Assessing compatibility of two laboratories: formulations as a statistical hypothesis testing problem

2013 Metrologia 50 49

(http://iopscience.iop.org/0026-1394/50/1/49)

IOP PUBLISHING METROLOGIA
Metrologia **50** (2013) 49–59 doi:10.1088/0026-1394/50/1/49

Assessing compatibility of two laboratories: formulations as a statistical hypothesis testing problem

Andrew L Rukhin

Statistical Engineering Division, Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

Received 7 August 2012, in final form 10 December 2012
Published 9 January 2013
Online at stacks.iop.org/Met/50/49

Abstract
A decision problem frequently occurring in metrology is that of compatibility of data obtained by two (or several) different laboratories, methods or instruments. One laboratory can be a prestigious national metrology institute whose data are considered to be a gold standard or a certified reference material interval. When each laboratory presents its results in the form of a coverage interval for the measurand, several statistical approaches to this conformity assessment problem are reviewed including the classical 'equality of means' hypotheses tests. A new compatibility hypothesis is formulated in terms of consistency of laboratory results and compliance with a maximum permitted uncertainty. The power functions of these tests are compared numerically. The Kullback–Leibler information number is suggested as a directional (asymmetric) interchangeability index.

(Some figures may appear in colour only in the online journal)

1. Introduction

The problem of establishing equivalence, conformance or compatibility of several measurement sets is an important area of statistical metrology. The recent survey [1] discusses about thirty publications in this field over the last fifteen years. It overviews standards organizations' current approaches to formal conformity testing, defining methodologies for assessing the compliance degree of user measurement/uncertainty with a specification standard.

The standard's nominal acceptability region is defined by a range contained between a lower limiting value (T_L) and an upper limiting value (T_U). Clearly acceptable user's coverage intervals are those falling entirely in the interval between T_L and T_U. Unacceptable results are those falling way below the T_L or considerably above the T_U. When user's interval overlaps with the acceptability interval, decision rules are required to judge whether the result is conformant or not. A common approach to deal with this case is to adopt a guard band which is an offset from the specification limits (T_L, T_U), extending or restricting them further to formal acceptance/rejection region boundaries. 'Simple' acceptance/rejection rules take the acceptance zone to be the specification zone. 'Relaxed' acceptance/rejection rules inflate the specification zone by use of an added (outward) guard band. 'Stringent' rules deflate the specification zone by use of an inward subtracted guard band.

Prominent among the publications on conformity testing, 'the most useful tools' [1] are the guidelines of ISO 10576-1 [2] and EURACHEM/CITAC [3] which follow principles set out in ASME [4]. Current recommendations of [2] take into consideration the effect of the user's interval on any decision rule in terms of producer's risk and user's risk. The most recently issued set of guidelines [3] endorses more relaxed (broader) acceptance zones and more stringent (narrower) rejection zones via mentioned guard bands. It extends the rule construction using probability (e.g. Gaussian) models for specification limits with and without the guard band regions.

These guides seem to refer to a statistical hypothesis testing situation. Indeed, the acceptance zone and the rejection zone are formulated as components of a decision rule. However, the null hypothesis which mathematically describes the compatibility in terms of population parameters is not formulated. An attempt to do so is presented in this paper which suggests that in addition to the equality of means, the relevant hypothesis restricts the ratio of uncertainties involved (or imposes the minimum for the measurement capability index).

2. Formulation

We discuss the conformance testing problem in the context of two labs each providing its summary data in the form of coverage intervals for the measurand. The issues related to the testing of compliance with imposed legal or regulatory limits involving inspection and the quality control process are not considered here. To fix the notation, let lab 1 report μ_1 for the mean and σ_1 for its uncertainty. It is assumed that this lab is in better agreement with SI units and/or with the standards established by national metrology institutes. In some cases the degrees of freedom ν_1 (typically large, if not infinite) is also provided. Under the normality condition, this lab's $(1-\alpha)100\%$ coverage interval for a measurand is $\mu_1 \pm z_{\alpha/2}\sigma_1$. Here for an error probability α, $0 < \alpha < 1$, z_α denotes $(1-\alpha)$th quantile of standard normal law with the distribution function Φ, $1 - \Phi(z_\alpha) = \alpha$. This quantity should be replaced by a percentile of a t-distribution if the degrees of freedom ν_1 is small.

The expansion factor 2, which approximately corresponds to $\alpha = 0.05$, is commonly used in metrology in the absence of other information about effective degrees of freedom. Then μ_1 is the certificate value, $2\sigma_1$ is the expanded uncertainty. In the motivating examples, lab 1 may be a gold standard providing specification limits $T_L = \mu_1 - 2\sigma_1$, $T_U = \mu_1 + 2\sigma_1$ or it may provide the certified reference material (CRM) coverage interval $\mu_1 \pm 2\sigma_1$ for a measurand.

