6 A ‘confidence level’ approach

5 PREREQUISITES

5.1 Initial characterisation of a new method
5.2 Common identification of ‘target’ and ‘non-target’ organisms
5.3 False positive and false negative results
5.4 Principles of Quality Assurance

5.4.1 Introduction
5.4.2 Media
5.4.3 Incubators
5.4.4. Membranes

6 COMPARISON OF METHOD A VERSUS METHOD B (STAGE 1)

6.1 Production of samples
6.2 Preparation of spiked samples

6.2.1 Protocol for generation of chlorine stressed-organisms using river water as the source of target organisms
6.2.2 Protocol for production of chlorine stressed-organisms from sewage effluent.

7 INTERPRETATION OF THE DATA (STAGE 1)

7.1 Statistical comparison of methods

7.1.1 Stage 1 - spiked samples (20-50 target organisms/portion)

8 STAGE 2 EVALUATION

8.1 Data analysis for stage 2

9 EVALUATION OF A METHOD FOR NEW USE IN A LABORATORY

9.1 Evolution of method introduction

10 COMPARING A MOST PROBABLE NUMBER (MPN) METHOD WITH AN ENUMERATION METHOD

11 COMPARING TWO MPN METHODS

12 REFERENCES

ANNEX A - WORKED EXAMPLES

COMPARISON OF METHODS FOR DRINKING WATER BACTERIOLOGY – CULTURAL TECHNIQUES

1. INTRODUCTION

Methods for drinking water bacteriology should be capable of serving their intended purpose: to detect or quantify a specified microbe or microbial group with adequate precision and accuracy. A recent ISO Technical Report [1]dealt with the validation of microbiological methods.

In some countries, the methods are specified in legislation and there is therefore no choice. In countries where the methods are not statutory, current regulatory and laboratory accreditation requirements for drinking water analysis require that alternative methods must be of 'equivalent or better' performance when compared with standard methods. Standard methods which have been validated are available from several sources, including those shown in the table below.

However, demonstration that methods are at least as reliable as standard methods is complex. This document provides a protocol for comparing the recoveries of confirmed target organisms by two methods (but is applicable to comparison of more than two) and takes the ISO Technical Report dealing with validation of microbiological methods forward to method comparison.

ISO	International Standards organisation
CEN	Comité Européan de Normalisation (European Committee for Normalization)
AOAC	Association of Official Analytical Chemists
APHA	American Public Health Association
The Microbiology of Water 1994 Part 1 – Drinking Water, Report on Public Health and Medical Subjects No. 71. HMSO. 1994	
Individual standards organisations of each country	

Sources of standard methods

2 SCOPE OF THE DOCUMENT

This document describes the procedures for the comparison of microbiological cultural methods used in drinking water bacteriology against specific criteria. A protocol gives detailed instructions, including the preparation of spiked samples, and recommends the number of compared measurements that will enable the evaluation of a previously validated method for **new** use in a laboratory. The comparisons involve two stages; Stage 1 compares the methods with target organisms at about 20 to 50 organisms per 100 ml, Stage 2 looks at lower levels of target organisms as have been traditionally used. Paired negative samples do not give any useful information on comparative recoveries of target organisms.

2.1 Statistical approach and acceptance criteria

The comparison of the trial method with the reference method must be made with an appropriate diversity of the organisms. The preparation of suitable test samples (Section 6.2) is very important and the waters used should come from several sources (each one to be referred to as a 'category of origin') and be collected over a period of time. In the main stage of the study (Stage 1) samples which give counts lying comfortably within the optimum range of the methods should be used and yield sufficient numbers to give statistical power to the comparison. For example with a membrane filtration method a suitable range would be 20 to 50 target organisms per unit test volume (typically 100 ml). The number of samples to provide sufficient statistical information to reach a conclusion in accordance with the criteria described in this protocol is recommended.

These criteria, described in detail in section 6, require a clear presentation of the data, a statistical comparison within the category of origin of the samples and finally an overall statistical comparison. The new method will be rejected if it shows significantly lower average counts than the reference method. It will be accepted, subject to Stage 2, if it is better or if it is "no different" and the 95% confidence interval for the average difference lies entirely above the value which would indicate that the trial method was finding 10% fewer organisms than the reference method.

Any test method, which is found acceptable, will then be tested in Stage 2 against the reference method with some samples where counts are low. This is to demonstrate that there is no gross change in the contrast between the two methods at contamination levels close to the detection level required for statutory monitoring samples.

3 BASIC CONCEPTS AND DEFINITIONS

A laboratory considering an alternative method to one currently in use needs to obtain sufficient comparative performance data to demonstrate the equivalence of the two methods before adopting the new method. This evidence may also be required by regulators where change is away from a recommended standard method.

A second laboratory may then repeat the process of full comparison and the new method may then become acceptable for use in that laboratory. Subsequent laboratories may also repeat the process until sufficient data have been obtained, pooled and reviewed to establish robustness.

3.1 Microbiological cultural methods

Methods are considered microbiological cultural methods when growth and multiplication are essential for detection or quantification of micro-organisms.

3.2 Validation

Validation of microbiological methods has recently been the subject of an ISO Technical Report, 'Validation of Microbiological Methods'. 1999. [1]

3.3 Definitions

3.3.1 Confirmed counts

That number of presumptive counts multiplied by the proportion confirmed to conform with the target organism.

3.3.2 Overdispersion

Variation in excess of Poisson randomness. Detected by the Poisson index of dispersion and measured quantitatively by estimating the parameter u (overdispersion factor) of the negative binomial distribution.

3.3.3 Poisson distribution

Fully random distribution of particle numbers when sampling a perfectly mixed suspension and there is no attraction or repulsion between organisms.

3.3.4 Precision

Closeness of agreement between independent test results obtained under stipulated conditions.

3.3.5 Presumptive counts

That number of organisms that produce a response typical of the target organism in or on primary detection media.

3.3.6 Primary validation

Establishment of the specifications for the performance of a new method and/or experimental verification that a method meets theoretically derived quality criteria.

3.3.7 Repeatability

Closeness of the agreement between the results of successive measurements of the same measure carried out under the same conditions of measurement.

3.3.8 Reproducibility

Closeness of the agreement between the results of measurements on the same measure carried out under changed conditions of measurement.

3.3.9 Robustness

Insensitivity of an analytical method to small changes in procedure.

3.3.10 Secondary validation

Demonstration by experiment that an established method functions according to its specifications in the user's hands.

4 SOURCES OF VARIATION AND ERROR

Several sources of error may complicate the statistical evaluation of comparing methods for drinking water bacteriology. These sources of error are sample variation, natural variation and systematic variation inherent in the methods.

4.1 Sample variation

A water source which is sampled for monitoring purposes may exhibit enormous variation in microbial content over time and between sampling sites, however close they are [2]. Therefore samples for the comparison of two methods should not be collected or prepared separately. A paired or split sample approach should be used, whereby a suitable sample is thoroughly mixed and two aliquots from this sample are examined at the same time, one by each method. On average, the theoretical expected number of organisms in both aliquots will be the same.

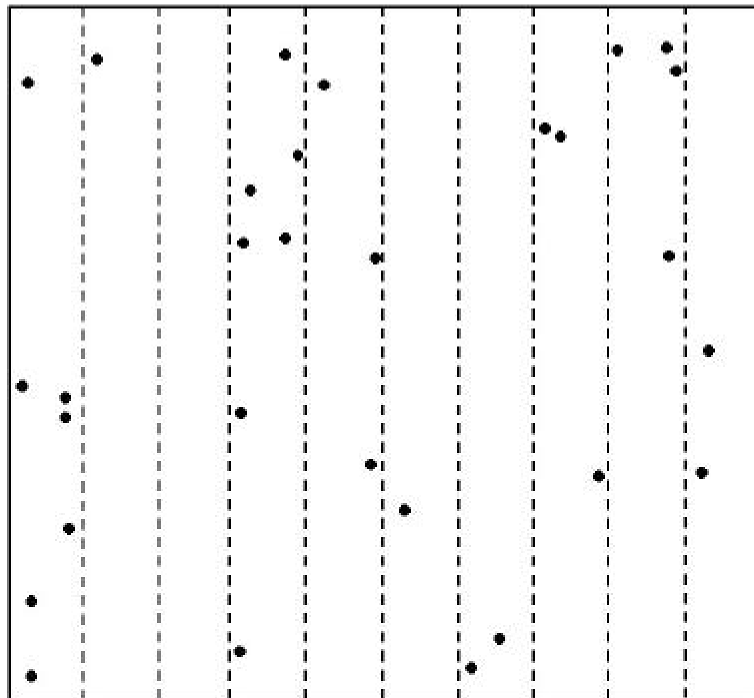


Figure 1: Random variation of organisms in subsamples.

4.2 Natural (e.g. random) variation

For each individual aliquot it is important to note that the exact number of organisms present may differ by chance. Figure 1 illustrates a volume of well-mixed water containing 30 organisms distributed at random. When the volume is divided into 10 aliquots the average number present is 3 but the range is 0 – 7. This random variation is the least possible variation which will be encountered in water microbiology, even when methodology is perfect. There may be additional overdispersion due to organism behaviour if there is a tendency for attraction or repulsion between organisms or between organisms and particles or between organisms and laboratory equipment.

This natural variation means that many samples have to be examined when establishing systemic variation, such as difference between methods. There has to be enough data to

average out the effects of natural variation. The size of this natural variation is illustrated in Figure 2, which shows results from 50 paired samples examined for the same parameter in one laboratory using the same methods. These were natural water samples being used for duplicate split sample Quality Control guidance charts [3], with the aim of checking that variation between split samples was no greater than random. In Figure 2 the correlation between pairs of counts appears low and the results are scattered. In fact, the product-moment statistic (r^2) measures only 0.39 even though this is an example of best achievable correlation in the water microbiology context. This illustrates that the statistic r is inappropriate. Correlation between paired microbial counts needs to be assessed and interpreted against the background of this inevitable variability.

