Acceptance testing procedure equivalence

B. Abbasi[1] R. Crawford[2] K. A. Haskard[3]
A. Olenko[4]

(Received 24 September 2011; revised 15 May 2012)

Abstract

This article addresses the issues of comparing different acceptance testing systems in an industrial setting, specifically in the dairy industry. The issues were two-fold: how to demonstrate that two different product testing systems were equivalent; and how to ensure that testing done by a customer or consumer on delivery of the product does not reject product deemed acceptable by the producer's testing system. Our comparison of sampling systems was focused around Operating Characteristic curves. Our results suggest that previous approaches are sound when data are normally distributed, although some refinement is possible. When data are not distributed normally, especially with multi-parameter distributions, the usual one dimensional Operating Characteristic curve method fails. In such cases, test methods can be compared by comparing acceptance surfaces in three dimensional plots. To address discrepancies between producer and consumer testing systems, especially if these arise because of different levels of variability

http://journal.austms.org.au/ojs/index.php/ANZIAMJ/article/view/4707 gives this article, © Austral. Mathematical Soc. 2012. Published June 13, 2012. ISSN 1446-8735. (Print two pages per sheet of paper.) Copies of this article must not be made otherwise available on the internet; instead link directly to this URL for this article.

between the two systems, an approach involving confidence intervals has the most appeal.

Contents

1 Introduction

This is the report of a 2011 Mathematics and Statistics in Industry Study Group (MISG 2011) working group. Fonterra, a leading multinational dairy company, asked the MISG 2011 participants to examine the following problem: how do we show that two sampling schemes—Scheme A (using the traditional end of run sampling point and test method A) and Scheme B (using an alternative sampling point, a different test method and possibly a different number of samples per production run)—give the same, or more generally equivalent confidence that the production lot meets specification.

In manufacturing or processing, many physical tests of the product are done to assess physical attributes or quality characteristics, and to ensure compliance with standards. Different tests or test procedures may be available at different stages of the process (including end-user or buyer testing at the delivery point). Test methods may vary in their precision, accuracy, cost, complexity, and the specific attribute(s) measured. Test results from several items or samples from a batch can be combined to characterise the batch as acceptable or otherwise. This is known as *acceptance sampling*.

The problem for the MISG was to investigate formal methods to compare different sampling systems, and how to declare them equivalent. Can a standard reference test method be replaced by a quicker, cheaper, easier alternative test and give equivalent results? Introduction of more highly technological equipment for testing, such as near-infrared spectrometry for example, may enable more accurate tests (through more accurate measurement and/or larger sample sizes). Does a different sampling scheme or acceptance criterion, or some combination of these, give an equivalent decision rule?

A further problem was to understand and investigate how to deal with or reduce differences between sampling systems, especially the occurrence of discrepancies between tests by the company before dispatch, revealing acceptable product, and tests on receipt by the buyer which sometimes reject a delivery. These problems can cost millions of dollars to resolve.

It quickly became clear that this seemingly simple problem is far from simple. It is not a case of a simple statistical test to compare two means or a problem in the classical theory of bioequivalence testing [11, 14]. First is the issue of an operational definition of equivalence. We might be interested in equivalence of individual tests (on individual items or samples of product), or equivalence of two sampling systems (on batches of product). How should we measure or describe such differences, and what size of difference becomes of practical importance? Variability is present in several different aspects: the product itself, measurement error of testing processes, and sampling error. There can be different accuracy of tests (laboratory accuracy) and different accuracy due to sample size, bias and precision differences, and lack of control of a customer's test. Because of sampling variation, there is always a nonzero probability of accepting a substandard batch (if present), and of rejecting a good batch; and even if two test methods are identical, with or without measurement error, they will not always accept or reject the same individual batches. Remember that "the main purpose of acceptance sampling is to decide whether or not the batch is likely to be acceptable, not to estimate the quality of the batch" [12].

The following section provides definitions and terminology, and describes and discusses the Operating Characteristic (OC) curve, which is a central feature of acceptance testing. In Section 3 several issues are discussed and approaches considered, including the meaning of test or sampling system equivalence. This section provides a mathematical framework and discusses variability and distributional assumptions and their consequences. Section 3 also summarises results about the OC curve for inspection by variables under a range of situations including measurement error, one-parameter and multi-parameter underlying distributions and the sensitivity of the OC curve to the distribution, and inter-rater agreement. The section ends with issues relevant to the consumer rejecting batches that have been accepted by the producer. The article concludes with a summary of the findings.

2 Definitions and terminology

We begin with some definitions. A *test method* is a well defined laboratory procedure producing a measured variable such as percentage moisture or fat content, or a binary variable such as acceptable or conforming to a standard, or non-conforming. When examining a batch of product, sampling is necessary. A *batch* is a collection of items or product produced at one time under the same conditions, perhaps all the product to be delivered to a given customer at a given time. A *sampling plan* specifies the number of items to be tested, and a rule for determining whether a batch is acceptable.

Two common forms of the test method are

1. for "*inspection by variables*" (such as fat or moisture content), the batch will be accepted provided the sample mean plus some specified multiple of the standard deviation is within some acceptable range, delimited by a specified critical value (the sample size, the multiple of the standard deviation and the critical value must be specified),
2. for *"inspection by attributes"* (conforming versus non-conforming), if more than a specified number of sampled items are non-conforming, the batch is rejected (sample size and the maximum allowable number non-conforming must be specified).

A complete *sampling system* consists of a laboratory test method together with a sampling plan.

Good summaries and pertinent comments on aspects of acceptance sampling were given by Grzegorzewski [5], Hald [7], and the electronic publication of the US National Institute of Standards and Technology [12]. Various specific requirements in the dairy industry are given by the ISO [8, 9].

2.1 Operating characteristic curves

Because practicalities necessitate that quality testing be done by sampling, it is impossible to be absolutely confident that a batch is completely acceptable—we can at best make probability statements about the quality of batches.

The operating characteristic or OC curve is a graph of the probability of accepting a batch against the proportion of the batch that is non-conforming or unacceptable. Therefore the OC curve is a plot of $\Pr(N)$ versus N/n (or sometimes just N for fixed n), where $\Pr(N)$ is the probability of accepting a batch that contains N non-conforming items when the total number of items is n.

The OC curve is obtained by applying acceptance criteria to samples with N non-conforming items in n samples. The sample curve can be determined in a theoretical manner or, if it is difficult, numerically by repeating sampling many times.

When small fractions are non-conforming, there is a large probability of accepting the batch, and the curve drops, typically in a reverse sigmoidal shape. Small probabilities of accepting batches are associated with a large proportion of non-conforming samples. Figure 1 illustrates a typical OC curve with some variations. Two important quantities are:

- acceptable quality level (AQL), the largest tolerable proportion non-conforming in a batch, the producer's baseline requirement for quality—in practice the batch will be deemed acceptable with some (specified, large) probability $1-\alpha$; and
- limiting quality (LQ), the largest proportion non-conforming in the batch that the consumer would tolerate, for which we wish to be confident of

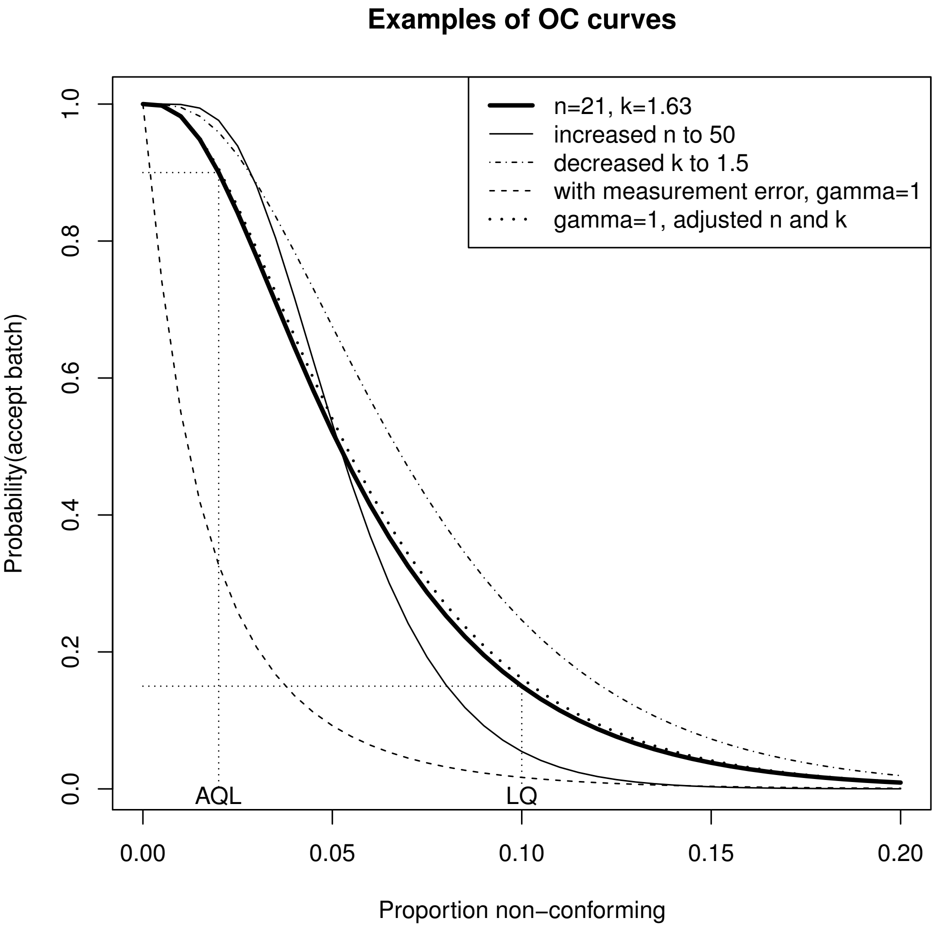

Figure 1: Example OC curves for inspection by variables.

rejecting the batch with some (specified, large) probability $1 - \beta$.