The data of lab 2 result in summary statistics, say, the sample mean $\bar{x}$ and its uncertainty, $u = u(\bar{x})$, which are lab's best estimates of its true mean μ_2 and of the standard deviation of $\bar{x}$. If lab 2 had performed n independent measurements, the degrees of freedom $\nu = \nu_2 = n - 1$ is typically smaller than ν_1. Then u^2 unbiasedly estimates σ_2^2/n, where σ_2^2 is the variance of lab 2 measurements. Under normality assumption the distribution of u^2 is that of $\sigma_2^2\chi^2(\nu)/\nu$ where $\chi^2(\nu)$ denotes a χ^2 random variable with ν degrees of freedom. Then the (biased) maximum likelihood estimator of σ_2^2/n is $\nu u^2/n$. The unknown bias of lab 2 is $\Delta = \mu_2 - \mu_1$, and its coverage interval is $\bar{x} \pm t_{\alpha/2}(\nu)u$ with $t_\alpha(\nu)$ denoting the $(1-\alpha)$th percentile of a t-distribution with ν degrees of freedom.

Classical statistics offers only limited guidance on assessing conformity. The next section reviews the methodology available for testing the equality of two means in the normal model. Our focus is on two-sided alternatives: $\mu_2 \neq \mu_1$ or $\Delta \neq 0$, although in some engineering or environmental applications alternatives of the form $\mu_2 \leqslant \mu_1 - 2\sigma_1$ or $\mu_2 \geqslant \mu_1 + 2\sigma_1$ are of interest.

3. Non-overlapping intervals and the Behrens–Fisher problem

In the notation of section 2, assume that lab 2's sample mean $\bar{x}$ is normal $N(\mu_2, \sigma_2^2/n)$, and u^2 is a multiple of a $\chi^2(\nu)$ random variable, $u^2 \sim (\sigma_2^2/n)(\chi^2(\nu)/\nu)$. Then one can use a simple t-test of the equality of the means: $\mu_2 = \mu_1$ for some value of the normal variance σ_2^2. Under this hypothesis, the ratio $(\bar{x} - \mu_1)/u$ has a t-distribution with $\nu = n - 1$ degrees of freedom. Conformity is rejected when

$$\frac{(\bar{x} - \mu_1)^2}{u^2} \geqslant t_{\alpha/2}^2(\nu). \tag{1}$$

However, σ_1 does not enter in this procedure which consequently ignores the some of available information.

ISO 10576-1 [2] explicitly states that user intervals should not be employed in the official designation of T_L and T_U, so that the tolerance is set for operational reasons and the decision process that follows from a result is intended to guarantee acceptable conformity of true values with the stated tolerance. Such a setting could lead to a frequent rejection of the compatibility hypothesis. Indeed, fairly often lab 2's interval, $\bar{x} \pm t_{\alpha/2}(\nu)u$, does not even intersect the CRM interval. A standard interpretation of two non-overlapping coverage intervals is that the two labs do not conform. Of course overlapping intervals do not imply that the hypothesis, $\mu_2 = \mu_1$, is to be accepted. See [5] for a well justified critique of studies that judge the significance of differences by examining the intersection of two intervals, and section 8 for numerical results on the power of this procedure.

Still, compliance testing performed on the basis of the intersecting intervals is promoted in metrology [6]. Such a test rejects conformity, i.e. the two intervals do not overlap, when

$$|\bar{x} - \mu_1| \geqslant 2\sigma_1 + t_{\alpha/2}(\nu)u. \tag{2}$$

It is possible that (1) holds, but (2) does not. However, (2) implies (1).

The null hypothesis: $\mu_2 = \mu_1$ for some unspecified values of σ_1 and σ_2, can be interpreted as the Behrens–Fisher problem [7]. This is a notoriously difficult theoretical question. The fact is that there is no unique 'optimal' test in this situation especially when n is not large. One has to specify the constant f to employ the critical or rejection region,

$$\frac{(\bar{x} - \mu_1)^2}{\sigma_1^2 + u^2} \geqslant f. \tag{3}$$

Rather embarrassingly for statistical theory, the known approximate solutions for f in (3) do not always agree. The classical solution leads to an awkward combination of the degrees of freedom both of which must be provided. When $\nu_1 = \infty$, this Welch–Satterthwaite's formula gives $f = t_{\alpha/2}^2(\nu_{\text{eff}})$, with $\nu_{\text{eff}} = \nu(1 + \sigma_1^2/u^2)^2$, which may result in a poor approximation if σ_1^2 is large [8].

One can argue that treating two labs symmetrically, as in (3), is not appropriate in our situation. Formally, test (3) with $f = t_{\alpha/2}^2(\nu)$ can be also derived from a Bayesian model suggested in the next section.

4. Bayesian approach

Here the unknown measurand μ is supposed to have a (prior) probability distribution, either because it is random according to the Bayes theory tenets or because it is a fixed unknown constant assigned a prior distribution reflecting the current state of knowledge. The situation described above suggests the prior

distribution for μ which is normal with the mean μ_1 and the variance σ_1^2. Indeed, in the CRM context, this is a classical model for the summary of extensive measurement work done to establish the certificate value which is encompassed in an interval of half-width $z_{\alpha/2}\sigma_1$ so that the CRM certificate gives an approximate $(1-\alpha)100\%$ confidence interval.