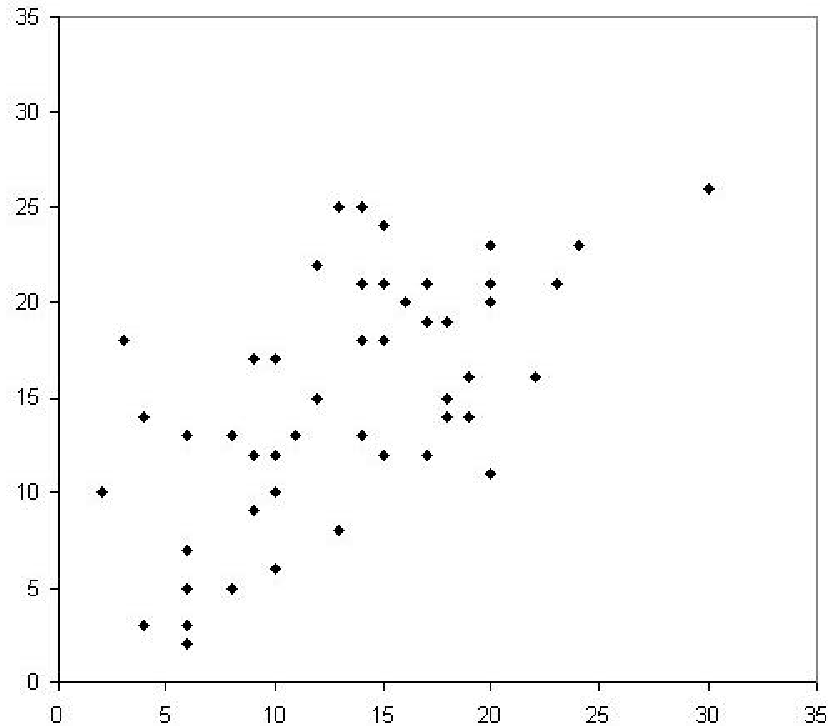


Figure 2: Pairs of replicate counts of total coliform organisms using 50 natural samples (100ml subsamples)

4.3 Other sources of variation

Other factors can affect either the numbers of organisms present or the number detected and reported in the result. Inadequate mixing of samples and inaccurate measurement of aliquot volumes affect numbers present. Errors in numbers reported can be introduced by laboratory procedures, including the methods being studied. A small amount of random measurement error is expected and accepted from every procedure. Excessive random errors would indicate an imprecise method and this should become apparent during the characterisation of a new method (see section 5). Non-random, systematic methodology error, due to inadequacy of the method, is the subject of these initial method validation procedures.

The challenge is to design and analyse a study in which the estimated effects of methodology errors can be separated from the estimated natural variation and random errors. The mathematical rationale for this is described in the next section.

4.4 Statistical detection of other sources of variation

(Adapted from Tillett and Lightfoot. 1995. [4])

Method comparison studies are designed and analysed to detect whether other sources of variation are present, and if they are microbiologically and statistically significant. The sources of variation in the relevant parameter are described as:

$$y_i = \mu + \varepsilon_i$$

where:

$i = 1$ or 2, representing the two aliquots in the paired sample;
y_i is the organism count;
μ is the mean value for the sample;
ε_i is the error.

This can be expanded to:

$$y_i = \mu_t + m_l + m_m + \varepsilon_{ti} + \varepsilon_{li}$$

where:

μ_t is the true mean value;
m_l is the laboratory effect;
m_m is the method effect [reference or trial method];
ε_{ti} is the random or natural error between aliquots;
ε_{li} is the random measurement error in the laboratory.

The laboratory effect plus the method effect is the systemic, average difference from the true mean when that method is used. Its size represents the bias and is inversely proportional to the "trueness" of the measurement.

The random errors reflect precision. The value of the difference between the paired counts is:

$$y_1 - y_2 = (\mu_t + m_l + m_{ref} + \varepsilon_{t1} + \varepsilon_{l1}) - (\mu_t + m_l + m_{trial} + \varepsilon_{t2} + \varepsilon_{l2})$$

$$= (m_{ref} - m_{trial}) + (\varepsilon_1 + \varepsilon_2)$$

The random errors should average zero if sufficient samples are examined. Thus the expected value of $y_1 - y_2$ is:

$$E(y_1 - y_2) = m_{ref} - m_{trial}$$

Any interaction between method and laboratory will be included in this expression but does not affect the conclusions about effectiveness of the trial method in this laboratory.

Because the absolute error values may be large (due to the natural random variation) the precision will be low and a large amount of data will be required for a powerful statistical estimate of $(m_{ref} - m_{trial})$.

4.5 Limitation of errors

Errors should be minimised or eliminated by implementing a Quality Assurance programme which includes the use of Internal Quality Control and participation in External Quality Assessment.

4.6 A 'confidence level' approach

It may not be statistically acceptable simply to take a predetermined number of samples to be examined at each stage of the comparison between two methods. Although it may be possible to theorise as to how many samples would give a 95 % probability of detecting an undesirable difference, the distributions of counts in the samples examined might turn out to be very different from expected. This document outlines a protocol which uses a 'confidence level' approach to provide the statistical power for the evaluation of an accepted method for new use in a laboratory.

The approach for comparing method A with method B is made by recording the difference from paired samples. These data are then progressively evaluated to see if the average results and confidence intervals are comparable.

5. PREREQUISITES

5.1 Initial characterisation of a new method

During method development certain optimum conditions must be established. Primary validation must have established the operational limits and performance characteristics of a new, modified or previously inadequately characterised method. The validation should have resulted in numerical and descriptive specifications for the performance and include a detailed and unambiguous description of the target of interest. An example of a validation experiment is shown in Figure 3.

Primary validation should establish the specificity, selectivity, relative recovery and other characteristics of the new method as described in the ISO Technical Report 'Validation of Microbiological Methods' 1999. [1] The method should have been tested against low numbers of organisms to show linearity over a range for the protocol. Additionally, primary validation must have established that the repeatability and reproducibility of the method is acceptable.

To avoid precision error laboratories should be participating in external quality control schemes.

Standard protocols for the bacteriological examination of water should give details of method performance characteristics and provide details of its specifications e.g. on the scope of the method; incubation robustness and time sensitivity; reliable working limits; target definition and identification. Additionally the protocol should offer advice and make demands on quality control of media and equipment. Protocols should provide laboratories with a structured procedure to assist the application of the methods and therefore obtain valid results. The format for colony count, for example, should include statements on sensitivity, selectivity, counting uncertainty, parallel plating, within-sample variation, and proportionality.

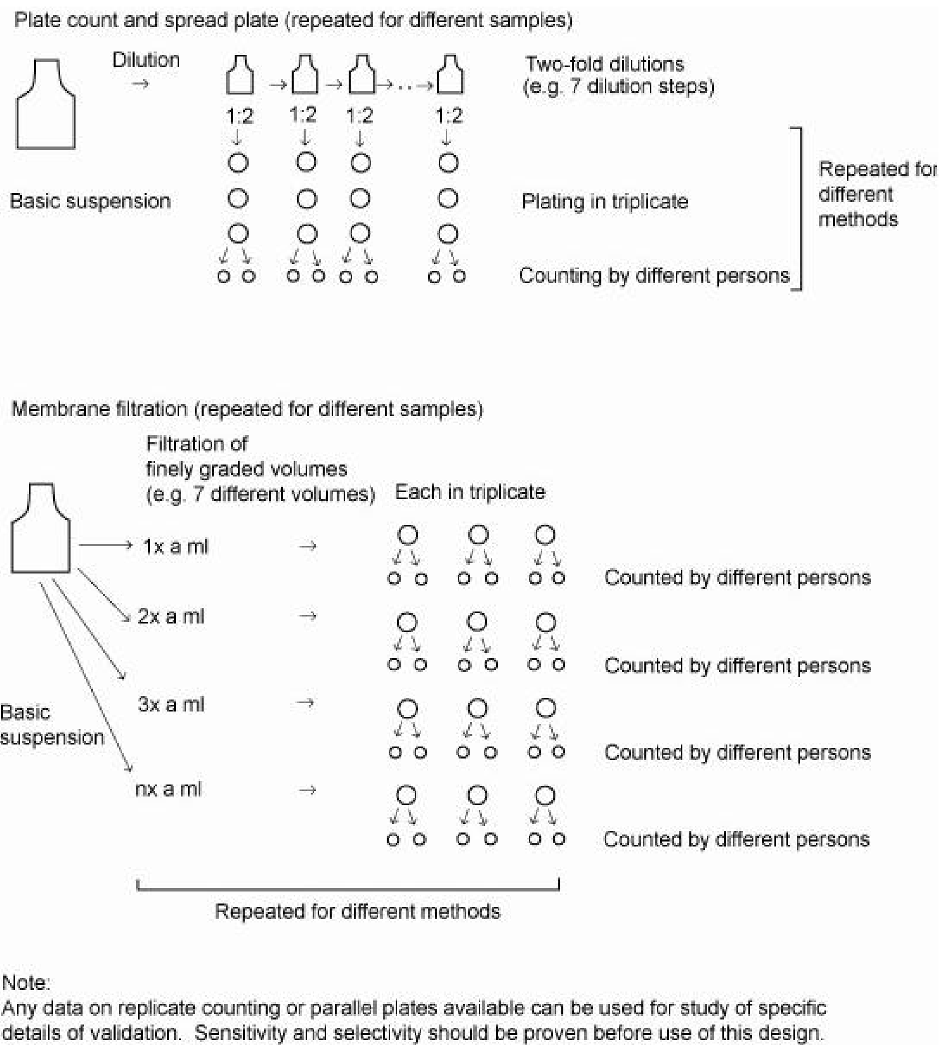


Figure 3: Example of a validation experiment

5.2 Common identification of ‘target’ and ‘non-target’ organisms

Microbiological methods are designed to detect and/or enumerate particular types of micro-organisms, the target-organisms. All other micro-organisms that may be present in the sample should be undetected or be readily differentiated, and therefore should not interfere with the analytical process. These are non-target organisms, also described as competitive or background flora.

The definition of target organisms should reflect the current microbiological understanding which may be reflected in legislation, and be sufficient to ensure common identification of ‘target’ and ‘non-target’ organisms for the two methods being compared.

5.3 False-positive and false-negative results

If a non-target organism is mistakenly identified as a target organism, a false-positive result is obtained. The reverse, a false-negative result, is obtained if a target organism is not identified by a characteristic or 'typical' reaction in the test. Note that a false-positive or a false negative result may be defined for single colonies, but also for the final result of the examination of a sample. The nature and concentration of non-target

organisms and target organisms may vary considerably between samples, certainly from different locations, but also in time at a certain location. This implies that a method that has been evaluated for a particular type of sample does not necessarily have universal applicability. To overcome this problem, methods have been prescribed in International Standards or legal requirements as a way of achieving a standardised approach to analysis by eliminating inter-method differences. It does not necessarily mean the prescribed method is optimum for all situations and the laboratory remains responsible for evaluating the performance of the method for the type of samples under investigation, and for seeking alternatives when necessary. The possible temporal variation of the performance of a method in relation to variable characteristics of the microflora should be evaluated as a part of the quality assurance programme.

5.4 Principles of Quality Assurance

5.4.1 Introduction

It is essential that any laboratory wishing to carry out a microbiological methods comparison according to this guidance must have appropriate quality systems in place. Accreditation, although setting good standards, does not guarantee a laboratory's performance. The sources of error that affect microbiological methods are:

- sampling error - the error due to the taking of subsamples
- natural error - the error due to the poisson distribution of organisms in a liquid matrix
- systematic error - the error inherent in the method used
- random laboratory error - the error that is particular to a laboratory or analyst

In an attempt to reduce the effects of the last two errors to a minimum it is essential that laboratories use external reference in the form of appropriate use of reference materials and take part in recognised external quality assessment schemes.