The AQL is smaller than the LQ, and these two values, together with the probabilities, place limits on the OC curve. Ideally, the OC curve drops off rapidly beyond the AQL, to minimise the chance of accepting a batch with proportion non-conforming greater than AQL.

The probability of rejecting a batch which is at AQL, denoted α as for Type I error rates in statistical hypothesis testing (wrongly rejecting the null hypothesis when it is true [1]), is called the *producer's risk* because it is the risk of rejecting a batch that is actually acceptable, and so is a cost to the producer. The probability of accepting a batch which is at LQ, denoted β as for Type II error rates in statistical hypothesis testing (failing to rejecting the null hypothesis when it is in fact false [1]), is called the *consumer's risk* because it is the risk that the consumer will be sent an unacceptable batch, where the consumer the limit of acceptability is assumed to be LQ.

For example, it might be required that at least 98% of a batch of butter has moisture content less than 16%. We desire a sampling system that has an OC curve with a high probability of accepting such a batch but a low probability of accepting a batch in which less than 98% has acceptable moisture content. There will be uncertainty, and safety margins must be built in. Further, we desire confidence that buyers or consumers, with their own sampling systems, will not reject batches that were deemed acceptable for delivery.

One form of the OC curve depends on a test statistic $\overline{X} + kS$ where $\overline{X}$ is the mean of a sample of items from the batch, S is the sample standard deviation, and the constant k and sample size n are part of the specification of the sampling system. This is described in more detail in Section 3.4.

Changes to the specification of the sampling system change the OC curve. Some examples are shown in Figure 1. Increasing the sample size n reduces the variance of the test statistic and makes the OC curve steeper—closer to an ideal curve, in which a batch is highly likely to be accepted if the proportion non-conforming is close to the AQL, but the probability drops away very quickly as the proportion non-conforming increases. If process variation within a batch is reduced, the effect will be similar.

Changing the multiplier k shifts the curve sideways. As shown in Figure 1, decreasing k moves the OC curve to the right, so that batches are more likely to be accepted. Conversely, increasing k makes it easier for batches to be rejected—the test statistic $\overline{X} + kS$ is more likely to extend beyond the specification limit L; the OC curve is shifted to the left.

Adding measurement error moves the OC curve to the left (see Figure 1), so for the same underlying proportion non-conforming, there will be a smaller chance of accepting the batch. This is of particular concern for the producer, if the consumer's test method has larger measurement error. Figure 2 shows examples of probability density functions for the underlying variable X, its sample mean based on a sample of n observations, and the distribution of the test statistic $\overline{X} + kS$, where k and n were chosen to give the bold OC curve shown in Figure 1. In Figure 2 the specification limit $L = 16\%$ for moisture in butter is indicated. These plots illustrate:

- the proportion of the individual items that will be non-conforming—this is the proportion of the fine solid curve that is beyond the limit $L = 16\%$;
- the distribution of the sample mean for a given sample size n (dashed curve);
- the distribution of the test statistic $\overline{X} + kS$—the dashed curve shifted to the right but also with increased variance because of the variability in S, showing the probability that the batch will be rejected, namely the proportion of the area under the bold solid curve that is beyond the limit $L = 16\%$.

For each n the test parameters k is chosen in such a way that, for a given proportion of non-conforming individual items, the probability of acceptance of the batch is the same (regardless of the scaling of the underlying distribution). For example, if we use the underlying normal distribution (Figure 1) then for

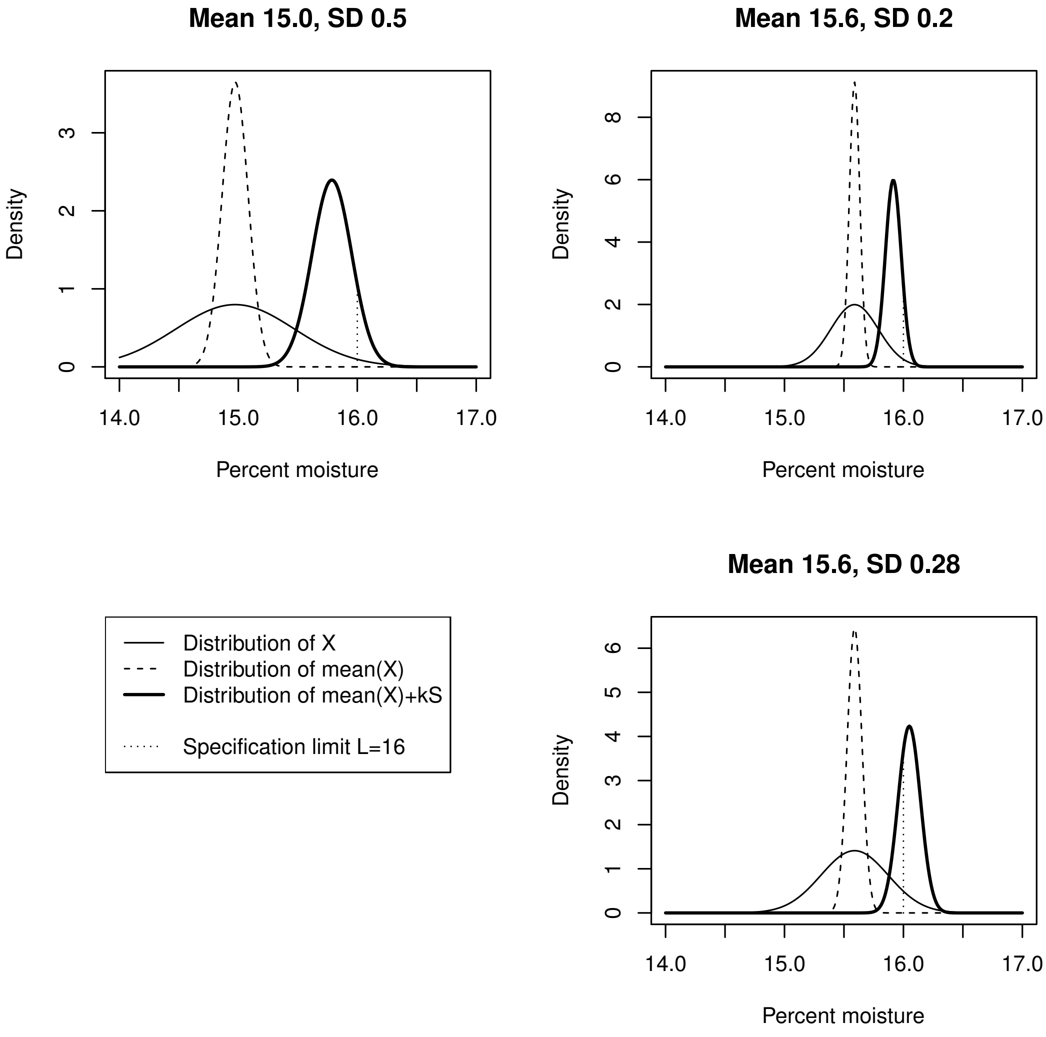

Figure 2: Density plots relevant to inspection by variables.

$n = 21$ we have $k = 1.63$.

The first two plots in Figure 2 are examples where the batch is at the AQL—the proportion of samples beyond the specification limit of 16% moisture content is $\alpha = 0.10$, and in both cases there is a 0.15 chance that the test statistic will be greater than the specification limit L. These two plots show that, if the variation between items within a batch is smaller, the whole batch can have values much closer to the critical limit (here $L = 16\%$ moisture) without increasing the risk of the batch being rejected. When variance is larger, an additional safety margin must be built in.