Assume that the measurements x_i, $i = 1, \ldots, n$, of lab 2 can be represented as $x_i = \mu + \Delta + \epsilon_i$ where Δ is the non-random bias of this lab, μ is the random measurand discussed above, and ϵ_i represents zero mean independent measurement error with some variance σ_2^2. Then the sample mean of lab 2's measurements has expectation $\mu_1 + \Delta$ and its variance is $\sigma_1^2 + \sigma_2^2/n$. Thus, this variance cannot be smaller than σ_1^2, which may have some appeal to metrologists who believe that by combining results of their measurements with another coverage interval, one hardly can diminish the overall uncertainty. The maximum likelihood estimator of the variance of the user's sample mean is now $\max(\sigma_1^2, u^2)$.

Under the hypothesis, $\mu_2 = \mu_1$, in the decomposition

$$\sum_i (x_i - \mu_1)^2 = \sum_i (x_i - \bar{x})^2 + n(\bar{x} - \mu_1)^2,$$

the first sum has the distribution of $\sigma_2^2\chi^2(n-1)$ and the second of $(\sigma_2^2 + n\sigma_1^2)\chi^2(1)$. In this situation the statistic $(\bar{x} - \mu_1)^2/u^2$ is distributed as $(1 + n\sigma_1^2/\sigma_2^2)t^2(n-1)$. Thus to test the conformance hypothesis in this situation using (1) would lead to an unacceptably high number of rejections especially when the ratio σ_2/σ_1 is small. Estimating the factor $1 + n\sigma_1^2/\sigma_2^2$ by $1 + \sigma_1^2/u^2$ results in a critical region (3) with $f = t^2_{\alpha/2}(\nu)$.

A better test (the so-called Wald test) is based on the above mentioned maximum likelihood estimator of the variance of $\bar{x}$. The two means are declared to be different when

$$\frac{(\bar{x} - \mu_1)^2}{\max(\sigma_1^2, u^2)} \geqslant t^2_{\alpha/2}(n-1). \tag{4}$$

Unlike the existing body of subjective or non-informative Bayesian techniques, we suggested here an informative objective prior distribution. Indeed, this prior reflects that certified values are based on averages of many repeated experiments, rather than on a subjective opinion about the measurand's distribution. Operationally the Bayes approach to compatibility testing coincides with the random-effects model [9]. In this context $\sigma_1^2 + \sigma_2^2$ plays the reproducibility error role while σ_1^2 represents the repeatability error [10].

The next section discusses a procedure motivated by the classical divergence (asymmetric 'distance') between probability distributions. This characteristic and the following null hypothesis involve the uncertainties ratio, σ_2/σ_1.

5. Information divergence and measurement capability index

Information theory and probability theory developed several concepts of divergence (or separation) between two probability distributions P and Q (or two densities p and q). The best known is the (Kullback–Leibler) *information* number,

$$K(Q, P) = E^Q \log\left[\frac{\mathrm{d}Q}{\mathrm{d}P}(X)\right] = \int \log\left[\frac{q}{p}(x)\right] q(x)\,\mathrm{d}x.$$

In general, $K(Q, P) \neq K(P, Q)$. However $K(Q, P) \geqslant 0$, and equality holds if and only if $Q = P$. This divergence plays an established important role in statistics, in particular, for testing goodness of fit, where P (theoretical model) and Q (empirical distribution) are not supposed to be exchangeable. The confidence regions covering two normal parameters are also based on the Kullback–Leibler number [11]. The widely used entropy of random variable X with density $q(x)$ is $-E^Q \log q(X) = -\int [\log q(x)] q(x)\,\mathrm{d}x$.

If $P = N(\mu_1, \sigma_1^2)$ and $Q = N(\mu_2, \sigma_2^2)$ are two Gaussian distributions,

$$K(Q, P) = \frac{1}{2}\left[\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2} + \log\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} - 1\right]. \tag{5}$$

In the setting of section 2, assume that the parameters μ_2, σ_2^2 are estimated on the basis of lab's data by their maximum likelihood estimators $\bar{x}$ and $\nu u^2/n$, while μ_1, σ_1 are treated as given. Then the estimated version of the information number is

$$\hat{K}(Q, P) = \frac{1}{2}\left[\frac{(\bar{x} - \mu_1)^2}{\sigma_1^2} + \log\frac{n\sigma_1^2}{\nu u^2} + \frac{\nu u^2}{n\sigma_1^2} - 1\right]. \tag{6}$$

When testing the null hypothesis: $\mu_2 = \mu_1, \sigma_2^2/n = \sigma_1^2$, (6) is closely related to the likelihood ratio test statistic, which has approximate χ^2-distribution with 2 degrees of freedom [7, theorem 12.4.2]. This test rejects when

$$\frac{(\bar{x} - \mu_1)^2}{n\sigma_1^2} + \log\frac{n\sigma_1^2}{\nu u^2} + \frac{\nu u^2}{n\sigma_1^2} - 1 \geqslant \frac{\chi^2_\alpha(2)}{n}.$$

Here and below $\chi^2_\alpha(\nu)$ denotes the $(1-\alpha)$th percentile of χ^2-distribution with ν degrees of freedom.

The boundary of this rejection region is depicted by the dotted line in figure 1 when $n = 4$, $\alpha = 0.05$, $\mu_1 = 0$, $\sigma_1 = 1$. The acceptance zone which can be interpreted as a confidence region for two normal parameters μ_1, σ_1^2, is centred at $\bar{x} = \mu_1, u = \sqrt{n/\nu}$. It looks nearly elliptic (although it is not symmetric about the line $u = \sqrt{n/\nu}$). In quality control an approximation of this region by a (rescaled) circle is used [12].