In addition, a quality assurance programme must be in place and documented in the quality manual used. Information about these and other techniques to ensure quality can be found in "Microbiological Analysis of Food and Water: Guidelines for Quality Assurance", 1998 [5].

Particular attention must be paid to media, incubators and membranes.

5.4.2 Media

The validation of reference materials has revealed that small changes in the composition of the medium can affect its performance, in particular its ability to recover target organisms [6]. There may also be batch to batch variation. The method comparison exercise should therefore use single batches of media that have been performance tested using reference materials. It should be prepared according to manufacturer's instructions, paying attention to the conductivity of the make up water, the pH of the medium (both before and after sterilisation), and minimising the time that the medium is exposed to high temperatures.

5.4.3 Incubators

Temperature of incubation for water microbiology methods has stringent limits defined in ISO standards and in documented national guidance. The temperature in incubators can vary from shelf to shelf and be affected by loading patterns. Incubators should preferably be fan assisted and should be validated using reference thermometers or thermocouples when first commissioned and be recalibrated on a regular basis. In use, the temperatures should be recorded daily preferably by continuous thermocouple monitoring.

5.4.4 Membranes

Membranes differ between manufacturers and between batches and this can affect the recovery of target organisms. Membrane batches should therefore be recorded and acceptance tests using reference materials carried out on each new batch. A single batch should be used for a method comparison exercise.

6. COMPARISON OF METHOD A VERSUS METHOD B (STAGE 1)

The comparison of two methods involves processing aliquots of the same samples by the two methods in parallel, and statistically analysing the results.

The methods must be compared using the type of samples for which the methods are intended. This, generally, will be water which has been subject to treatment, usually including disinfection. The most commonly used disinfectant is chlorine. Because of the high quality of most treated water supplies it will, generally, be necessary to produce samples artificially which mimic the effect of inadequate treatment. Protocols for the production of suitable samples for chlorine disinfected supplies are given in Section 6.2. For alternative disinfectants such as ozone it will be necessary to determine by experiment suitable conditions for sub-optimal disinfection permitting the survival of suitable numbers of target organisms.

There are two stages: Stage 1 compares the two methods with numbers of target organisms within the ideal operating range (20 – 50), and Stage 2 uses low numbers of target organisms to assess equivalence close to the limit of detection. Data on linearity should have been obtained during initial validation experiments.

Initially, a minimum of 150 samples should be included. The methods should be presented with the volume relevant to the standard in question. This is generally 100 ml and this volume is used in this document for illustrative purposes. No dilutions should be made. Spiked samples should be tested over different days, testing up to 10-15 samples per day.

6.1 Production of samples

There are a number of ways of producing suitable samples for a method comparison exercise based on chlorinated waters and these are listed in order of preference.

1. Chlorinated tap water + river water with the addition of chorine to produce chlorine-stressed organisms and a final concentration of chlorine of approximately 0.1 – 0.5 mg/l (protocol 6.2.1).

2. Through treatment samples (e.g. after granulated activated carbon treatment or post rapid gravity filter) if necessary, with a final concentration of chlorine of approximately 0.1 mg/l.

3. Chlorinated tap water + sewage effluent with the addition of chlorine to produce chlorine-stressed organisms and a final concentration of chlorine of approximately 1.2-2.5 mg/l (protocol 6.2.2).

4. Naturally contaminated unchlorinated groundwater with the addition of chlorine to produce chlorine-stressed organisms, with a final concentration of chlorine of approximately 0.1 mg/l.

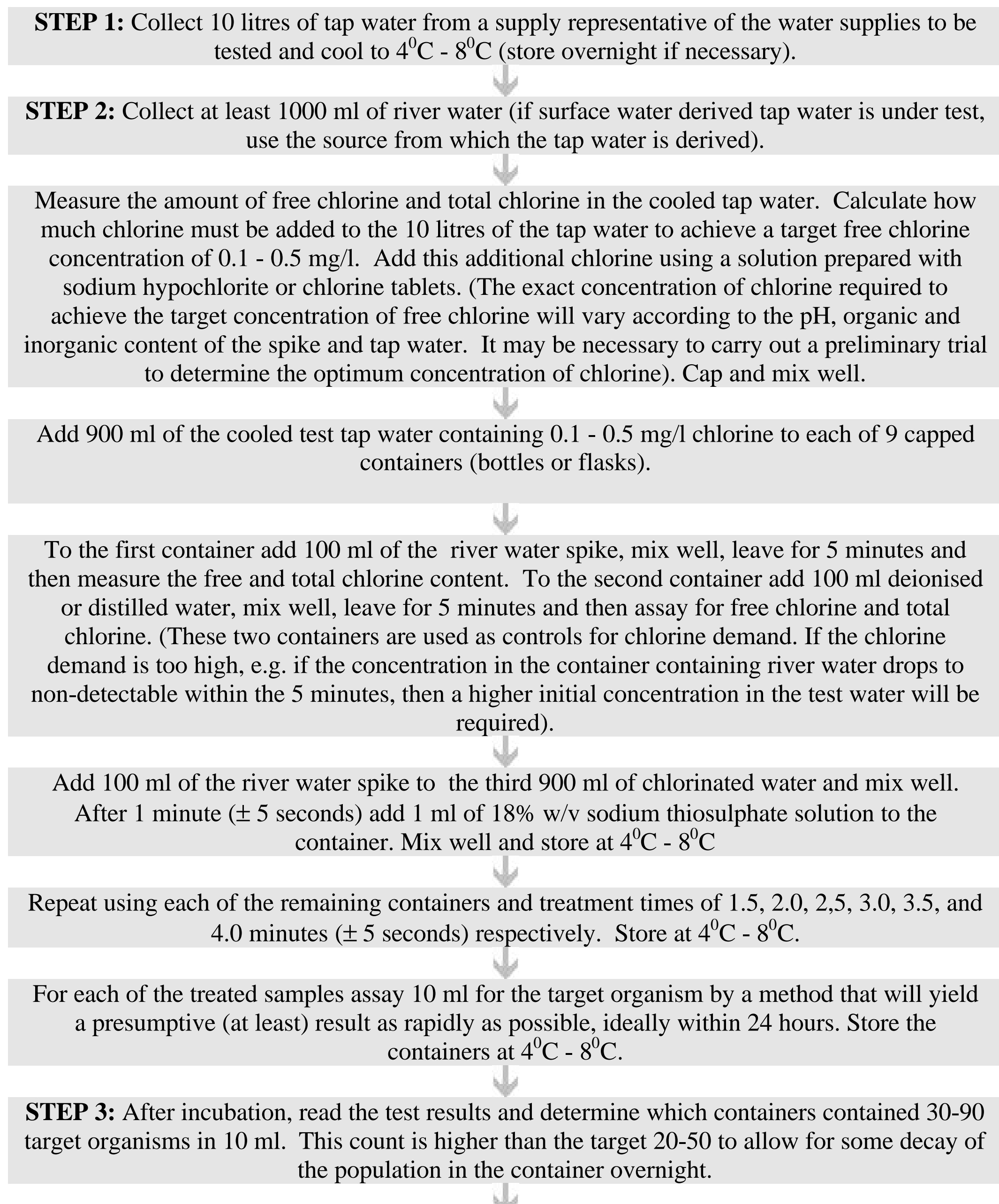
In certain situations, it may be necessary to compare unchlorinated but environmentally stressed organisms. In that case, sublethally injured organisms may be prepared by prolonged storage of sewage effluent or river water, instead of chlorinating these samples.

6.2 Preparation of spiked samples

Spiked samples should be prepared which contain chlorine-stressed target organisms, organisms closely related to target organisms; and non-target organisms.

Ideally, samples spiked with target organisms should contain 20-50 organisms per100 ml. A minimum of 100 target and 100 non-target organisms per method should be identified to confirm it is behaving as originally validated. However, identification of 100 non-target organisms is dependent upon the frequency of colonies growing on the media under test that do not produce colonies similar to the target organism(s), and 100 may not be achievable.

6.2.1 Protocol for generation of chlorine-stressed organisms using river water as the source of target organisms



Preparation of replicate test samples: For each container with water having counts of organisms within the 30-90 target range, add 900 ml amounts of fresh test tap water under test to separate clean 1000 ml containers. Add sufficient sodium thiosulphate to neutralise any residual chlorine and mix well.

For each of the selected assayed spiked samples add 100 ml to the 900 ml of dechlorinated tap water. Mix well.

Each litre sample now provides up to 10 replicate 100 ml drinking water samples suitable for testing two or more methods in parallel by one or more analysts.

The volumes of assayed spike samples and dechlorinated tap water used to prepare the water samples for testing can be increased proportionately to produce larger volumes for larger numbers of test replicates (e.g. addition of 300 ml of assayed spike sample to 2700 ml of dechlorinated tap water will generate sufficient volume to conduct 15 paired 100 ml samples for analysis by two methods).

6.2.2 Protocol for production of chlorine-stressed organisms from sewage effluent.

STEP 1: Collect 10 litres of tap water from a supply representative of the water supplies to be tested and cool to 4^0C - 8^0C (Store overnight if necessary).

↓

STEP 2: Collect at least 1 litre of sewage effluent and allow to stand for one hour at 4^0C - 8^0C to ensure solids have settled.

↓

Determine the level of residual chlorine, if any, in the cooled tap water.

↓

Prepare a chlorine solution of 12 - 25 mg/l by dissolving the appropriate amount of chlorine generating tablets or hypochlorite solution in 1 litre of distilled or deionised water. This will be used to add sufficient chlorine solution to the 9 litres of sewage effluent spiked test water to achieve a target chlorine concentration of 1.2 - 2.5 mg/l. (The exact concentration of chlorine required will vary according to the pH, organic and inorganic content of the spike and tap water. It may be necessary to carry out a preliminary trial to determine the optimum concentration of chlorine). Cap and mix well.

↓

Taking care not to disturb any settled solids transfer 500 ml of the sewage effluent into a clean separate 10 litre container (one fitted with a tap will make the following steps easier to carry out) containing a magnetic stirrer bar (or other stirring mechanism). Add 8.5 litres of the test tap water at 4^0C - 8^0C (from Step 1) resulting in 9 litres of sewage effluent spiked test water. Cap the container, mix the contents thoroughly by inverting several times, stand the container on a magnetic stirrer and stir vigorously.

↓

Start the clock and add the 1 litre of chlorine solution to the container making the final volume of 10 litres. The exact volume of chlorine solution may have to be adjusted to provide the target concentration of free chlorine. Mix the contents vigorously by rolling backwards and forwards for 1.5 -2 minutes. Place container on the stirrer and continue stirring vigorously. After 3 minutes (± 5 seconds), take a 500 ml sample into a capped vessel (bottle or flask) containing 1 ml of 18% w/v sodium thiosulphate. Mix well by inverting several times to ensure the chlorine is rapidly neutralised by the thiosulphate. Continue taking 500 ml volumes from the 10 litre container at one-minute intervals until 16 samples have been taken.