The top two plots in Figure 2 illustrate just two of the many ways in which batches could arise with the same proportion non-conforming. In each case, the probability of acceptance is the same—it depends only on the proportion non-conforming. The two parameters of the underlying distributions, assumed normal, are transformed into a single measure, namely the proportion of the batch that is non-conforming to specification, and it is this proportion alone (regardless of the underlying mean and variance) that, via the OC curve, determines the chance of the batch being accepted or rejected.

The final plot in Figure 2 shows the effect of measurement error that is not allowed for—the observations X have the same mean as the plot immediately above it, but they have a larger variance, simulating the effect of measurement error. Here the measurement error has variance equal to the underlying variance of the actual moisture content values, in other words the variance is doubled, corresponding to $\gamma = 1$ (see Section 3.4). In this scenario we see that considerably more than 10% of the batches will be rejected, although the distribution of the true moisture content (excluding the measurement error) is the same as in the plot immediately above it. This third plot has OC curve equal to the dashed curve in Figure 1, while the two plots in the top line of Figure 2 have the bold OC curve in Figure 1.

3 Approaches considered

Previous work has of course been done in this area, but Fonterra was looking for simpler, clearer or more directly justifiable methods to demonstrate when different sampling systems are equivalent. Initially the discussion group began with a relatively clean slate, to facilitate fresh ideas, not influenced or contaminated by previous approaches. As various ideas and approaches were suggested and investigated, it became apparent that most had already been considered and the industry representative revealed various relevant documents and previous work. This confirmed that the approaches already considered are indeed sensible; however, sequential testing by producer then consumer had seldom been examined.

We have two main problems. First, how to compare two sampling systems, and second, how to modify different sampling systems so they produce the same decision with a specified level of confidence, or otherwise ensure that consumers do not reject batches previously accepted by the producer.

It is natural to compare acceptance sampling systems via their OC curves. In the following subsections we consider some issues of general relevance and examine some suggested approaches.

3.1 What does equivalence mean?

Even if sampling systems are equivalent in the sense that they both give the same probability of accepting a batch with a given proportion non-conforming, that is, they have identical OC curves, they will not in general reject the same batches. For example, if there is a 10% chance of rejecting a batch of a given quality, on average one in ten such batches would be rejected but it will not necessarily be that each method rejects the same one in ten. The producer would send 90% of such batches to the consumer, and for each of these the consumer (if using a sampling system with identical OC curve) would independently reject one in ten of these, on average.

Identical OC curves mean only that *on average*, or in the long run, the two sampling systems behave equivalently. They are equivalent in a statistical sense, not on an individual decision basis.

A natural approach to comparing sampling systems is to compare their OC curves. If the OC curve are the same the long term behaviour of the test will be the same. The difference between two systems could be quantified based on area between the OC curves, maximum distance apart, or differences at specified points on the curve—noting that not all parts of the curve have equal importance.

In the next section we formulate our general approach, and then consider some applications to OC curves.

3.1.1 Statistical decision equivalence of tests

As previously mentioned, we would like to gain information on two decision methods about certain characteristics of batches. The decision rules can be based on different statistical principles or formulae and can use measurements of different specification parameters. However, it is desired that these methods give the same decision given identical quality characteristics. For example, the first method might be based on heating butter, the second method might use some chemical tests, but both methods must answer the same question: does a batch of butter have moisture content less than 16%?

Let Σ be a vector of specification characteristics. Let us denote by K_1 the set of values of theoretical specification characteristics for which “an individual item satisfies the requirements”, and by K_2 the set of values for which “an individual item does not satisfies the requirements”. Therefore K_1 and K_2 are disjoint. It may happen that some values of specification characteristics are not in $K_1 \cup K_2$ (there is a “gap” between K_1 and K_2). For example, if additional tests are required for values of specification characteristics between K_1 and K_2 in order to decide about acceptance or rejection of an individual item.

We denote by $\hat{\theta}_1 = \hat{\theta}_1(X_1, \ldots, X_n)$ the test statistic of the first method, and $\hat{\theta}_2 = \hat{\theta}_2(X_1, \ldots, X_n)$ the test statistic of the second method. C_1 and C_2 are acceptance regions for the first and the second method respectively; that is, if $\hat{\theta}_1 \in C_1$ ($\hat{\theta}_2 \in C_2$) then we accept batches, based on the first (second) method's results.

Individual items can be either conforming to specification (for example, moisture content 16% or less—the set K_1) or non-conforming (set K_2 consists of values greater 16%). The statistic $\hat{\theta}_1$ ($\hat{\theta}_2$) relates to a sample of observations from a batch. The decision required is whether to accept the batch (is the statistic $\hat{\theta}_1$ ($\hat{\theta}_2$) in the acceptance region C_1 (C_2)?), and this will depend on the extent to which the individual items in the sample meet the specification. The measurements might or might not be direct measurements of moisture content, and the two tests might measure different characteristics, but both have the aim of rejecting batches that do not meet the maximum 16% moisture content criterion.

We have the same decision for both methods if $\hat{\theta}_1 \in C_1$ and $\hat{\theta}_2 \in C_2$, or $\hat{\theta}_1 \in \overline{C}_1$ and $\hat{\theta}_2 \in \overline{C}_2$, where $\overline{C}$ is the complement of C. From a practical point of view we are mainly concerned with the first type of decision.

How can we define statistical decision equivalence of two methods? We suggest the following.

Definition 1. *The method based on defining statistic $\hat{\theta}_2$ is statistically decision equivalent (simply called "equivalent" in the remainder of this article) to the method based on $\hat{\theta}_1$ if:*

1. $\inf_{\Sigma\in K_1} \Pr(\hat{\theta}_1 \in C_1 \mid \Sigma) \geqslant 1-\alpha\,, \quad \inf_{\Sigma\in K_1} \Pr(\hat{\theta}_2 \in C_2 \mid \Sigma) \geqslant 1-\alpha\,,$
2. $\sup_{\Sigma\in K_2} \Pr(\hat{\theta}_1 \in C_1 \mid \Sigma) \leqslant \beta\,,$

where $1-\alpha$ is chosen to take large probability values (close to 1), and β takes small probability values (close to 0).

Note that the two methods are not mathematically equivalent (symmetric):

we have a condition on substandard batches being misidentified as good ones only for method 1 (Fonterra's method).

This definition of equivalence does not mean that we are trying to find almost equal OC curves. OC curves can be quite different for equivalent methods. Equivalence in Definition 1 means that:

1. with high probability $1-\alpha$ both methods accept a batch for the set K_1 of values of specification characteristics; and
2. with high probability $1-\beta$ a batch is rejected for the set K_2 by method 1.

Of course, Definition 1 holds in those cases where OC curves are very close.

Definition 1 is similar to classical approaches in hypotheses testing theory [4] and the theory of statistical decision [2, 10]. However, in our definition we have two statistics $\hat{\theta}_1$ and $\hat{\theta}_2$, and two acceptance regions C_1 and C_2, which are in general different. Also, the second condition only deals with the first statistic (method), because incorrect acceptance of batches by customers (the second method) does not reduce Fonterra's profit, and therefore is of no concern to the producer.

In theory, if both methods and underlying data distributions are known then probabilities in the above definition can be calculated theoretically or found numerically by simulations (if it is difficult to derive exact formulae). In the latter case, one needs to simulate data from a known underlying distribution for the chosen Σ, compute test statistics $\hat{\theta}_1$ and $\hat{\theta}_2$ for each simulation, and use the empirical probabilities:

$$\Pr\left(\hat{\theta}_i \in C_i \mid \Sigma \in K_1\right) \approx \frac{\text{number of } \hat{\theta}_i \text{ in } C_i}{\text{total number of simulations}}\,, \quad i = 1, 2\,. \tag{1}$$

Example 1. Suppose we use underlying normal distributions, the OC curve approach, and there are three methods (see more details in Sections 3.6 and 3.7.2). Suppose that Figure 3 gives the OC curves. Let $K_1 = [0, 25]$, $K_2 = [150, 200]$, $1-\alpha = 0.8$, $\beta = 0.4$, the acceptance regions be $C_1 = C_2 = \{1\}$, $\overline{C}_1 = \{0\}$, and the test statistics $\hat{\theta}_1$ and $\hat{\theta}_2$ take on only two values: 0 (the