The main difficulty is that the hypotheses like the equality of the means: $\mu_2 = \mu_1$, or the equality of distributions: $Q = P$, hardly address the right question when testing compliance. This author believes that a relevant version of the compatibility hypothesis is $\mu_2 = \mu_1$ and the unknown σ_2 (or rather $\sigma_2/\sqrt{n}$) is not considerably larger than σ_1. When assessing compliance with any standard, very large values of σ_2/σ_1 are not acceptable as they lead to poor performance of virtually all available procedures. For large u tests (1), (2), (3) or (4) cannot reject the null compatibility hypothesis no matter how far apart are $\bar{x}$ and μ_1. For example, as is well known to metrologists, by claiming a large uncertainty u, lab 2 can always accomplish overlap of its interval with lab 1's interval.

Therefore the assertion, $\mu_2 = \mu_1$, has little meaning unless it is accompanied by a restriction $\sigma_2 \leqslant B\sigma_1$. Here B $(B \geqslant 1)$ is the maximum allowable upper bound on the relative uncertainties of two labs' measurements. This bound can be defined through the *measurement capability index* which is defined here as the ratio of expected widths of two coverage

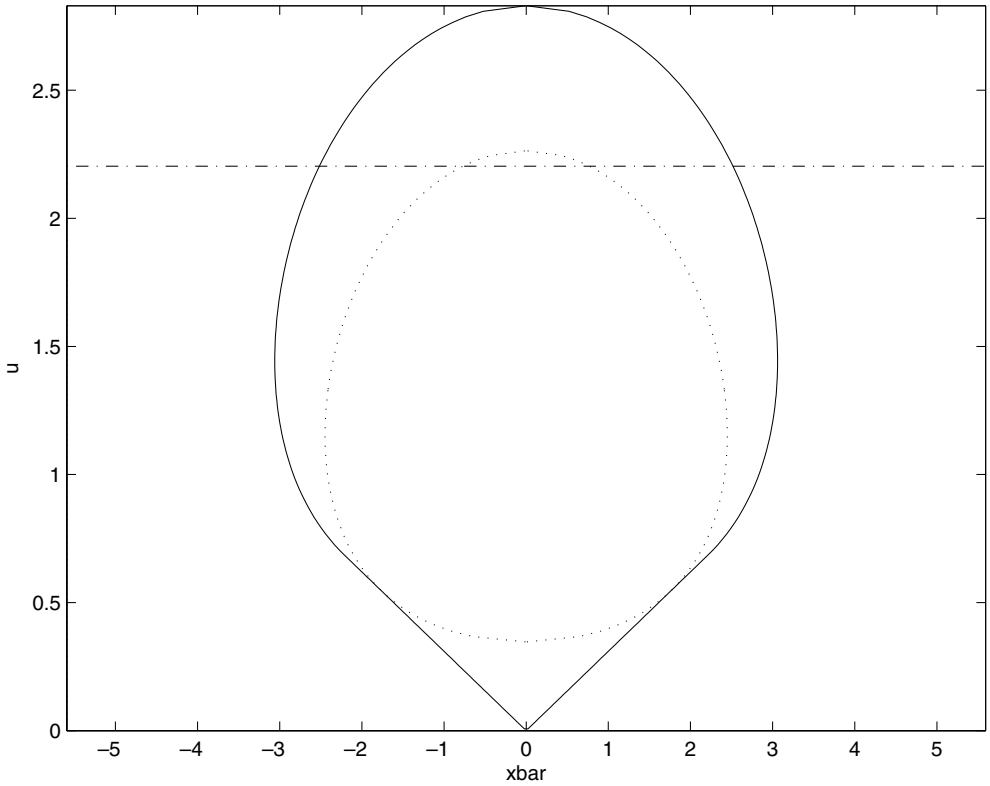

Figure 1. Boundaries of the acceptance regions for the likelihood ratio test of the equality of normal distributions (the dotted line) and of the null hypothesis: H_0: $\mu_2 = \mu_1, \sigma_2 \leqslant B\sigma_1$ when $B = 2.5$, $n = 4$, $\alpha = 0.05$, $\mu_1 = 0$, $\sigma_1 = 1$. The Mood's test boundary is shown as the dashed–dotted line.

Table 1. CRM intervals for SRM 1944 in $\mu g\,kg^{-1}$ units and the test results.

Analyte	μ_1	$2\sigma_1$	(1)	(2)	(3)	(4)	(7)	(8)
Phenanthrene	5270	220	9	5	5	8	10	6
Fluoranthene	8920	320	6	3	3	5	10	7
Pyrene	9700	420	7	4	4	6	12	10
Benz*[a]*anthracene	4720	110	8	6	6	5	12	11
Benzo*[ghi]*perylene	2840	100	3	1	1	4	5	6
PCB 52	79.4	2.0	9	6	6	9	9	8
PCB 118	58.0	4.3	13	5	4	7	13	5
PCB 153	74.0	2.9	11	6	7	8	10	8
PCB 180	44.3	1.2	9	4	5	8	9	8
PCB 209	6.81	0.33	6	3	3	6	6	5
Hexachlorobenzene	6.03	0.35	5	3	3	6	7	6
cis-chlordane	16.51	0.83	8	8	8	8	10	9
trans-nonachlor	8.20	0.51	8	7	7	6	11	10
4,4′-DDT	119	11	6	4	4	5	8	6

intervals, $C_m = 2\sqrt{n}\sigma_1/[t_{\alpha/2}(\nu)\sigma_2]$. Indeed, for the desired measurement capability index C_m, $B = 2\sqrt{n}C_m^{-1}/t_{\alpha/2}$. See [13] for the discussion of other capability characteristics in quality control problems where the traditional definition of the index is $(T_U - T_L)/(6\sigma_2)$ with the recommended value exceeding 1.5.