↓

Assay 10 ml volumes of each 500 ml chlorinated sewage spiked tap water samples for the target organism by a method that will yield a presumptive (at least) result as rapidly as possible, ideally within 24 hours. Store the sample containers overnight at 4^0C – 8^0C.

↓

STEP 3: After incubation of assays read test results and determine which containers have 30-90 target organisms in 10 ml. This count is higher than the target 20-50 to allow for some decay of the population in the container overnight.

↓

Preparation of replicate test samples: For each container with water having counts of organism within the 30-90 target range, add 900 ml of the fresh test tap water under test to a clean 1000 ml container. Add sufficient sodium thiosulphate to neutralise any chlorine and mix well.

↓

For each of the selected assayed spiked samples add 100 ml to 900 ml of dechlorinated tap water. Mix well.

Each litre sample now provides up to 10 replicate 100 ml simulated contaminated drinking water samples suitable for testing two or more methods in parallel by one or more analysts.

The volumes of assayed spike samples and dechlorinated tap water used to prepare the water samples for testing can be increased proportionately to produce larger volumes for larger numbers of test replicates (e.g. addition of 300 ml of assayed spike sample to 2700 ml of dechlorinated tap water will generate sufficient volume to conduct 15 paired 100 ml samples for analysis by two methods).

7. INTERPRETATION OF THE DATA (STAGE 1)

Pilot work with the preparation of samples is essential. It is necessary to ensure that as many as possible of the study samples give counts within the required range. Once the study has commenced all sample results must be recorded. If any result is higher than was intended (for example, too numerous to count, which is equivalent to 100 for membrane filtration or all tubes positive for MPN) data analysis may be different. If the paired results by both methods are too numerous to count then they can be omitted from the analysis because they contribute no information about whether the test method gave a higher or lower result than the reference method. However, if only one of the methods produces such results then the paired result must be analysed. The exclusion of such results could bias the conclusions. In some situations, it is possible to allocate an arbitrary high number to ">" results (e.g. 181 for >180). However, if there are many such results then it will force the researcher to use non-parametric statistical techniques because there will be no proper estimate of the difference in counts for these paired results. Clearly the study will be much easier to analyse and interpret if the results fall within the target range, and it is sensible to define this study range as well below the upper limit of the range of application of the method. It is impossible to prepare samples which never exceed a stated number – there are many factors which can affect the content of an aliquot – and so events which are higher than expected should still lie within the range of application.

Study samples which yield zero counts by either method must also be recorded and included in the analysis, with the exception of paired samples where count is zero by both methods. These can be excluded from the analysis because they contribute no information about whether the test method gave a higher or lower result than the reference method. Negative samples play an important role in the primary validation of a new method, to ensure that false–positive results are not a problem, but are not informative in this quantitative method comparison.

7.1 Statistical comparison of methods

When performing confirmation, representative colonies should be picked for confirmation in relation to existing identification rates. A maximum of 5 colonies per plate should serve the purpose, i.e. in the same microcosm 5 x 10 colonies. It may be necessary to confirm more colonies if confirmation rates are uncertain. A discussion of the statistics of confirmation is given in Section 6.3.3. in the Report on Public Health and Medical Subjects No. 71. HMSO. 1994 [7].

7.1.1 Stage 1 - spiked samples (20-50 target organisms/portion)

The protocols for preparation of spiked samples (section 6.2.1 and 6.2.2) are expected to give 20-50 target organisms per portion. The water may be stored and aliquots tested by both methods on different days, when the quantity of water to be used on any day shall be thoroughly mixed and then appropriate volumes drawn off for use in the two methods. These two results will be recorded together and analysed as a "paired sample" split between the trial and reference method, as shown in the worked examples (Annex A).

It is preferable that these samples come from a variety of categories of origin (see 2.1). Each set of samples belonging to a particular category of origin will have been suitably prepared, generally by one of the procedures described in Section 6.1. Each category of

origin will involve material from a particular source (e.g. a particular river or section of river, a treatment works etc.) although material can be collected over a period of time. These categories of origin are referred to as "sources" for convenience, although it should be noted that the samples are not directly from the particular source but they have involved combinations and manipulation of materials.

Suppose these sources number N (and sensible values of N might be 4 to 10).
The researcher must test enough samples from the N categories of origin to give statistical information that answers the questions:

Q1: Is the relative performance of the two methods similar for all N categories of origin?

Q2: If it is, then is the trial method finding significantly fewer organisms on average?

Q3: Even if it is finding a similar number or more, does the 95 % confidence interval for the estimated difference between average counts exclude the value where the trial method is finding ≥ 10 % fewer organisms than the reference method?

It is likely that these questions could be answered with at least 15 samples per source and a total of not less than 150. **However, unpredictability of organism numbers makes it difficult to predict the statistical information in a fixed size trial.** If the confidence interval for question 3 is still too wide, in other words if the comparison is inconclusive, then more samples should be examined as suggested in the flowchart below.
If stage 1 indicates that the trial method is acceptable so far then proceed to further stage(s) of evaluation.