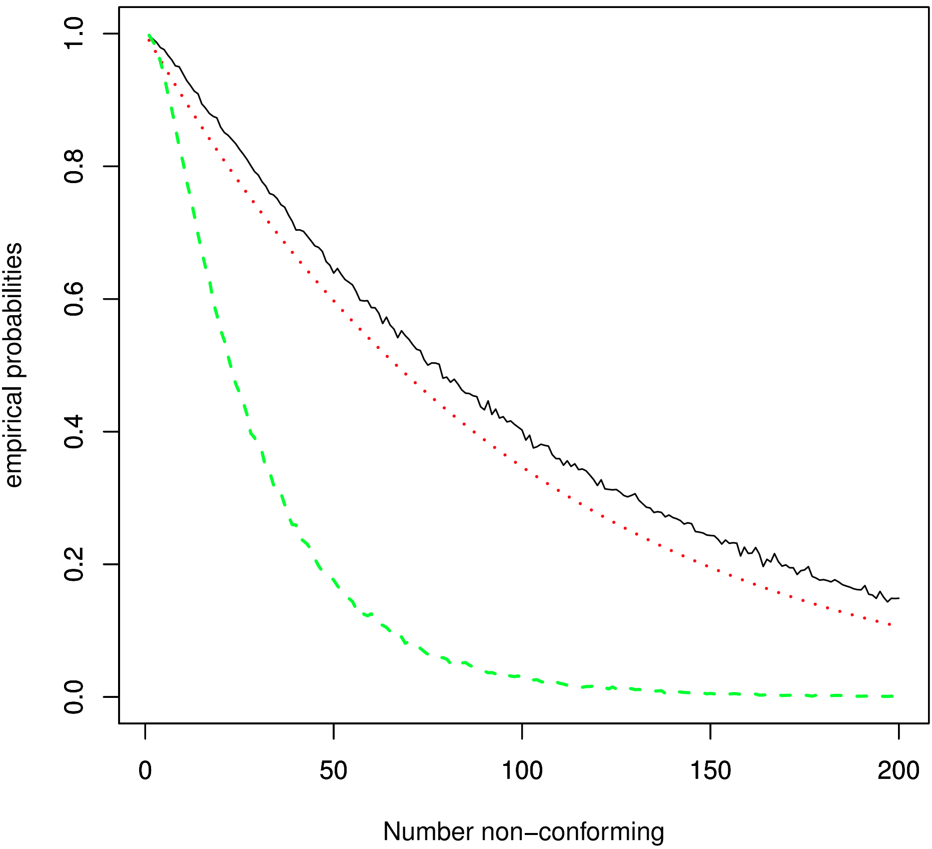

Figure 3: Red and black OC curves are for equivalent tests.

batch does not conform to some standard) and 1 (the batch conforms to the standard). Then Figure 3 clearly shows that method 2 is equivalent to method 1 (red and black OC curves), but method 3 (green OC curve) is not equivalent. R code for this example is given in appendix A.

3.2 Distributional assumptions

Standard theory about OC curves typically assumes that the observed variable X is normally distributed, and that the test statistic such as $\overline{X}+kS$ follows (approximately) a normal distribution. The values for n_0 and k_0 in Section 3.4 are estimated in this way. However, even if X is normally distributed, we know that S is not, so it is clear that this is an approximation. Furthermore, the assumed mean and variance for $\overline{X}+kS$ are both approximations. For the case with no measurement error and variance unknown, to derive n and k we assume that $\overline{X}+kS$ has mean approximately $\mu+k\sigma$ and variance approximately $(1+k^2/2)\sigma^2/n$, where the n sampled observations X are assumed to arise from a normal distribution with mean μ and variance σ^2 [15, Equations (17) and (18)].

Section 3.7 examines the effect of non-normality on the OC curve.

3.3 Variability of sampling system

Different variability of sampling systems is a likely source of differences between systems. This could arise because of different sample sizes or because of different measurement error variation. When sampling is used and measurement errors exist, there is always the possibility that a batch accepted by the producer will be rejected by the consumer. The underlying process variation, for example the variation of the true percentage moisture between items in the batch, should be the same in all sampling systems (except possibly due to deterioration during transport or storage—this is a separate issue which we do not address), but the actual value observed would be more variable if individual measurements are less accurate.

The larger the sample size and the smaller the variance of the observations (namely the smaller the measurement error variance), the closer we can allow the true mean of the process to go to the tolerable limit. Reducing the process variation, namely making the product more consistent, has the same effect because it reduces the chance of items outside of specification occurring.

It is process variation that is of importance, and measurement error is a nuisance, but we cannot usually separately estimate them. However, if we make multiple measurements on sampled items then it is possible to decompose the total variance into process variance and measurement error. This is utilised in Sections 3.4 and 3.5.

One suggestion for modifying a sampling system to achieve (statistically) the same decision as another sampling system is to use confidence intervals rather than point estimates for the batch acceptance criterion. This helps in situations with different variability. A sampling system with small variability can confidently assess a batch as acceptable. If the consumer's sampling system has greater variability, due to measurement error or otherwise, there is a greater chance they will reject a truly acceptable batch. Deriving a confidence interval for the rejection criterion makes clear the inherent uncertainty. Using the confidence interval approach, the parties could agree that if the whole confidence interval is outside some specified limit, there will be no dispute, but if the confidence interval spans both acceptable and unacceptable values of the criterion, further testing or negotiation is necessary.

3.4 Theoretical OC curves

Inspection by attributes applies when individual items have a binary outcome: conforming to specification, or non-conforming. An acceptance sampling system specifies a sample size and a maximum number of non-conforming items that the sample may contain for the batch to be accepted. Inspection by variables applies when the variable X of interest is a single measurement on a continuous scale (for example moisture content in butter) and a sample item is said to conform to specification if it is smaller than a specified upper limit L, for example a maximum of 16% moisture in butter. (Lower specification limits can be treated analogously, such as a minimum requirement of 80% fat in butter.) Because more information is utilised, inspection by variables can result in smaller sample sizes required, albeit at the expense of making some distributional assumptions. Inspections by attributes or variables can be described by an Operating Characteristic or OC curve.

In this section we focus on inspection by variables. The sampling system specifies a sample size n and a multiple k such that a batch will be accepted provided the statistic $\overline{X}+kS$ is no larger than the specification limit L, where $\overline{X}$ is the sample mean and S is the sample standard deviation. The values n and k are determined such that there is probability $1-\alpha$ of accepting a batch with proportion non-conforming at the AQL, and probability β of accepting a batch when the proportion non-conforming is LQ. Typically we assume that the variable X follows a normal distribution with mean μ and variance σ^2. If σ^2 is known, we derive the following values for n and k respectively, choosing the solution with the smallest sample size n:

$$n_0 = \left(\frac{\zeta_{1-\alpha}+\zeta_{1-\beta}}{\zeta_{1-\text{AQL}}-\zeta_{1-\text{LQ}}}\right)^2, \quad k_0 = \frac{\zeta_{1-\beta}\zeta_{1-\text{AQL}}+\zeta_{1-\alpha}\zeta_{1-\text{LQ}}}{\zeta_{1-\beta}+\zeta_{1-\alpha}}, \tag{2}$$

where ζ_p is the pth quantile of the standard normal distribution, that is, $\Pr(Z<\zeta_p)=p$ where $Z\sim N(0,1)$ [15]. When σ^2 is unknown, a similar argument leads to a modified n, increased by a multiplicative factor $(1+k_0^2/2)$. The results are summarised in Table 1.

Typically, the OC curve assumes there is no measurement error. When measurement error is present the OC curve changes, and an OC curve designed on the assumption of no measurement error will not give the results intended. We assume that when there is measurement error, we observe

$$X=\mu+B+M$$

where $\mu+B$ is the true value of the variable of interest (such as percent moisture), $B\sim N(0,\sigma_B^2)$, and M is measurement error, $M\sim N(0,\sigma_M^2)$, so that σ_B^2 is purely process variance or variance of the actual values within the batch, and σ_M^2 is measurement error variance. We define the ratio of measurement error standard deviation to process standard deviation to be $\gamma=\sigma_M/\sigma_B$.

Table 1: Sampling system specifications for inspection by variables under different scenarios for measurement error and known or unknown variances. The three entries in each cell are: test statistic $\overline{X}+kS$ (accept the batch if the statistic is less than L), multiplier k and sample size n, based on Equation (2).

measurement error	variances known	variances unknown
none	$\overline{X}+k\sigma$ k_0 n_0	$\overline{X}+kS$ k_0 $n_0(1+k_0^2/2)$
present $m=1$	$\overline{X}+k\sigma_B$ k_0 $n_0(1+\gamma^2)$	$\overline{X}+kS_B$ (but S_B unknown) k_0 $n_0(1+k_0^2/2)(1+\gamma^2)$
equivalent to	$\overline{X}+k^*\sigma$ $k^*=k_0/\sqrt{1+\gamma^2}$ $n_0(1+\gamma^2)$	$\overline{X}+k^*S$ $k^*=k_0/\sqrt{1+\gamma^2}$ $n_0(1+k_0^2/2)(1+\gamma^2)$ or $n_0(1+k_0^2/2+\gamma^2)$
present $m\geqslant 2$	$\overline{X}+k\sigma_B$ k_0 $n_0(1+\gamma^2/m)$	$\overline{X}+kS_B$ k_0 $n_0[1+\gamma^2/m+k_0^2/(2m)]\times$ $\times\left[(m+\gamma^2)^2+\gamma^4/(m-1)\right]$

Wilrich [15] derived OC curves under various scenarios: without and with measurement errors, with variances known or unknown, and (when measurement error is present) with one ($m=1$) or multiple ($m\geqslant 2$) independent measurements per sampled item. Table 1 summarises his results, with the addition of an approximate result for the case when there is measurement error, the variances are unknown, and a single measurement is made for each sampled item, so that σ_B cannot be estimated.