In some cases, B can be ascertained on the basis of the physical/chemical nature of the measurand, on validation data, on previous repeatability and reproducibility studies, or possibly on proficiency level of a particular lab. When testing compliance with legal or regulatory specification limits, fairly large values for C_m are anticipated. Czaske [14] indicates that in this situation 'the usual values of C_m are between 2 and 5'. However, when testing conformity with the CRM, one may encounter smaller values of this characteristic.

Thus we formulate the following null hypothesis as the conformance testing problem, H_0: $\mu_2 = \mu_1, \sigma_2 \leqslant B\sigma_1$ for a given B. An implication of this hypothesis is that even identical values μ_2 and μ_1 are not acceptable if σ_2 is considerably larger than σ_1. Figure 1 shows the boundary of the acceptance region for the likelihood ratio test of H_0 when $B = 2.5, n = 4$, $\alpha = 0.05, \mu_1 = 0, \sigma_1 = 1$.

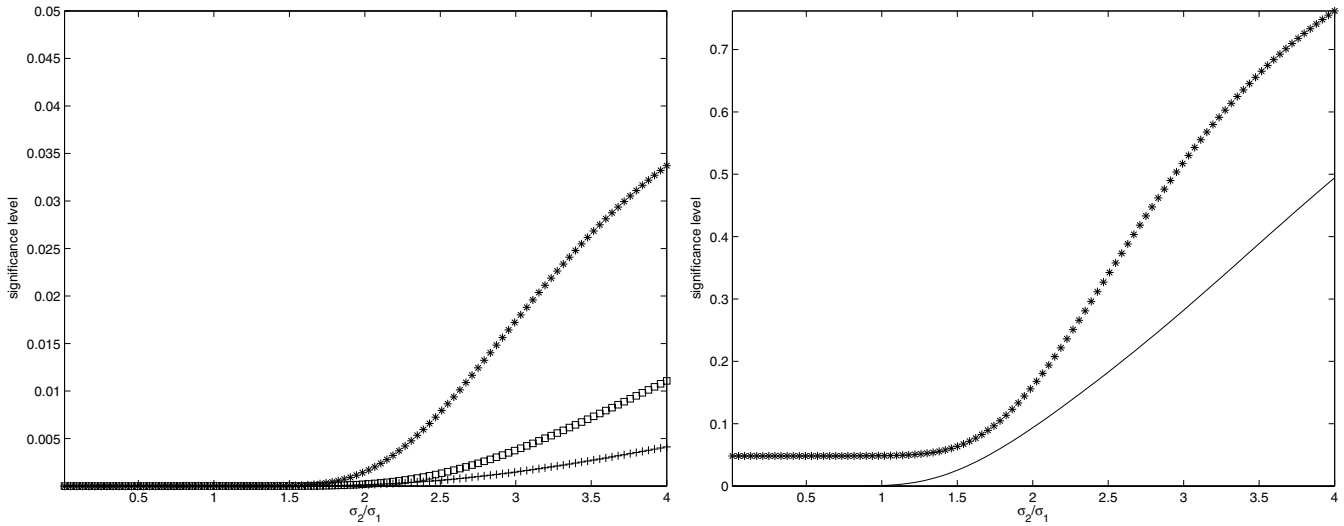

Figure 2. The significance level of test (1) (solid line on the top), of (2) (line marked by squares), of (4) (line marked by $*$), and of test (3) (line marked by +) (left panel) when $n = 5$, $\alpha = 0.05$, $\mu_2 = \mu_1$, $\sigma_2/\sigma_1 \leqslant 4$. The same characteristic is shown for test (7) (line marked by $*$), and for test (8) (solid line) in the right panel.

If the unknown variance of $\bar{x}$, σ_2^2/n, is estimated by $\nu u^2/n$, H_0 will be rejected when

$$\min_{0<b\leqslant B}\left[\frac{(\bar{x}-\mu_1)^2+\nu u^2}{b^2\sigma_1^2}+\log\left(\frac{b\sigma_1}{\sqrt{\nu}u}\right)^2-1\right]\geqslant\frac{\chi_\alpha^2(2)}{n}.$$

This fact can be seen from the general form of the rejection region for a null hypothesis which is a union of sub-hypotheses (the so-called intersection-union test) [15]. This rejection region can be written as

$$\frac{(\bar{x}-\mu_1)^2}{\nu u^2}\geqslant \mathrm{e}^{\chi_\alpha^2(2)/n}-1, \qquad \text{if } (\bar{x}-\mu_1)^2+\nu u^2\leqslant B^2\sigma_1^2, \tag{7}$$

$$\frac{(\bar{x}-\mu_1)^2+\nu u^2}{B^2\sigma_1^2}+\log\left(\frac{B\sigma_1}{\sqrt{\nu}u}\right)^2-1\geqslant\frac{\chi_\alpha^2(2)}{n},$$
$$\text{if } (\bar{x}-\mu_1)^2+\nu u^2> B^2\sigma_1^2.$$

The acceptance region of (7) is a balloon-shaped figure in figure 1 which contains the acceptance region for the likelihood ratio test of the hypothesis of equality of normal distributions. Small u's are not included in this region if $(\bar{x}-\mu_1)^2/\sigma_1^2$ is large.