EVALUATION OF DATA

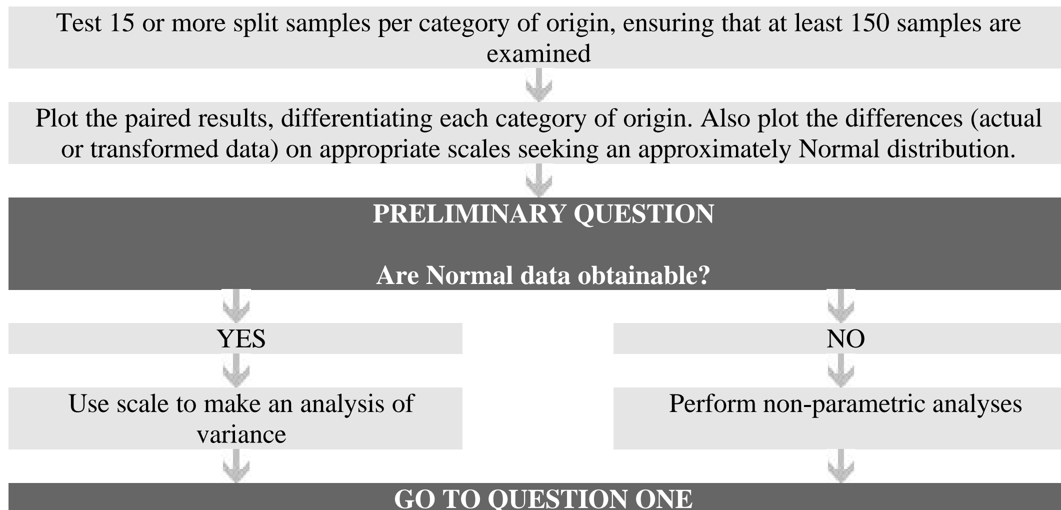


Figure 4: Stage one of the statistical comparison of methods – Part 1

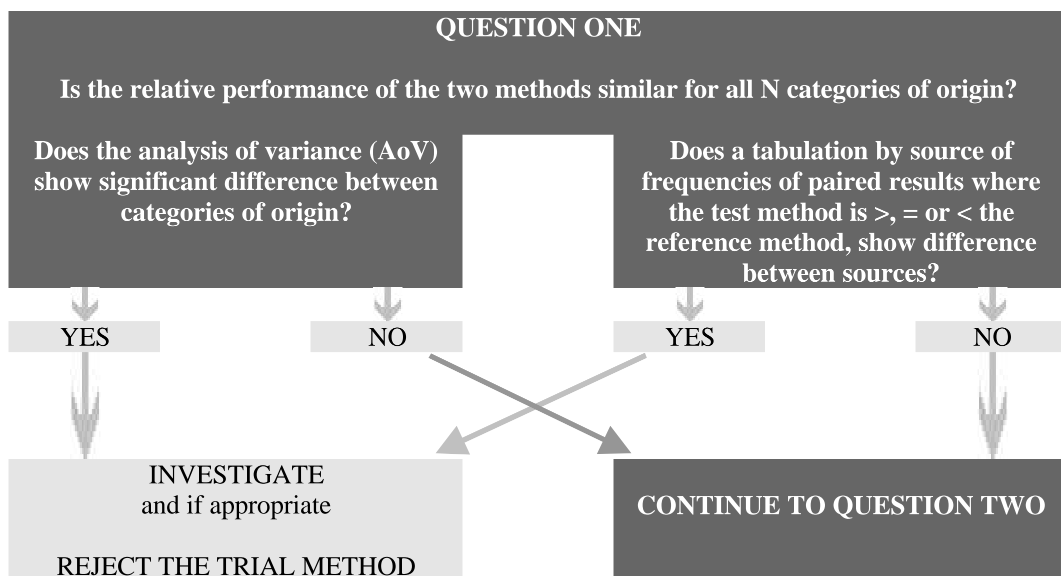


Figure 5: Stage one of the statistical comparison of methods – Part 2

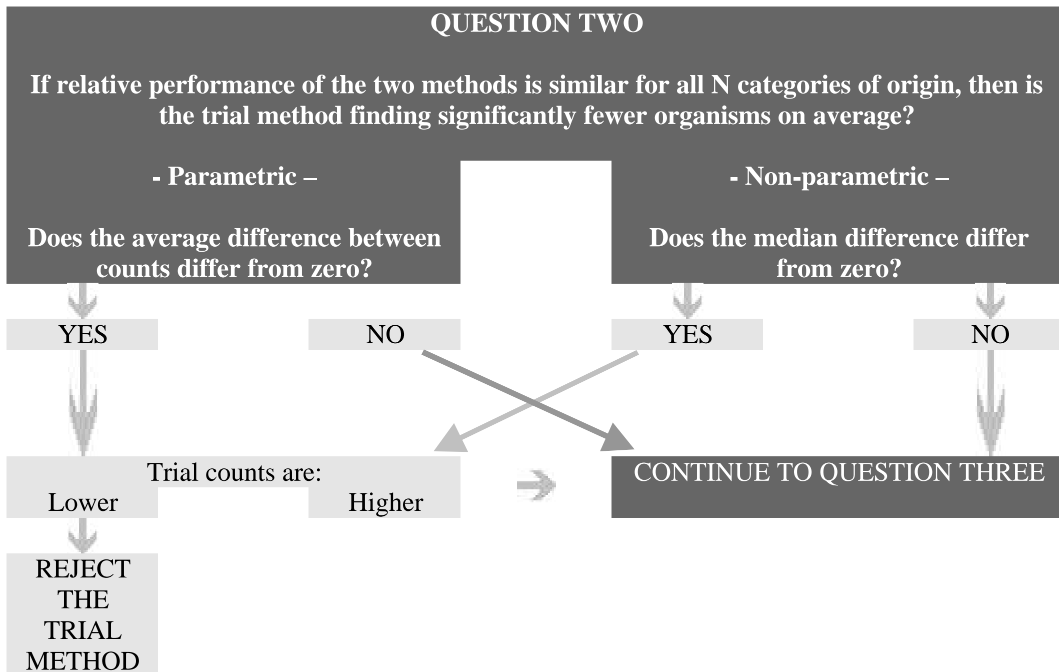


Figure 6: Stage one of the statistical comparison of methods – Part 3

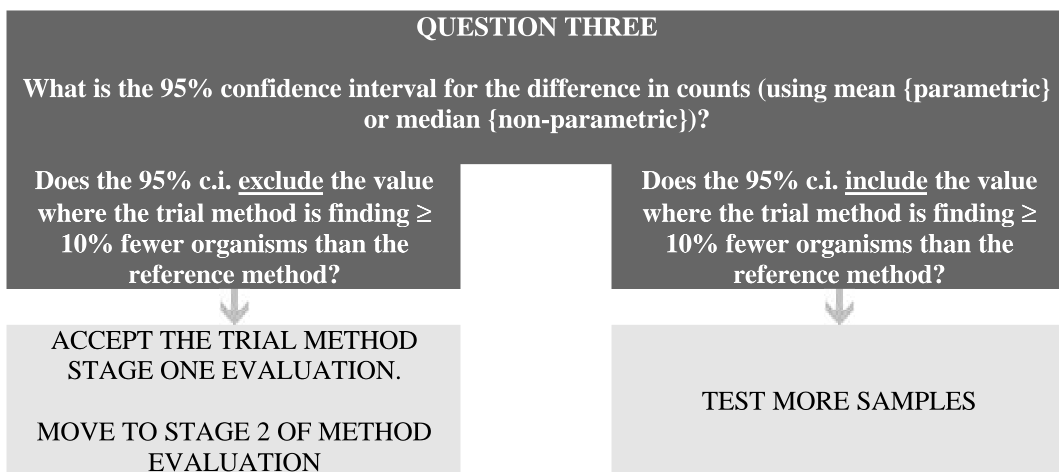


Figure 7: Stage one of the statistical comparison of methods – Part 4

8. STAGE 2 EVALUATION

Stage 2 evaluation is undertaken if satisfactory results are obtained from stage 1 and consists of paired analyses of spiked samples containing less than 20 target organisms per unit test volume. The process is repeated using spiked samples expected to give low numbers of relevant organisms per portion. This is to ensure that the comparison remains valid at lower levels, approaching those nearer to statutory testing but not so low as to result in the comparison being based upon presence/absence.

At least 30 paired sample results are required, giving counts in the 1 to 10 range, with one or more organisms detected by at least one method. The samples can be prepared in the same way as for Stage 1, but with extra dilution. Successive two-fold dilutions of the same sample are permitted, but the samples should not all come from one derivation category. Again, they should include an appropriate diversity of organisms.

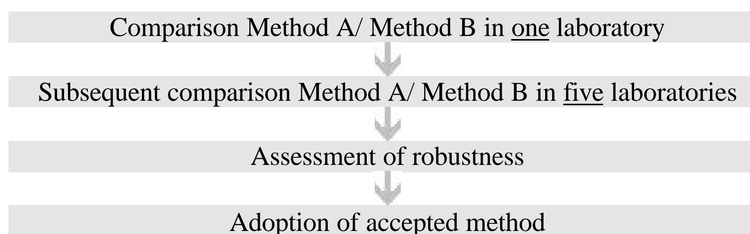
8.1 Data analysis for stage 2

All the results should be plotted. If appropriate, a regression model can be estimated to check that the slope and intercept do not show a significant inferiority of performance by the test method. With low counts it may be a problem to justify a parametric approach and it becomes more efficient to use a non-parametric analysis, as shown in the worked examples (Annex A). The proportion of paired results where the test count exceeds the reference count should not be significantly lower than a half. Thirty samples should give an estimate of the proportion, with confidence intervals which are no wider than plus and minus 20 %. If there is any trend towards a poorer relative result by the test method in Stage 2 as compared with Stage 1 then further samples may be required to demonstrate that the test method is not performing more poorly that the reference method with low counts. The aim should then to be to establish confidence intervals of ± 10 %, which might require about 100 samples.

9. EVALUATION OF A METHOD FOR NEW USE IN A LABORATORY

Any laboratory considering introduction of a method for new use should initially become familiar with the method e.g. by testing against pure cultures in suspension in water samples (n = 20–50 organisms per unit test volume) using target and non-target organisms.

9.1 Evolution of method introduction



A new method will need to undergo full comparative testing, using the procedure outlined in this document, in at least five laboratories before being regarded as potentially of general applicability. Where adequate comparative assessments in a single laboratory indicate that the new method is at least comparable to a reference method, then it could be adopted by that testing laboratory whether or not other laboratories have carried out similar comparisons. Once five laboratories have demonstrated equivalent or better performance of a new method compared to the reference method, wider adoption by other laboratories can be undertaken, once an assessment of robustness of the data from the five laboratories has been undertaken. This may result in fewer numbers of samples being needed for testing by the subsequent laboratories. The comparisons in the first five laboratories will require examination of at least 180 samples (150 Stage 1 + 30 Stage 2) in each laboratory. The samples are run in duplicate, and plates read blind in duplicate. Ideally all four ways of preparing samples listed in 6.1 should be used but samples should at least be representative of those in the laboratory area, i.e. those samples that are likely to be tested by the new method. Data from these comparisons should be submitted to the appropriate authority, so that the data can be pooled and reviewed following further statistical appraisal. By pooling the data, both positive and negative findings, it is possible to assess more confidently the robustness, repeatability and reproducibility of the method.

Once the robustness, repeatability and reproducibility have been established the method is generally accepted. Therefore the number of samples that **subsequent** laboratories are required to compare can be reviewed in the light of the expanding database. It is recommended, however, that a minimum of 30 spiked samples, using wild strains of organisms, are compared with the reference method.

10. COMPARING A MOST PROBABLE NUMBER (MPN) METHOD WITH AN ENUMERATION METHOD

The same two-staged plan should be used as outlined in the Sections 6 and 7. In situations where the MPN method is the test method being considered for adoption the aim will be to demonstrate that the MPN method is not finding significantly fewer of the relevant organisms than the enumeration method and, even when that is the case, that the average difference has been demonstrated accurately. The required level of accuracy is such that the 95 % confidence interval for the average difference in counts should not include the situation where the MPN method is finding 10 % fewer organisms. In situations where the MPN is the reference method then the investigation will be checking that the enumeration method does not give significantly lower counts and, if this is so, that the 95 % confidence interval for the average difference in counts should not include the situation where the enumeration method is finding 10 % fewer organisms.

The design of the study is exactly the same as for the comparison of two enumeration methods. Water samples should be prepared which involve organisms from a variety of origins, typical of the work of the laboratory. The same stages of the flow chart are used, the only difference being the factors influencing the choice of statistical methods. Paired analysis must be used but in this situation it will be comparing a count with a probable number.

The range of values achievable with an MPN is discontinuous within the range of the method. For example, in the 11 tube series (1x50 ml; 5x10 ml; 5x1 ml) if 9 tubes are positive (i.e. 1, 5, 3) the MPN is 91 per 100ml and if 10 tubes are positive (i.e. 1, 5, 4) the MPN is 160 per 100ml. It is impossible to get a result between 91 and 160. Probability studies have shown which counts or ranges of counts, would correspond to which MPNs for 11 tube (1x50 ml; 5x10 ml; 5x1 ml) and 15 tube (5x10 ml; 5x1 ml; 5x0.1 ml) series [8]. One approach could be to group the results from the counting method and compare them with the corresponding MPN and use non-parametric analyses. For example, with the 11 tube example given here it has been shown that true counts of 69 – 110 would most probably give a tube result of 1, 5, 3 and therefore an MPN of 91. True counts of 111 – 175 would most probably give a tube result of 1, 5, 4 and therefore an MPN of 160. So enumeration results of 69 – 110 could be analysed as “equal” to an MPN of 91 etc. Alternatively, the MPN can be regarded as the end result given to the customer and compared directly with the count from the paired result. Careful plots of the results must be made and consideration given to using non-parametric analyses along similar lines to those used in the worked examples with enumeration methods in Appendix 1b of Annex A.

It is important to be aware that the large gaps in the possible values of an MPN appear when most of the tubes are positive. The problems of the discontinuities can be reduced by using samples where a few, but less than half the tubes show reaction. Otherwise the MPNs will be approximate and the comparison with the enumeration method could become biased. Multiple tube methods which require large numbers of tubes at several dilutions will be more accurate than MPN methods with fewer dilutions and replicate tubes.

As with a comparison of two enumeration methods it is imperative that, once the study is underway, all results are collected for analysis and the only results which can be legitimately discarded are those where both methods find nothing or where both methods fail to give an estimate because the enumeration method showed overgrowth and the MPN method gave all tubes positive.

11 COMPARING TWO MPN METHODS

The same two-stage study plan should be used as outlined in the flow chart. The points raised in section 10 will apply to both methods but the principles of the comparison remain the same. Again the only difference may be in the choice of statistical methods, which should be decided after thorough scrutiny of data summaries and plots. It is likely that non-parametric methods will be needed. The preparation of samples should aim for numbers such that a few, but less than half the tubes show reaction in the reference MPN method.

12. REFERENCES

1. Technical Report ISO TR 13843. 1999. Validation of Microbiological Methods.

2. Public Health Laboratory Service Water Surveillance Group
   Preliminary study of microbiological parameters in eight inland recreational waters.
   Letters in Appl. Microbiology 1995 **21**, 267 – 271

3. Lightfoot NF, Tillett HE, Boyd P and Eaton S. Duplicate split samples for internal quality control in routine water microbiology. Lett. Appl. Microbiol. 1994; **19:** 321-324.

4. Tillett HE and Lightfoot NF. 1995. Quality control in environmental microbiology compared with chemistry; What is homogeneous and what is random? Wat Sci Tech 31: 471-77.

5. Lightfoot NF and Maier EA (eds). 1998. Microbiological Analysis of Food and Water. Elsevier Science, Amsterdam.

6. Mooijman KA., in't Veld PH., Koekstra JA., Heisterkamp SH., Havelaar AH., Notermans, SHW., Roberts D., Griepink B. and Maier E., 1992.
   Development of microbiological reference materials, Commission of the European Communities, Community Bureau of Reference, Report EUR 14375 EN, ISSN 1018-5593.

7. Report on Public Health and Medical Subjects No. 71. 1994. Methods for the Examination of Waters and Associated materials. The Microbiology of water.
   Part 1 – Drinking water. HMSO

8. Tillett HE and Coleman RE. 1985. Estimated numbers of bacteria in samples from non-homogeneous batches of water. J Appl Bacteriol 59: 381-88.