In Table 1, the equivalence in the $m=1$ case is exact when variances are known (because $\sigma^2 = \sigma_B^2(1+\gamma^2)$), and approximate when variances are unknown (taking S^2 approximately equal to $S_B^2(1+\gamma^2)$ and assuming γ is known approximately, if γ is not known one can use some upper bounds for the ratio σ_M/σ_B instead of γ). Unknown variances give rise to an adjustment for the presence of measurement error whereby an OC curve approximately equivalent to an OC curve for the corresponding no-measurement-error case is obtained by using the same criterion $\overline{X}+kS$ but with the multiplier and sample size modified as follows:

$$k^* = k/\sqrt{1+\gamma^2}, \quad n^* = n(1+\gamma^2) \text{ or } n^* = n\left(1+\frac{\gamma^2}{1+k^2/2}\right). \qquad (3)$$

This requires γ to be, at least, known approximately.

Figure 1 illustrates the effect of adjusting k and n (using the second equation for n^* in (3)) for the example $\gamma = 1$. The modifications give an OC curve (the sparsely dotted line) very close to the nominal curve appropriate when there is no measurement error (the bold line). Wilrich [15, 16] states that if the standard deviation of measurement error is unknown, the OC curve is changed in an uncontrolled manner, and he does not see fit to consider the approximation shown in Table 1.

If the actual ratio σ_M/σ_B is larger than the assumed γ, then too many acceptable batches will be rejected—the probability of accepting a batch which meets the AQL will be less than $1-\alpha$. In general, if one uses an OC curve based on the assumption of smaller measurement error (or no measurement error), then for any proportion p non-conforming, the probability of accepting the batch will be less than the intended probability.

Conversely, if the actual ratio σ_M/σ_B is smaller than the assumed γ, then too many unacceptable batches will be accepted—the probability of accepting a batch with proportion non-conforming equal to LQ will be greater than β. In general, if one uses an OC curve based on the assumption of larger measurement error than is actually present, then for any proportion p non-conforming the probability of acceptance will be larger than the intended probability. This can work to the producer's favour if in the consumer's sampling system the sample size n and multiplier k are modified using an upper bound for the value of γ. Specifically, replace k and n from the sampling system assuming no measurement error with k^* and n^* defined by Equation (3).

3.5 Decomposing the variance

In the example of moisture content of butter, it is the actual moisture content that is important, not the measurement of it, if measurement error is present. Ideally we would like to base our acceptance criterion on the process variance, excluding the measurement error variance. This can be done, as indicated by Wilrich [15], if two or more measurements are taken of each sampled item. The two variance components σ_B^2 and σ_M^2 are estimated separately by a variance decomposition from an analysis of variance or by the method of Residual Maximum Likelihood [13]. We then use only the estimate S_B^2 of the process variance in our acceptance criterion, namely $\overline{X} + kS_B$, noting that $\overline{X}$ will still incorporate measurement error. Table 1 shows appropriate choices of k and n.

Fonterra provided two data sets with multiple measurements on each sample and we analysed these to determine variance components. In one data set, 14 equivalent samples, duplicated, were sent (blind) to 15 respected laboratories worldwide, to determine some physical characteristic. The standard deviation of the measurement error was typically less than 5% of the standard deviation of the process, and could probably be ignored without detriment.

In the second data set, five different microbiology tests were compared, on 524 samples, duplicated and again blind. Most of the samples were spiked with vastly different levels of micro-organisms to give a large range measured values. In this data set, the standard deviation of the measurement error was around 10% of the variation between the samples for all five tests, but because the latter variation was artificially inflated by the spiking, the measurement error would be, in practice, much larger relative to process error. This measurement error would be too large to ignore when comparing sampling schemes.

3.6 OC curves in equivalence for the one parameter case

For test methods which use data or some function of data from underlying probability distributions with only one parameter (only one element in Σ) OC curves may be used as described above to determine equivalence of these methods. This approach is applicable when the proportion non-conforming (on the horizontal axis of the OC curve) depends on a single parameter, such as μ when the variance σ^2 is known.

In Fonterra's example a batch of 1000 cartons of butter needs to be inspected for moisture. Fonterra's specification limit for moisture is at most $16\,\text{g}/100\,\text{g}$. Two methods are investigated.

1. Take samples $X_1, \ldots, X_7$ from seven cartons and test for moisture content. Calculate the empirical average $\overline{X}$ and standard deviation S of these samples. If $\overline{X} + 1.5S < 16$, then accept the batch.
2. Take samples $X_1, \ldots, X_{10}$ from ten cartons and test for moisture content. If there are no out of specification test results, then accept the batch.

If we assume that the variables X_i are from the normal distribution $N(\mu, \sigma^2)$, we have a two parameter underlying distribution. However, if we consider the number of out of specification items Σ as a parameter for OC curves, then we have an underlying probability distribution with only one parameter. In the latter case, for given Σ, the parameter μ is a function of σ^2, such that for various σ^2 in the underlying distribution $N(\mu, \sigma^2)$ the probability of acceptance of the batch is same.

We use formula (1) to calculate empirical probabilities. As in Example 1 we plot OC curves for both methods with specification characteristic Σ, see Figure 4. For these plots X_i were simulated from normal distributions with $\sigma = 1$ (the first plot) and with $\sigma = 2$ (the second plot). In each case the mean μ was chosen to get a specified value of Σ. It is clear from Figure 4 that method 2 is equivalent to method 1, because the two curves are almost identical and therefore both inequalities in Definition 1 are satisfied. Similarly, for other values of the parameter σ we obtain identical OC curves. Therefore, simulation results support the equivalence of these two methods under normality assumptions.

We used $10\,000$ simulations to build each plot. Method 1 OC curve is graphed as the solid line and dots are used for method 2 curve. The left and right plots are practically identical as we chose μ to get same parameter Σ for both cases. R code for this example is given in appendix B.

3.7 Limitation of the OC curve approach

3.7.1 Sensitivity to the underlying distribution

We now consider two test methods from Section 3.6 for data with non-normal underlying probability distributions. We show that in this case OC curves do not show equivalence of the methods.

Example 2. We use underlying uniform distributions. Let butter satisfy requirements if not more than 25 samples out of 200 have moisture content more than 16%, that is, $K_1 = [0, 25]$. Let the batch not satisfy requirements if at least 150 samples out of 200 have moisture content more than 16%, that is, $K_2 = [150, 200]$. We assume that $1-\alpha = 0.8$, $\beta = 0.4$, the acceptance regions are $C_1 = C_2 = \{1\}$, $\overline{C}_1 = \{0\}$, and the test statistics are determined by the same methods as before. The empirical probabilities obtained by simulation are shown in Figure 5. R code for this example is given in appendix C.

It is clear that method 2 is not equivalent to method 1 (red and black OC curves). Indeed, for the set K_1 and $1 - \alpha = 0.8$ the first condition in Definition 1 does not hold. Therefore, the OC curves approach is very sensitive