An alternative procedure, which is an analogue of Mood's test [11], accepts the null hypothesis H_0 for the values of $\bar{x}$ and u such that $(\bar{x}-\mu_1)^2/\sigma_1^2\leqslant B^2\chi_{\alpha_1}^2(1)/n$ and $\nu u^2/\sigma_1^2\leqslant B^2\chi_{\alpha_2}^2(\nu)$, with $(1-\alpha_1)(1-\alpha_2)=1-\alpha$. Thus, its rejection region is

$$\frac{(\bar{x}-\mu_1)^2}{\sigma_1^2}\geqslant\frac{B^2\chi_{\alpha_1}^2(1)}{n} \qquad \text{or} \qquad \frac{\nu u^2}{\sigma_1^2}\geqslant B^2\chi_{\alpha_2}^2(\nu). \tag{8}$$

In figure 1 the acceptance region of Mood's test is a rectangle whose side lines and the lower boundary coincide with the plot boundary, and whose upper boundary is the dashed–dotted straight line. We took $\alpha_1=\alpha_2=\sqrt{0.95}$.

6. Non-conformity testing and interchangeability characteristics

In the notation of section 2, the compliance probability is defined as

$$\begin{aligned}P_c &= P\left(|X-\mu_1|<z_{\alpha/2}\sigma_1\right)\\ &= \Phi\left(\frac{\mu_1-\mu_2+z_{\alpha/2}\sigma_1}{\sigma_2}\right)-\Phi\left(\frac{\mu_1-\mu_2-z_{\alpha/2}\sigma_1}{\sigma_2}\right)\end{aligned} \tag{9}$$

with X representing the random normal measurement of the lab 2, $X\sim N(\mu_2,\sigma_2^2)$. As was alluded to earlier, this probability is quite small when $\sigma_2>\sigma_1$. Perhaps for this reason, the ASME *Guidelines* [4] make reference to a common decision rule for industrial application, the so-called $N:1$ rule. This procedure mandates that the user interval cannot exceed $1/N$ of the specification zone. This procedure corresponds to $C_{\mathrm{m}}\approx N$, and N is typically taken to be 3 or 4, which may require a fairly substantial number n of repeats.

Motivated by the fact that (9) takes its largest values when $|\mu_1-\mu_2|/\sigma_2\leqslant c_0$, Wang and Iyer [16] suggested to test the validity of the claim, $|\mu_1-\mu_2|/\sigma_2\geqslant c_0$, which is a non-conformance hypothesis. Designating non-conformity as a null hypothesis is commonly suggested when testing bioequivalence where a generic drug (which has to establish itself) must be compared with the brand name drug. A similar non-compliance null hypothesis with σ_2 replaced by σ_1 has been advocated in metrology as well [6].

The main argument in favour of non-conformance as the null hypothesis is that a statistical test can offer evidence only against it. Indeed, in a hypothesis testing situation, large p-values do not necessarily affirm the validity of the null hypothesis but merely the lack of evidence to the contrary. It is also believed that once the null hypothesis is rejected, one cannot alter this declaration. See [6] for a further discussion and [17] for a review of available statistical techniques.