ANNEX A WORKED EXAMPLES

The following examples of comparisons between test methods and reference methods, which use colony counts from membranes, contain real data which have been selected to illustrate relevant situations. The data have been anonymised and their provenance adapted.

EXAMPLE A1. (Target organism = confirmed total coliform organism counts)
Stage 1:

A total of 150 samples were split between the test and the reference method. Confirmed counts per 100 ml were recorded. The sample material came from five geographically distinct backgrounds, all relevant to the routine work of the laboratory. The data are listed in Appendix 1a. The aim was to produce counts in the range 20 to 50, but in practice the range was wider. Median (and range) of counts were:

reference method: 30 (5 -180)
test method: 30 (3-170)

The data are plotted as counts in Figure 7 and on logarithmic scales in Figure 8. The paired differences were computed (test – reference) first using counts in Figure 9 then using log counts in Figure 10.

Comment
The test method appears to be comparable to the reference method in that there are as many, if not more, observations above the line of equivalence as below in Figure 7.
The paired differences appear clustered around zero but some differences are large in magnitude. This is not unusual in "real life" microbiology where occasional observations display variation much greater than random (Poisson) variation. The median difference (trial count minus reference count) was 1.0, range –61 to +90.

Preliminary Question: are the data Normal?
The distribution of the differences appears symmetric but with long tail ends in Figure 9. Working with logarithms of counts the tails are reduced, but are still affected by the "outlying" values in Figure 10. Normal scores for the differences, using both scales, were computed and plotted against the differences Figure 11. In this example the NSCORES function of the Minitab statistical software package was used. The plot will be linear for perfectly Normal data. The data from original counts are S-shaped in Figure 11 but are nearer to linear with logarithmic data although the affect of the "outliers" can still be seen in Figure 12.

A cautious response to question 1 would be to use non-parametric analysis because of the outliers, although parametric analysis of the differenced log data would be acceptable.

Parametric analysis

Question 1:
The differences were compared between the sample groups using one way analysis of variance. The output from Minitab is shown in Table 1. There is no significant difference between the groups.

Question 2:
The mean of the 150 paired differences (using log counts) is 0.021, standard error 0.019. Comparing this with the null hypothesis of zero average difference, $t_{149} = 1.10$, $p = 0.27$.

There is no significant difference, on average, between counts from the two methods.

Question 3:

The third question asks whether the lower end of the 95 % confidence interval (c.i.) for the average difference includes the situation where the test method is finding 10 % fewer organisms than the reference method.
In this example the 95 % c.i. for the average difference is $0.021 \pm t_{149, 0.05}\ 0.019 = -0.016$ to 0.058

The question is saying that the unacceptable situation occurs when, on average, test count (y) is 90 % of reference count (x). i.e.

$$y = 0.9\,x$$
on the logarithm scale:
$$\log(y) = \log(0.9) + \log(x)$$
thus the difference in counts becomes:
$$\log(y) - \log(x) = \log(0.9) + \log(x) - \log(x)$$
$$= \log(0.9) = -0.046$$

In this example the confidence interval for the average difference lies above this value of –0.046 (being –0.016-0.058).

CONCLUSION - accept the trial method for further consideration in Stage 2.

Non-parametric analysis

Question 1:
The differences were compared between the sample groups using ranks, as shown in Table 2. The probabilities test the null hypothesis that any discrepant paired result, where the counts were not identical, was as likely to give a higher count by the test as by the reference method. They were calculated using exact binomial probabilities, parameter $p = 0.5$, two-tailed. Alternatively McNemar's test can be used, but this becomes approximate when the number of discrepant pairs is less than about 15.
There was no significant difference between methods in any of the groups of samples and no indication of disagreement in findings between groups.

Questions 2 and 3:
The median of the 150 paired differences was comparing this with the null hypothesis of zero average difference and the 95 % c.i. estimated. The Mann-Whitney test in Minitab was used. The median difference (test count minus reference count) was:
1 with 95 % c.i. – 4 to + 6. The median count by the reference method was 30. The lower point of the c.i. would represent an average 13 % deficit in counts by the test method (i.e. {(30-4)/30} x 100).

CONCLUSION – a small amount of extra data should be sought and included in a repeat analysis to ensure than the lower confidence interval is reduced to – 3, or whatever become equivalent to 10 % below the overall reference median count.

Comment
The cautious approach, which is a perfectly correct one, leads to the conclusion that more data need to be collected for Stage 1 of the trial. If the parametric approach is used then the confidence interval is narrower and it is acceptable to go on to Stage 2. It should be noted that both analyses have reached the same conclusion that the test method is giving a very slightly higher average count but that it is not significantly better or worse than the reference method.

Stage 2:
Six more samples from each group were obtained using the same five geographically distinct sources of material. These 30 paired results are listed in Appendix 1b and plotted in Figure 13. There is a scatter of points above and below the line of equivalence for all groups of data. These data appear to agree with the findings from Stage 1. Analysis confirms that there is no significant difference between method:

Parametric:

Counts by test method:	mean = 4.2	median = 3	range 0 to 14
Counts by reference method:	mean = 3.8	median = 2.5	range 0 to 14

Difference = (test count – reference count).
Mean difference = 0.37, standard error 0.67. Comparing this with the null hypothesis of zero average difference, $t_{29} = 0.55$, $p = 0.6$.
The test method appears comparable to the reference method at Stage 2.
The 95 % c.i. for the difference in counts is $0.37 \pm t_{29, 0.05}\ 0.67 = -1.0$ to $+1.8$
N.B. the lower limit would imply a test mean count of (4.2 − 1.03) which is 83 % of the reference method mean. This is above the target of demonstrating, at Stage 2, that the test is not 20 % worse than the reference method.

Non-parametric:
Test count compared with reference count was:

<	=	>	total	
11	5	14	30	(p=0.5)

CONCLUSION - accept the test method as being comparable to the reference method in this laboratory.

EXAMPLE A2. (target organism = confirmed *E. coli* counts)

Stage 1:
These data come from another trial. The results presented here represent a smaller study but the evidence already throws the suitability of the test method into doubt and illustrates how a large study is not needed to demonstrate that a method is probably unsatisfactory. The data are listed in Appendix 1c.

Forty five samples were split between the trial and the reference method. Confirmed *E. coli* counts per 100 ml were recorded. The sample material came from three geographically distinct backgrounds, all relevant to the routine work of the laboratory. The data are listed in Appendix 1c. The aim was to produce counts in the range 20 to 50, but in practice the range was wider. The median (and range) of counts were:

reference method: 30 (10 - 84)
test method: 28 (3 - 78)

The data are plotted as counts in Figure 14. The paired differences were computed using log counts as in the previous example. Both the non-parametric approach (Table 3) and the parametric analysis of variance using differenced log counts (Table 4) show that the second group of data displayed significantly worse results by the trial as compared with the reference method. Of the 15 paired results, 11 were higher by the reference method, two the same and two lower (p = 0.02). The 95 % c.i. for the mean of the test minus reference log counts lies entirely below zero.

Comment
The results from the other two data groups show no significant difference between methods. But the poor performance of the test method with group 2 data suggest that it is not worth investigating further, unless there is a very good reason for this poor performance which can be rectified.

CONCLUSION - reject the test method as not suitable for use in this laboratory.

Appendix 1a

Data for Example A1

Stage 1 – Data by Source

Source 1 reference	Source 1 test	Source 2 reference	Source 2 test	Source 3 reference	Source 3 test	Source 4 reference	Source 4 test	Source 5 reference	Source 5 test
96	106	13	7	35	41	8	10	25	25
22	24	15	21	27	24	11	3	60	52
22	30	21	28	20	23	76	70	19	20
40	40	23	34	40	108	15	17	11	12
40	20	14	35	20	13	19	33	10	8
15	23	5	29	40	16	27	28	9	10
6	11	8	74	70	84	21	21	7	11
80	80	16	12	37	81	21	30	19	26
10	15	84	52	50	15	50	81	58	75
60	40	71	10	80	32	50	60	43	39
63	43	9	69	57	95	34	19	142	140
50	50	80	78	80	60	30	25	8	10
80	170	60	60	14	10	41	47	20	30
14	14	33	16	40	15	22	24	180	170
40	38	24	14	60	126	100	98	10	30
60	40	*	*	*	*	17	13	10	20
21	32	*	*	*	*	75	77	13	12
31	45	*	*	*	*	11	7	32	24
86	72	*	*	*	*	14	17	10	10
14	23	*	*	*	*	33	33	70	76
*	*	*	*	*	*	41	37	15	17
*	*	*	*	*	*	6	20	27	28
*	*	*	*	*	*	30	17	21	21
*	*	*	*	*	*	27	23	21	30
*	*	*	*	*	*	26	18	50	81
*	*	*	*	*	*	15	8	34	19
*	*	*	*	*	*	18	15	25	30
*	*	*	*	*	*	20	28	24	22
*	*	*	*	*	*	58	30	100	98
*	*	*	*	*	*	17	21	13	17
*	*	*	*	*	*	58	56	75	77
*	*	*	*	*	*	21	19	14	17
*	*	*	*	*	*	30	38	41	37
*	*	*	*	*	*	64	64	27	23
*	*	*	*	*	*	41	44	39	50
*	*	*	*	*	*	25	25	17	21
*	*	*	*	*	*	17	7	58	56
*	*	*	*	*	*	52	36	21	16
*	*	*	*	*	*	67	63	30	38
*	*	*	*	*	*	12	32	52	36
*	*	*	*	*	*	44	34	63	67
*	*	*	*	*	*	32	19	34	44
*	*	*	*	*	*	64	68	25	25
*	*	*	*	*	*	39	36	64	68
*	*	*	*	*	*	57	32	39	33
*	*	*	*	*	*	39	42	39	42
*	*	*	*	*	*	40	51	40	51
*	*	*	*	*	*	21	29	12	16
*	*	*	*	*	*	24	35	41	80
*	*	*	*	*	*	41	80	32	27