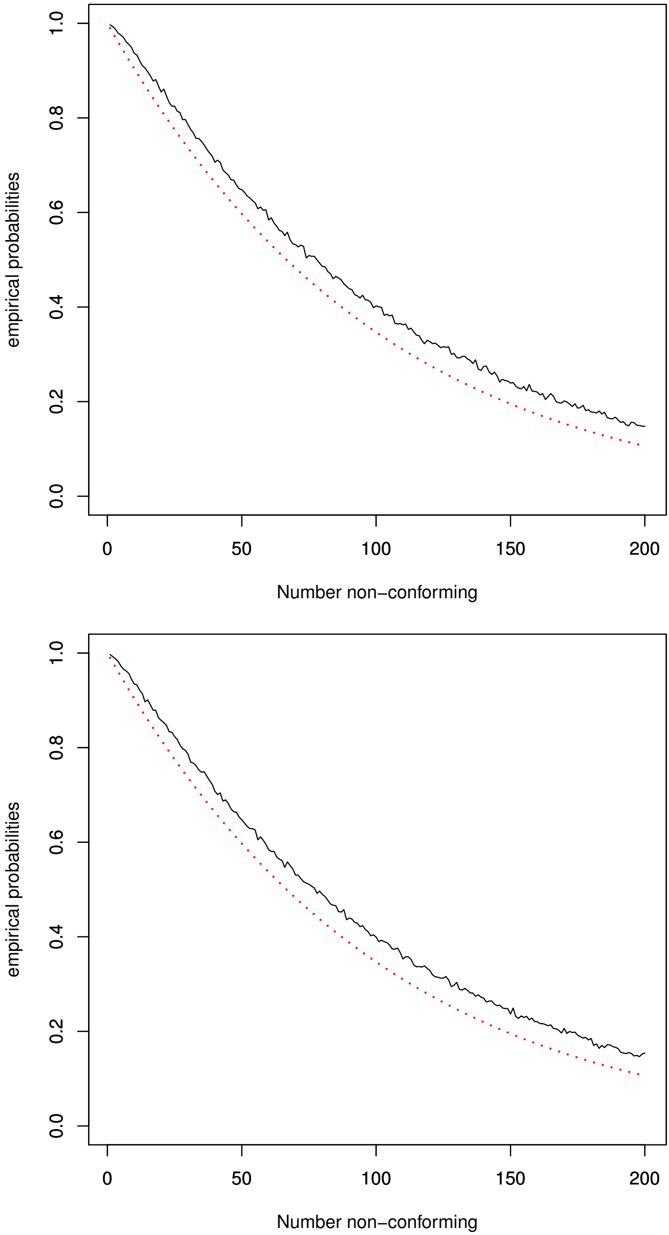

Figure 4: OC curves for $\sigma = 1$ and $\sigma = 2$.

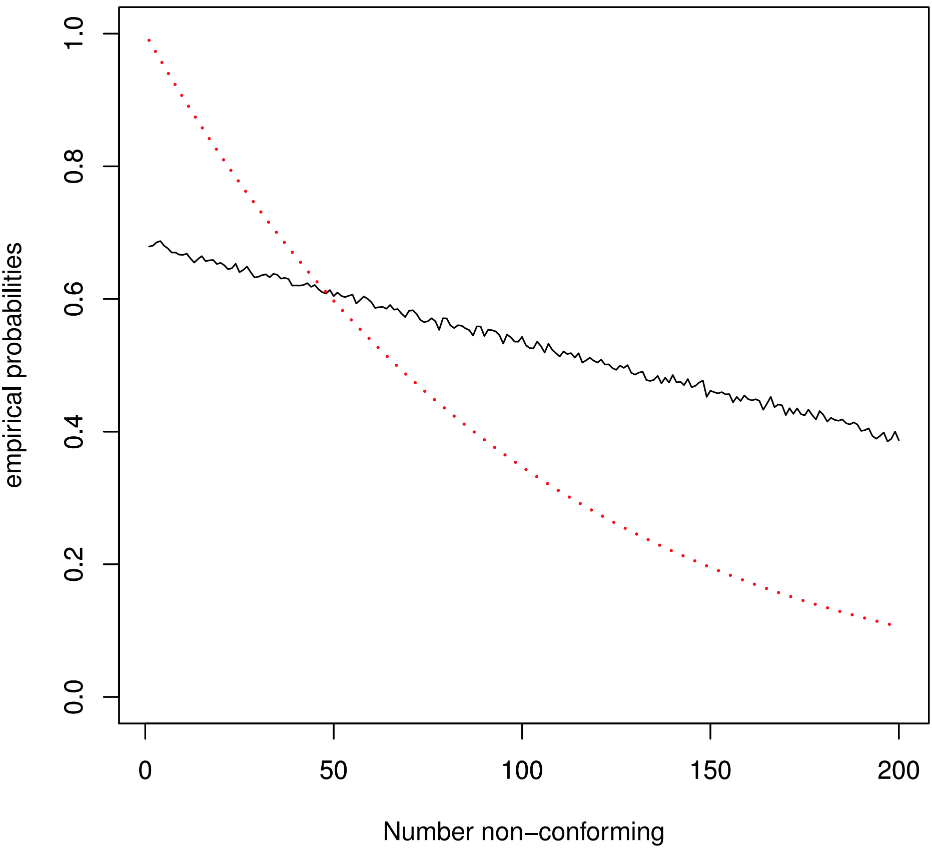

Figure 5: OC curves for methods 1 and 2, and for uniformly distributed data.

to underlying data distributions and in many cases experimental information on underlying distributions is required to derive reliable results.

3.7.2 Multiparameter cases

In this section we consider test methods and underlying probability distributions for cases in which there is not a one-to-one correspondence between distributional parameters and the proportion non-conforming. We show that it is impossible to use OC curves in these cases (as it was described in Sections 3.1.1 and 3.6) to determine equivalence of two methods.

Example 3. Let us add a third method to the two methods in Section 3.6:

3. Take samples $X_1, \dots, X_7$ from seven cartons and test for moisture content. Calculate empirical average $\overline{X}$ of these samples. If $\overline{X} + 2 < 16$, then accept the batch.

Similarly to Section 3.6, we assume that the variables X_i are from the normal distribution $N(\mu, \sigma^2)$. Section 3.6 showes that for methods 1 and 2 it is possible to use Σ as the sole parameter and for different σ OC curves do not change.

However, if we consider the number of non-conforming items Σ as the only parameter, then it is impossible to calculate acceptance probabilities for method 3. The reason is that Σ does not completely determine acceptance probability distributions in method 3. We have to add more parameters to uniquely determine acceptance probabilities in method 3.

For example, the additional parameter σ allows us to uniquely specify the distribution. However, if we change σ we do not have the same OC curve in method 3 as in methods 1 and 2, see Figure 6.

For $\sigma = 2$ method 3 is equivalent to methods 1 and 2, as we have larger acceptance probabilities in method 3 (see the second plot in Figure 6). However, for $\sigma = 1$ method 3 is not equivalent to methods 1 and 2 (see the first plot in Figure 6).

R code for $\sigma = 1$ is given in appendix A. For $\sigma = 2$ the same code can be used but with `s<-2`.

In general, the one dimensional OC curve approach cannot be used if methods use data with multiparameter distributions. An exception is test methods, for example methods 1 and 2 with normally distributed data, as discussed in Section 3.4, in particular in relation to Figure 2.

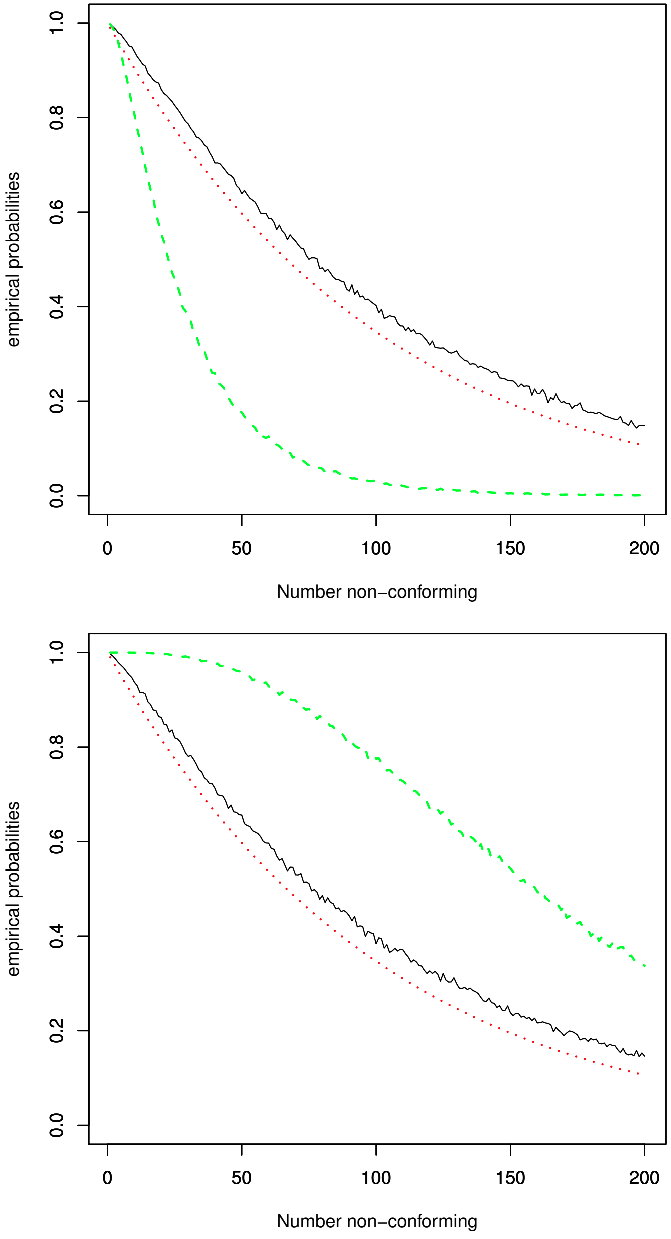

Figure 6: OC curves for $\sigma = 1$ and $\sigma = 2$.