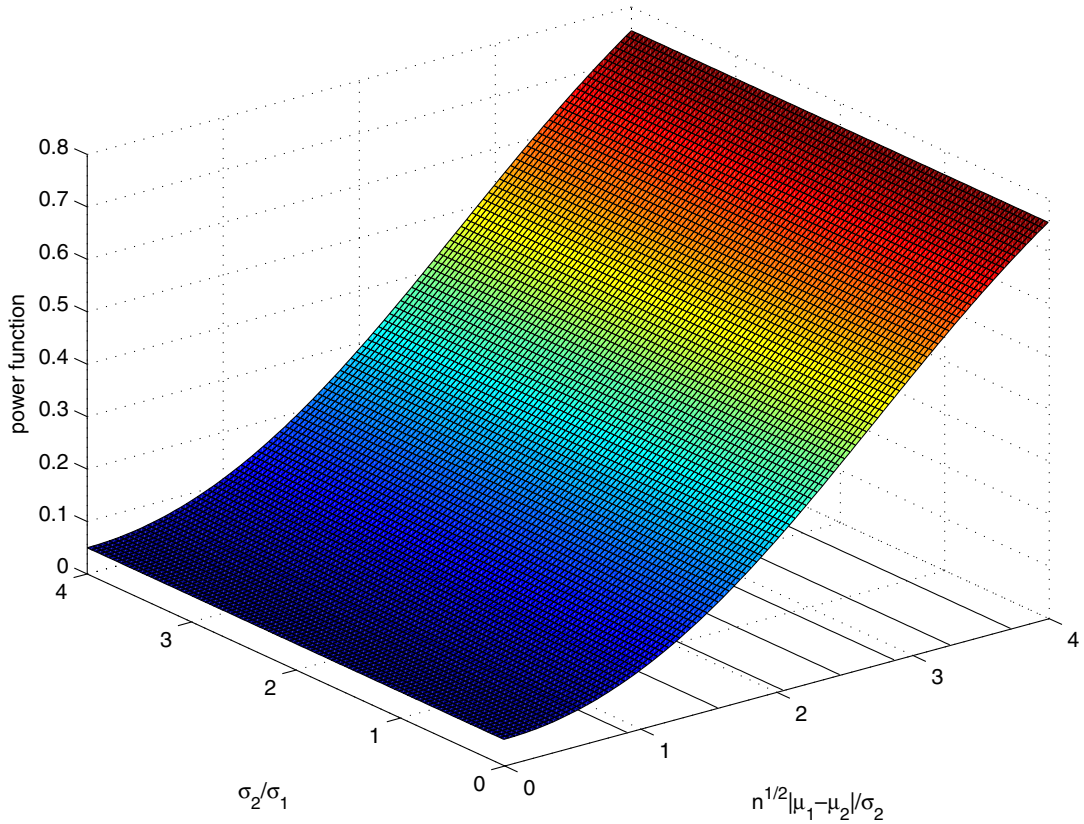

Figure 3. The power function of test (1).

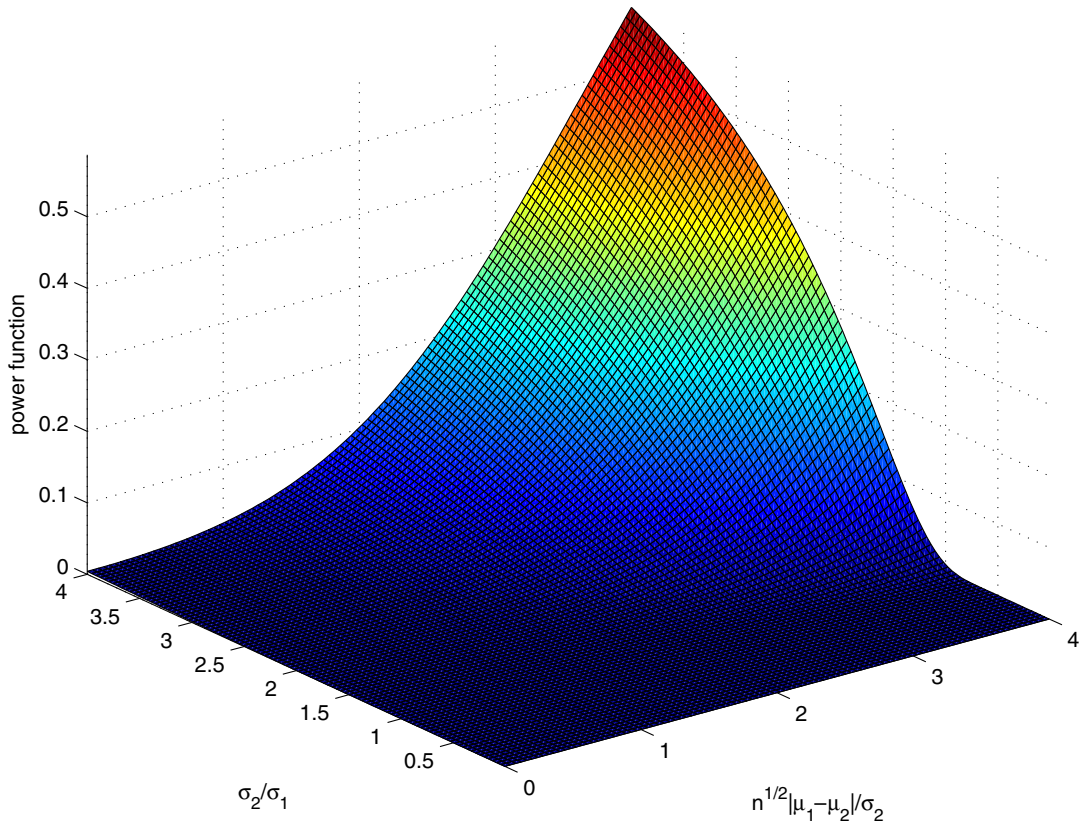

Figure 4. The power function of test (2).