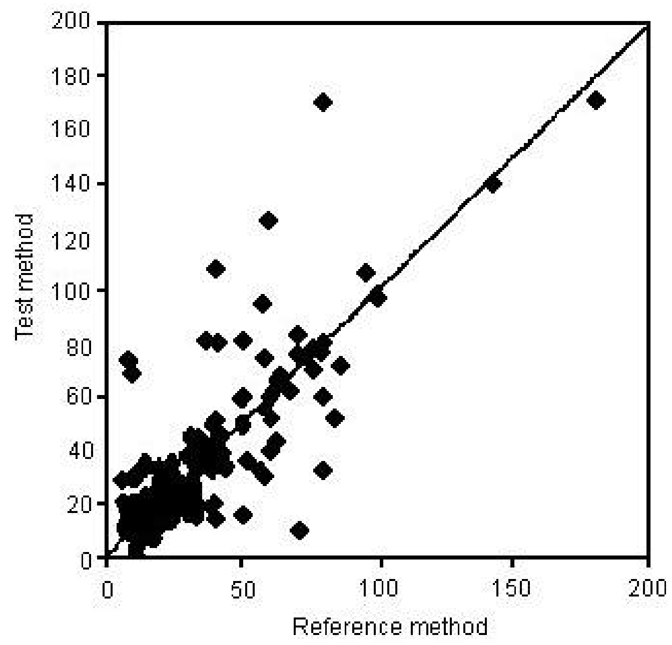


Figure 7: 150 Paired counts

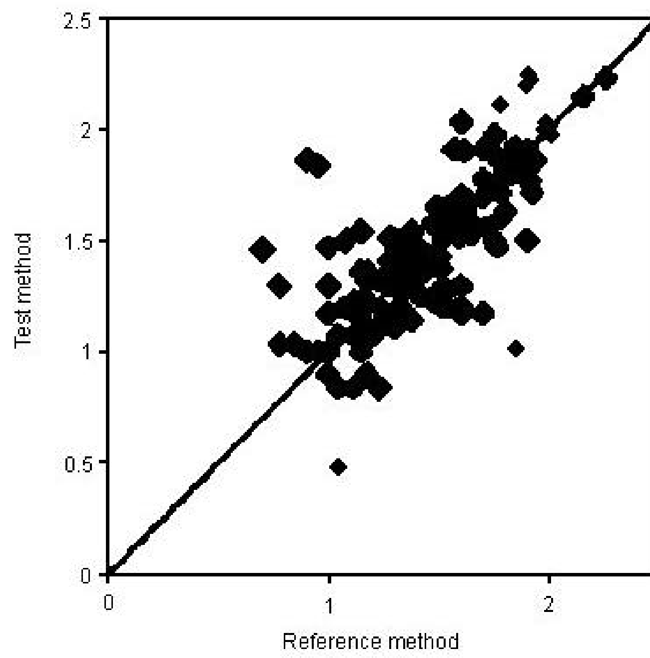


Figure 8: Log scale

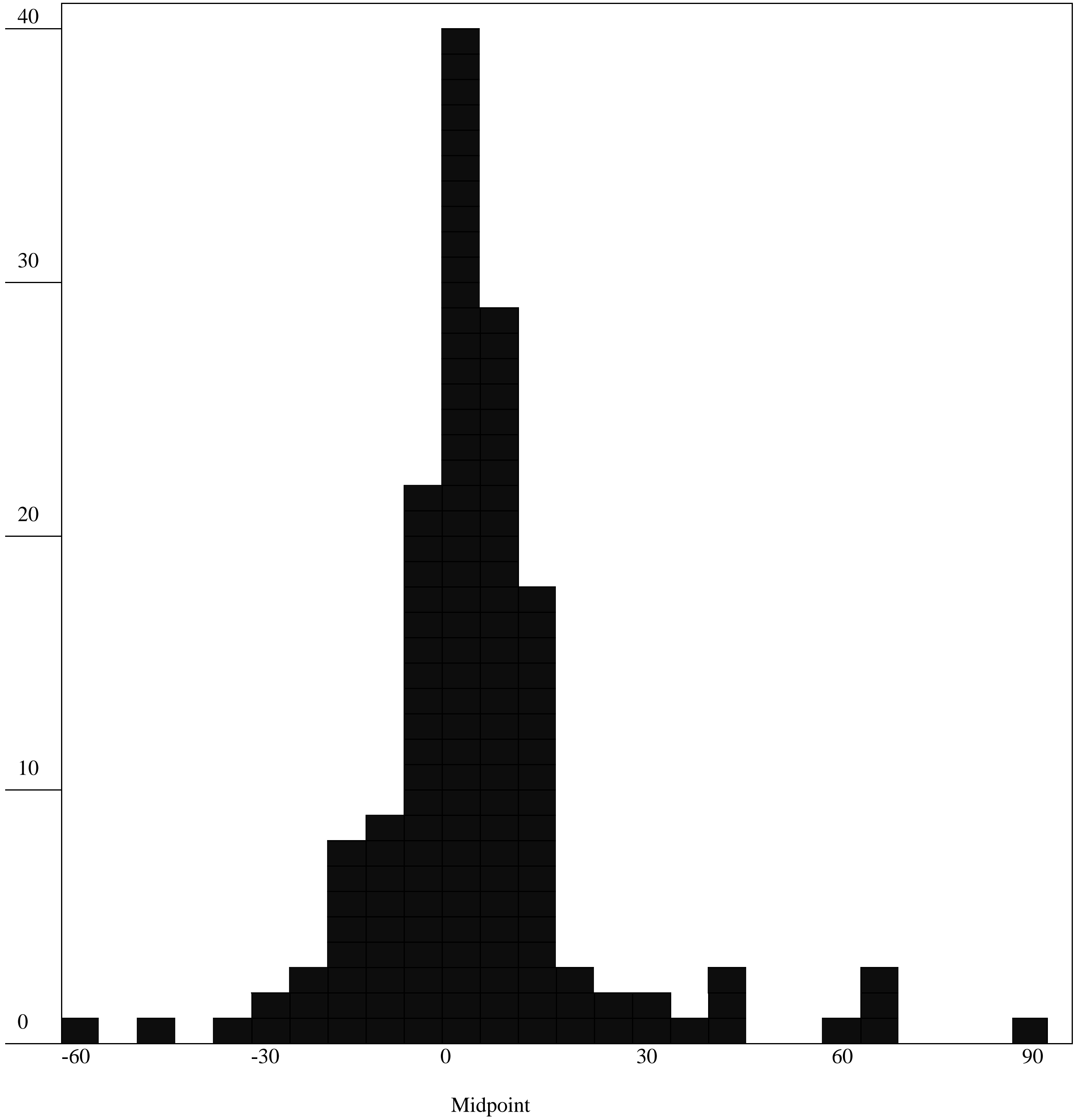


Figure 9: Histogram of Differences

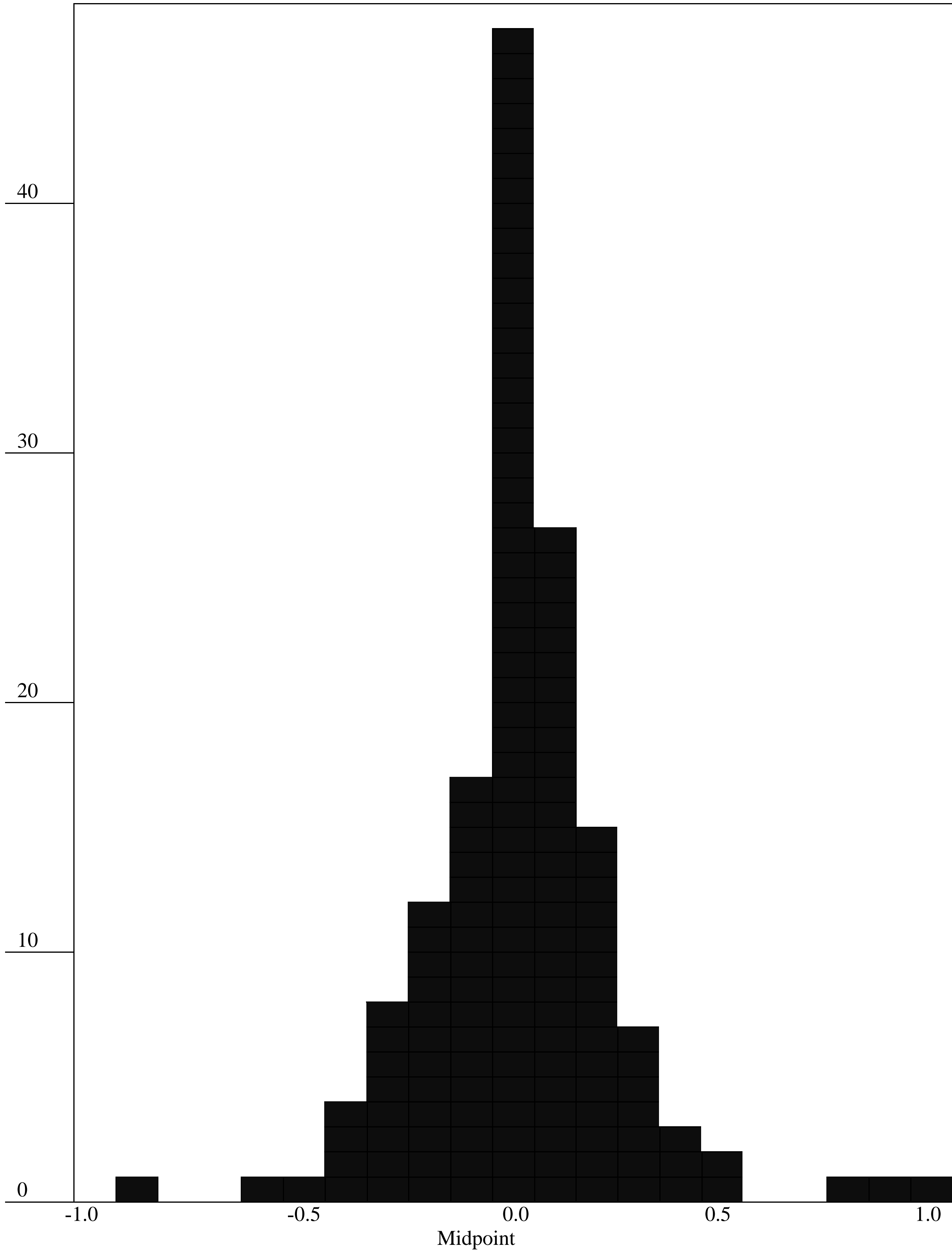


Figure 10: Histogram of log difference

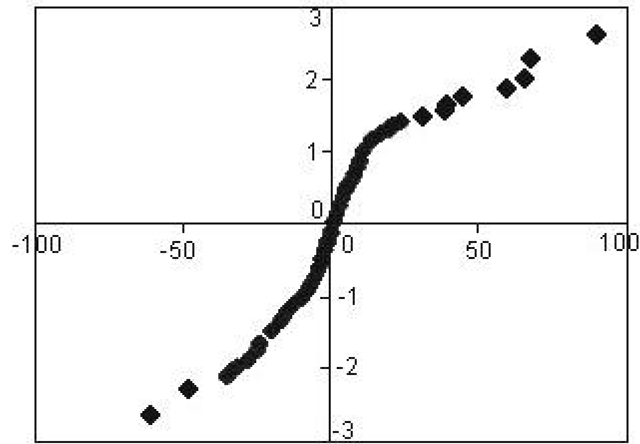


Figure 11: Normal plot for differenced counts

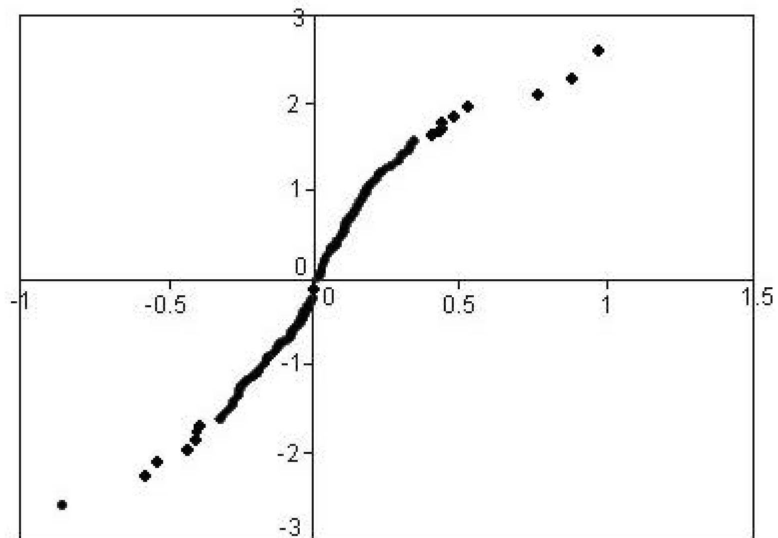


Figure 12: Normal plot differenced log counts

Appendix 1b
Example A1
Data for Stage 2

Group	Reference	Test
1	4	0
1	2	1
1	8	7
1	1	4
1	0	2
1	2	4
2	0	3
2	4	1
2	0	6
2	14	10
2	13	7
2	4	7
3	5	9
3	3	3
3	1	0
3	1	1
3	3	2
3	7	14
4	9	6
4	1	0
4	6	11
4	2	2
4	0	1
4	6	7
5	1	1
5	3	3
5	13	3
5	0	2
5	2	3
5	0	6

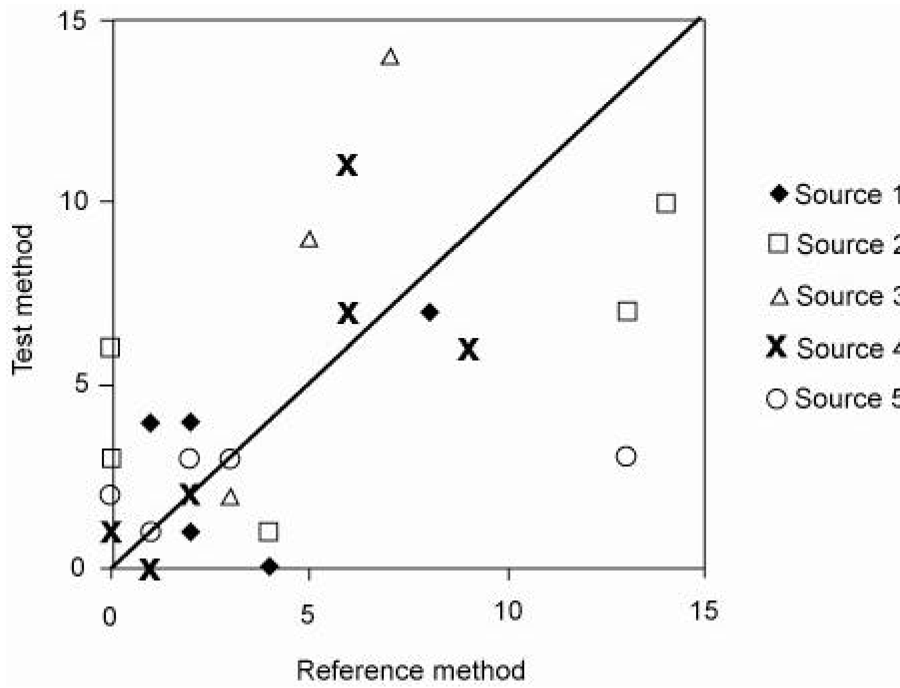


Figure 13: Stage 2 samples

Appendix 1c
Data for Example A2
Stage 1 Data by Source

source	reference	test
2	76	51
2	27	18
2	21	12
2	18	18
2	33	37
2	30	11
2	30	10
2	22	3
2	15	10
2	32	15
2	20	23
2	23	10
2	20	12
2	32	18
2	24	24
3	35	15
3	23	40
3	40	40
3	40	16
3	30	44
3	37	18
3	50	18
3	57	34
3	14	28
3	40	51
3	44	60
3	15	15
3	33	60
3	26	29
3	30	40
1	32	24
1	18	21
1	21	28
1	28	27
1	10	32
1	19	60
1	16	9
1	84	52
1	71	30
1	80	78
1	60	60
1	33	46
1	24	44
1	55	50
1	20	30

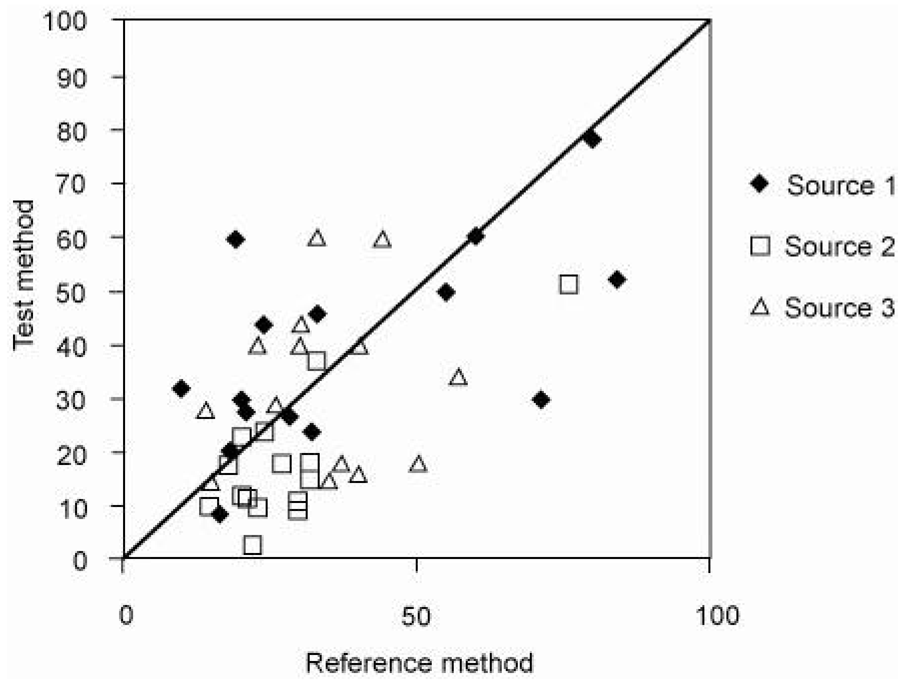


Figure 14: Exercise 2 – 45 paired counts

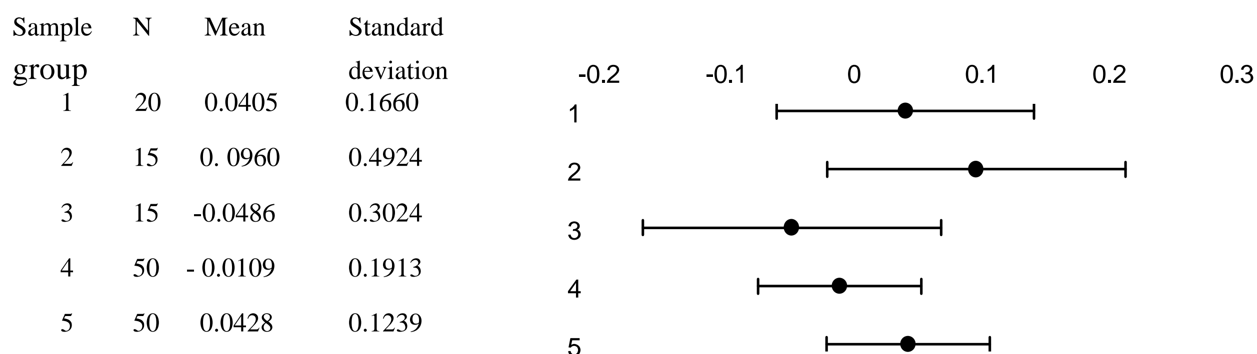