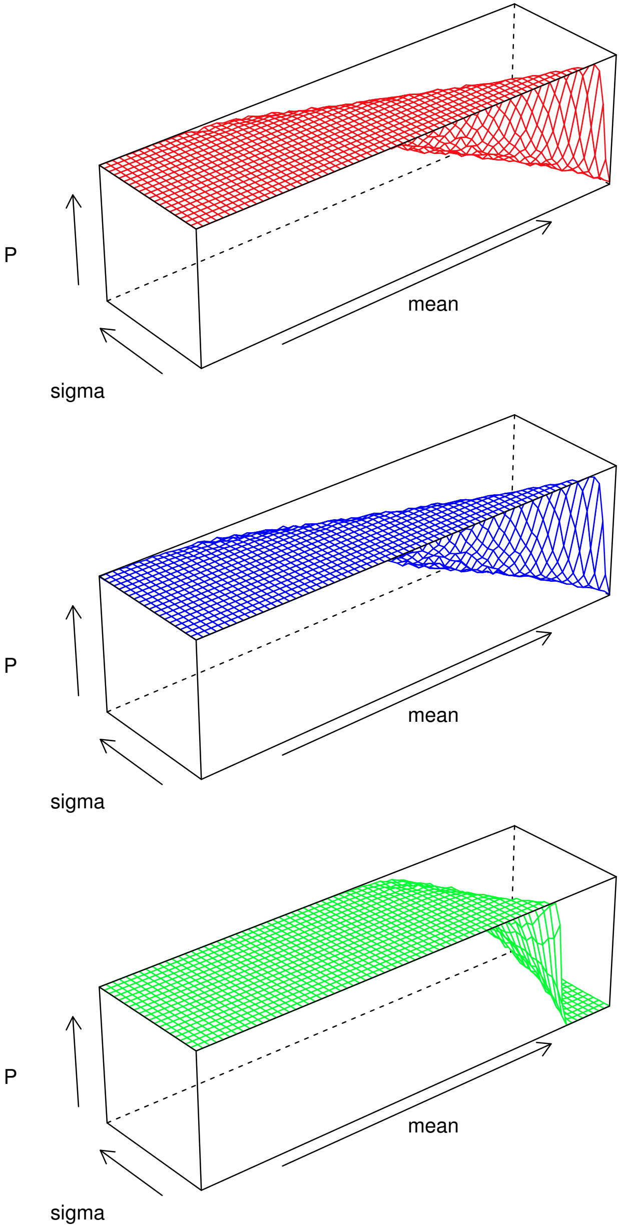

Figure 7: OC surfaces for methods 1–3 and underlying $N(\mu, \sigma^2)$ distribution.

3.7.3 OC surfaces

Some of the above mentioned limitations of the OC curves approach are overcome by producing sets of OC curves for different values of parameters of underlying distributions and investigating their properties. For two parameter cases this is done most efficiently using OC surfaces, by producing a three dimensional plot of acceptance probabilities for all possible values of two underlying parameters. If the number of parameters is greater than two, then such cases can be analysed by producing OC surfaces for various pairs of parameters.

Example 4. Let the variables X_i be from the normal distribution $N(\mu, \sigma^2)$. We plot acceptance probabilities as a function of μ and σ^2. For methods 1, 2, and 3 from Sections 3.6 and 3.7.2 we obtain OC surfaces shown in Figure 7. They illustrate the equivalence of methods 1 and 2 and the difference of method 3. R code is given in appendix D.

3.8 Inter-rater agreement

There are various classical statistical methods to determine inter-rater agreement, which are potentially useful for solving equivalence problems [6]. Unfortunately these methods rely on assumptions which are not appropriate for Fonterra's data. We demonstrate this using one such measure of inter-rater reliability, Cohen's kappa coefficient [3].

Cohen's kappa is a statistical measure of agreement between two raters which classify items into mutually exclusive categories. It is defined by

$$\kappa = \frac{\Pr(a) - \Pr(b)}{1 - \Pr(b)} ,$$

where $\Pr(a)$ is the relative observed agreement among raters, namely the number of decisions in agreement out of the total number of items, and $\Pr(b)$ is the probability of chance agreement. If the raters are in complete agreement, then $\kappa = 1$. If there is no agreement other than what would be expected by chance, then on average $\kappa = 0$.

For Fonterra’s equivalence problem we analyse batches. Each batch is assessed by two methods and each method either accepts the batch (”Yes”) or rejects (”No”). Therefore we represent the data as follows.

		Method 2	
		Yes	No
Method 1	Yes	k_1	k_2
	No	k_3	k_4

The observed relative agreement is

$$\Pr(a) = \frac{k_1 + k_4}{k_1 + k_2 + k_3 + k_4} .$$

The probability of random agreement is

$$\Pr(b) = \frac{k_1 + k_2}{k_1 + k_2 + k_3 + k_4} \cdot \frac{k_1 + k_3}{k_1 + k_2 + k_3 + k_4} .$$

In most cases for Fonterra’s data we have k_1 much larger than k_2, k_3 and k_4. Therefore the value of $1 - \Pr(b)$ is likely to be close to 0, and κ is unlikely to be statistically significant. Moreover, Fonterra would not in general dispatch batches which did not pass Fonterra’s tests, so $k_3 = k_4 = 0$, and Cohen’s kappa is degenerate.

Cohen’s kappa requires a reasonable number of both positive and negative decisions, rather than the imbalance expected in our setting where we expect the product to be generally good. Producers will not intentionally make bad product, which makes it difficult to determine reliably whether sampling systems agree in a range of situations. One suggestion to cope with this was to combine samples from several batches with a wider range of quality to get greater variability. Another possibility might be to adjust our criteria to be more stringent, so that a larger proportion of batches are notionally rejected. However, based on these observations we concluded that Cohen's kappa and other classical measures of inter-rater agreement are not helpful measures of agreement the context of comparing acceptance sampling systems.

3.9 Ensuring the consumer does not reject batches received

Without making the quality impractically high or extremely consistent, or the consumer having a test with very low sensitivity, it is impossible to ensure the consumer never rejects a batch accepted by the producer. This is because acceptance is based on probabilities and sampling, and if any items in the batch are not within specification, they could by chance be selected in the consumers sampling system. However we can reduce the chance of the consumer rejecting a batch deemed acceptable on the basis of the consumer's sampling system.

There are several ways this could be done. First, as intimated above, the producer could ensure that all product is within specification by having very stringent quality controls and providing only consistently high quality product—but this is an unfair burden on the producer if over specification product is required purely because of a consumer's inferior acceptance sampling system.

Alternatively, if the producer has knowledge of the characteristics of the consumer's sampling system, the producer can modify their own sampling system to ensure that the batches that are accepted, and consequently dispatched, have proportion non-conforming that the consumer has a very small chance of rejecting. However, once again, this puts the burden on the producer and may result in rejecting too many truly good batches. Similarly to the examples in this article, producers can use simulations for two producer methods to check their equivalence and investigate areas of OC curves/surfaces which are most affected by changes in methods.

Further, if the consumer's sampling system has particularly large measurement error, there is still a chance that good quality items could be deemed unacceptable. Thus, without some control over the parameters of the consumer's sampling system, it will be very difficult to guarantee they will not reject a batch deemed by the producer to be acceptable. Therefore, most suggestions depend on some influence on the characteristics of the consumer's sampling system.

These suggestion include having more relaxed limits such as AQL and/or allowing larger α, and possibly the specification limit L, although, as in the case of moisture and fat content of butter, this may be fixed by legislation. The contract between the producer and consumer could require certain standards in their acceptance testing, or demonstration within a specified degree of certainty (by a method agreed in advance) if the consumer wishes to dispute the acceptability of a batch. The use of confidence intervals for the proportion of the batch non-conforming is one way to examine the consumer's estimate, with the level of uncertainty due to large variability in the sampling system clearly evident. The batch could be agreed to be substandard if the confidence interval was entirely outside of the specification range. If there is ambiguity, to some extent the onus could be on the consumer to demonstrate (to an agreed confidence level) that the batch is unacceptable (for example based on a larger sample size).

There is always a trade off between rejecting good batches and sending poor batches, or batches that the consumer has a greater chance of rejecting (whether they are truly substandard or not). The cost of rejecting good batches needs to be weighed up against the cost of disputes if the consumer challenges.

4 Conclusions

We analysed some aspects of equivalence of acceptance testing systems. It was shown that classical statistical methods to quantify inter-rater reliability are not appropriate in the context of sequential testing. A general formulation for equivalence of tests was suggested. The method of inspection by variables and the Operating Characteristic (OC) curve were described and discussed. This method is well accepted but rests on assumptions of normality (at least approximately). Measurement error presents difficulties, and proper allowance must be made to avoid inappropriate rejections of acceptable batches. The effect of measurement error on the OC curve can be demonstrated, and an adjustment is possible if an estimate of the measurement error is available.

We demonstrated that the approach based on OC curves is sensitive to the type of distribution, and, except when data are normally distributed, in general the OC curves approach is not a reliable method, in particular for multiparameter distributions. We recommend using OC surfaces for two parameter distributions. Examples of R code for simulations to support our conclusions are given in the appendices.

It would be of interest to apply these results to different kinds of test methods in a range of settings. Another interesting area for further investigations is to adopt the approach to sequential tests. It would be interesting to explore robust testing (criteria) under a wide range of underlying distributions.

Acknowledgements We are grateful for the referees' comments, which helped to improve the style of the presentation. We thank David Shteinman, who contributed to MISG discussions on this project. We also thank John Shepherd and his RMIT colleagues for organizing such a stimulating MISG 2011. We are also grateful to Fonterra for the problem, data, and information and materials that are not in Open Access.

A R code for example 1

```
m <- 10000 # number of samples to simulate
k<-200
s<-1
n = 1:m # vector: n = 1, 2, ..., m; simulation number
good <- numeric(m) # initialize for use in loop
emp_probab1<- numeric(k)
emp_probab2<- numeric(k)
emp_probab <- numeric(k)
for (p in 1:k){
for (i in 1:m){
pick<-rnorm(7,mean = 16-s*qnorm(1-p/1000,mean=0,sd= 1),sd = s)
good[i] <- as.numeric(mean(pick)+1.5*sd(pick)< 16)}
emp_probab[p]<-mean(good) # approximates P}
for (p in 1:k)
{emp_probab1[p]<-choose(1000-p,10)/choose(1000,10)}
for (p in 1:k){
for (i in 1:m){
pick<-rnorm(7,mean = 16-s*qnorm(1-p/1000,mean= 0,sd= 1),sd = s)
good[i] <- as.numeric(mean(pick)+2< 16)}
emp_probab2[p]<-mean(good) # approximates P}
plot(emp_probab,ylim=c(0,1), pch=24 ,
+ ylab = expression("empirical probabilities"))
par(new=TRUE)
plot(emp_probab1,ylim=c(0,1), pch="*", col=’red’,
+  ylab = expression(""))
par(new=TRUE)
plot(emp_probab2,ylim=c(0,1),col=’green’,ylab = expression(""))
```

B R code for example 2

```
m <- 10000 # number of samples to simulate
k<-200
s<-1 # s<-2 for variance =2
n = 1:m # vector: n = 1, 2, ..., m; simulation number
good <- numeric(m) # initialize for use in loop
emp_probab1<- numeric(k)
emp_probab2<- numeric(k)
emp_probab <- numeric(k)
for (p in 1:k){
for (i in 1:m){
pick<-rnorm(7, mean = 16-s*qnorm(1-p/1000,mean=0,sd= 1),sd = s)
good[i] <- as.numeric(mean(pick)+1.5*sd(pick)< 16)}
emp_probab[p]<-mean(good) # approximates P}
for (p in 1:k)
{emp_probab1[p]<-choose(1000-p,10)/choose(1000,10)}
plot(emp_probab,pch="*", ylim=c(0,1),
+ ylab = expression("empirical probabilities"))
par(new=TRUE)
plot(emp_probab1,ylim=c(0,1),col='red', ylab = expression(""))
```

C R code for example 3

```
m <- 10000 # number of samples to simulate
k<-200
good <- numeric(m) # initialize for use in loop
emp_probab1<- numeric(k)
emp_probab2<- numeric(k)
```

```
emp_probab <- numeric(k)
for (p in 1:k){
for (i in 1:m){
pick<-runif(7, min=6+p/200, max=16+p/200)
good[i] <- as.numeric(mean(pick)+1.5*sd(pick)< 16)}
emp_probab[p]<-mean(good)}
for (p in 1:k)
{emp_probab1[p]<-choose(1000-p,10)/choose(1000,10)}
plot(emp_probab,ylim=c(0,1),pch="*",
+ ylab = expression("empirical probabilities"))
par(new=TRUE)
plot(emp_probab1,ylim=c(0,1),col=’red’, ylab = expression(""))
```

D R code for example 4

```
m <- 1000 # number of samples to simulate
k1<-70
k2<-20
emp_probab1<- matrix(data = NA, nrow = k1, ncol = k2)
emp_probab2<- matrix(data = NA, nrow = k1, ncol = k2)
emp_probab3<- matrix(data = NA, nrow = k1, ncol = k2)
for (del1 in 1:k1){
for (del2 in 1:k2){
good<-0
for (i in 1:m){
pick<-rnorm(7, mean = 1.9+del1*0.2, sd = 0.1+del2*0.2)
good<- good+as.numeric(mean(pick)+1.5*sd(pick)< 16)}
emp_probab1[del1,del2]<-good/m # approximates P method 1
}}
```

```
for (del1 in 1:k1){
for (del2 in 1:k2){
good<-0
for (i in 1:m){
pick<-rnorm(10, mean = 1.9+del1*0.2, sd = 0.1+del2*0.2)
good <- good+as.numeric(max(pick)< 16)}
emp_probab2[del1,del2]<-good/m # approximates P method 2
}}
for (del1 in 1:k1){
for (del2 in 1:k2){
good<-0
for (i in 1:m){
pick<-rnorm(7, mean = 1.9+del1*0.2, sd = 0.1+del2*0.2)
good <- good+as.numeric(mean(pick)+2< 16)}
emp_probab3[del1,del2]<-good/m # approximates P method 3
}}
library(lattice)
wireframe(emp_probab1,col='red',xlab = "mean",
+ ylab = "sigma",zlab = "P")
wireframe(emp_probab2,col='blue',xlab = "mean",
+ ylab = "sigma",zlab = "P")
wireframe(emp_probab3,add=TRUE,col='green',xlab = "mean",
+ ylab = "sigma",zlab ="P")
```

References

[1] L. J. Bain and M. Engelhardt. *Introduction to Probability and Mathematical Statistics*. Duxbury Press, 2000. M29

[2] N. N. Cencov. *Statistical Decision Rules and Optimal Inference*. American Mathematical Society, 2000. M36

[3] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement.* **20**, 1960, 37–46. M51

[4] D. R. Cox and D. V. Hinkley. *Theoretical Statistics.* Chapman and Hall/CRC, 1979. M36

[5] P. Grzegorzewski. Acceptance sampling plans by attributes with fuzzy risks and quality levels. In: P-Th. Wilrich and H. J. Lenz, editors, *Frontiers in statistical quality control,* Vol. 6, pages 36–46. Physica-Verlag, 2001. M26

[6] K. Gwet. *Handbook of Inter-Rater Reliability.* Advanced Analytics, LLC., 2010. M51

[7] A. Hald. *Statistical theory of sample inspection by attributes.* Academic, 1981. M26

[8] ISO 8196-1:2009 Milk—Definition and evaluation of the overall accuracy of alternative methods of milk analysis. M27

[9] ISO 8196-2:2000 Milk—Definition and evaluation of the overall accuracy of indirect methods of milk analysis. M27

[10] N. A. Nechval, K. N. Nechval, E. K. Vasermanis. Statistical decision equivalence principle and its applications. *Proceedings of International Conference RelStat04*, 2004, 81–89. M36

[11] S. K. Niazi. *Handbook of Bioequivalence Testing.* Informa Healthcare, 2007. M25

[12] *NIST/SEMATECH e-Handbook of Statistical Methods.* http://www.itl.nist.gov/div898/handbook/pmc/section2/pmc21.htm M25, M27

[13] H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58** (3), 1971, 545–554. http://www.ams.org/mathscinet-getitem?mr=0319325 M43

[14] S. Wellek. *Testing Statistical Hypotheses of Equivalence.* Chapman and Hall/CRC, 2002. M25

[15] P.-Th. Wilrich. Single sampling plans for the inspection by variables in the presence of measurement error. *Allgermeines Statistisches Archiv*, **84**, 2000, 239–250. M38, M40, M41, M42, M43

[16] P.-Th. Wilrich. Statistical concepts of capability of detection. *Appl. Stochastic Models Bus. Ind.* **18**, 2002, 339–346. http://www.ams.org/mathscinet-getitem?mr=1932646 M42

Author addresses

1. **B. Abbasi**, School of Mathematical and Geospatial Sciences, RMIT, Melbourne, Australia.
mailto:babak.abbasi@rmit.edu.au

2. **R. Crawford**, Fonterra, New Zealand.
mailto:Rob.Crawford@fonterra.com

3. **K. A. Haskard**, Australian Mathematical Sciences Institute, Melbourne, and Data Analysis Australia, Nedlands, Australia.
mailto:kathy@daa.com.au

4. **A. Olenko**, Department of Mathematics and Statistics, La Trobe University, Melbourne, Australia.
mailto:a.olenko@latrobe.edu.au