Table 1: Example A1, Stage 1, parametric analysis.

Analysis of Variance of differences of log counts

	DF	SS	MS	F	p
Sample group	4	0.2393	0.0598	1.12	0.349
Error	145	7.7427	0.0534		
Total	149	7.9821			

Individual 95% Confidence levels for
Mean based on Pooled standard deviation

Sample group	N	Mean	Standard deviation
1	20	0.0405	0.1660
2	15	0. 0960	0.4924
3	15	-0.0486	0.3024
4	50	- 0.0109	0.1913
5	50	0.0428	0.1239



Pooled standard deviation= 0.2311

Table 2 Example A1, Stage 1, paired comparison, non-parametric analysis.

Test method compared with reference method count

Sample group	<	=	>	ALL	p*
1	6	4	10	20	0.4
2	7	1	7	15	0.6
3	8	0	7	15	0.5
4	23	4	23	50	0.6
5	17	4	29	50	0.052
ALL	61	13	76	150	

* two-tailed, binomial probability

Table3. Example A2, non-parametric analysis

Test method compared with reference method

Sample group	<	=	>	ALL	p*
1	7	1	7	15	0.5
2	11	2	2	15	0.02
3	5	2	8	15	0.3
ALL	23	5	17	45	

* two-tailed, binomial probability

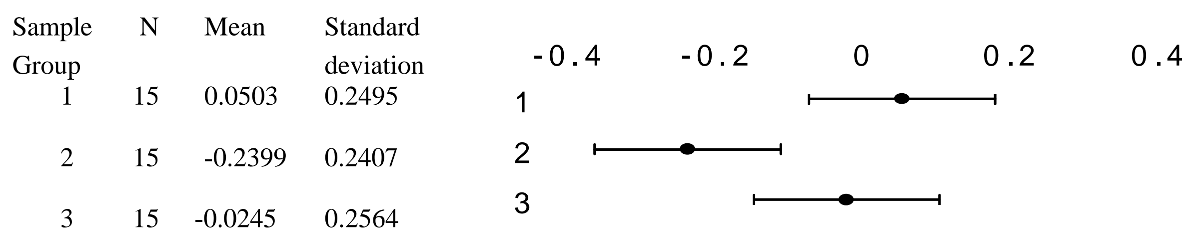
Table 4: Example A2, parametric analysis.

Analysis of Variance of differences between log counts

	DF	SS	MS	F	p
Sample group	2	0.6812	0.3406	5.50	0.008
Error	42	2.6027	0.0620		
Total	44	3.2839			

Individual 95% Confidence levels for
Mean based on Pooled standard deviation

Sample Group	N	Mean	Standard deviation
1	15	0.0503	0.2495
2	15	-0.2399	0.2407
3	15	-0.0245	0.2564



Pooled standard deviation= 0.2489