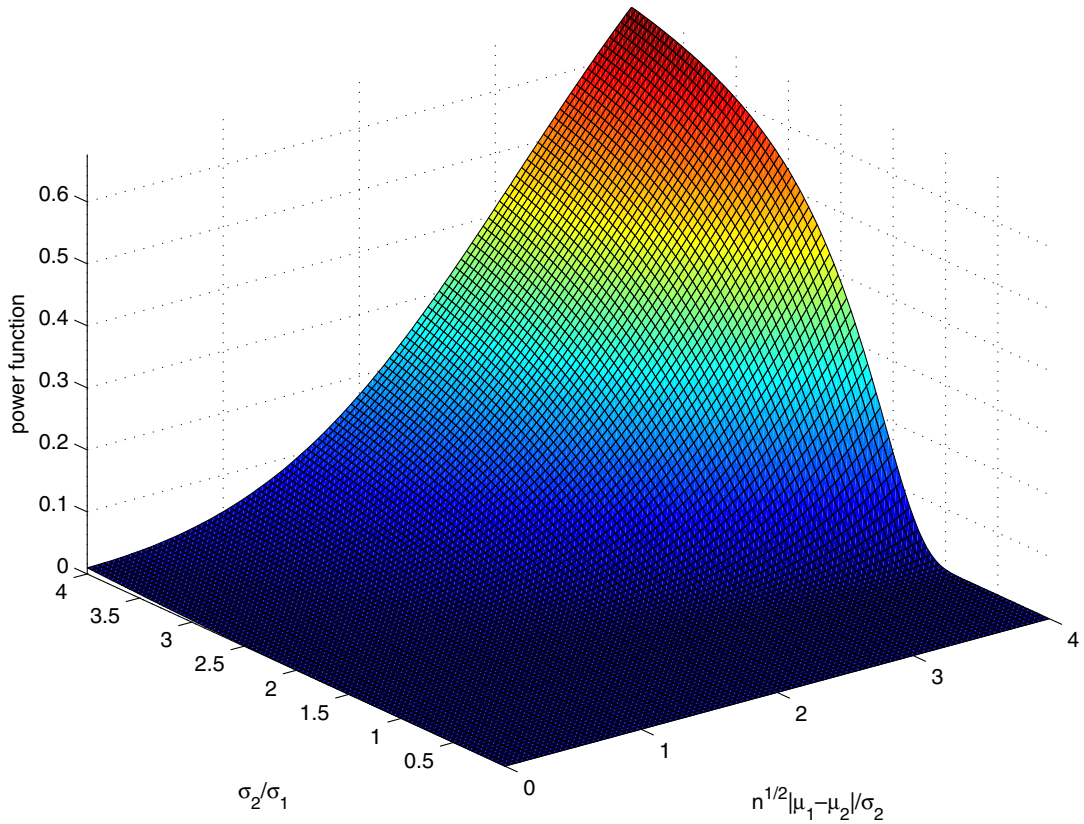

Figure 5. The power function of test (3).

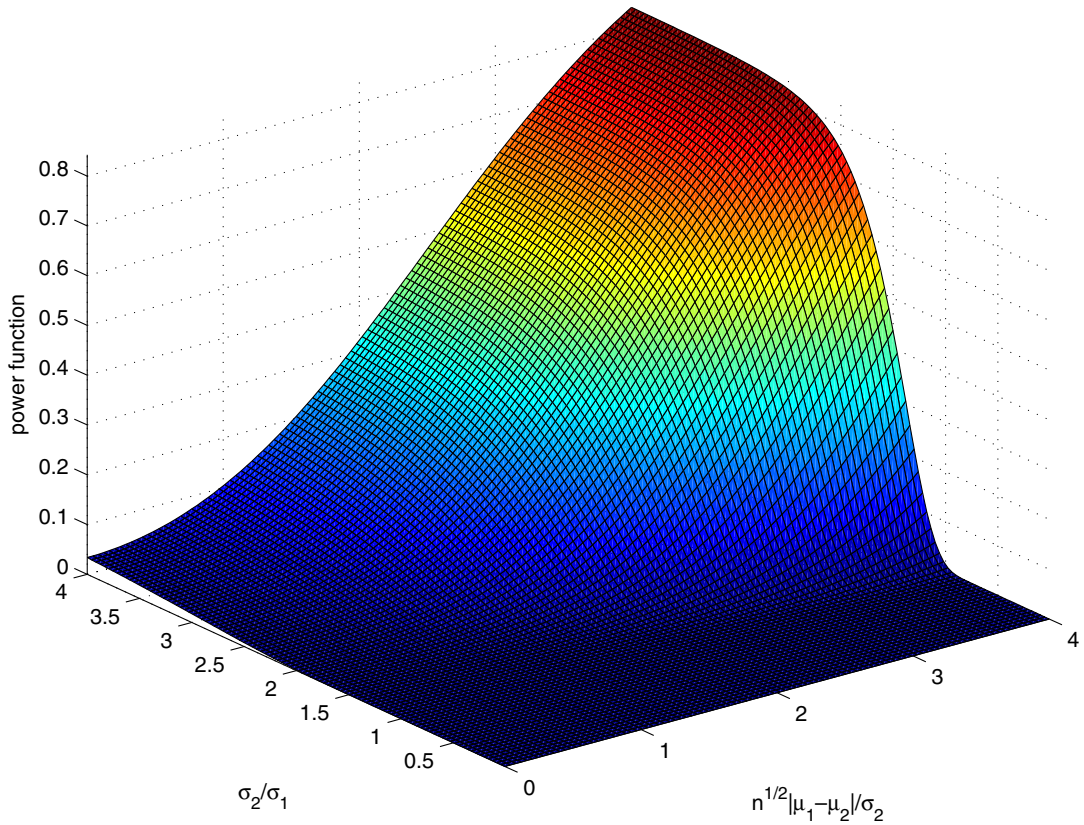

Figure 6. The power function of test (4).

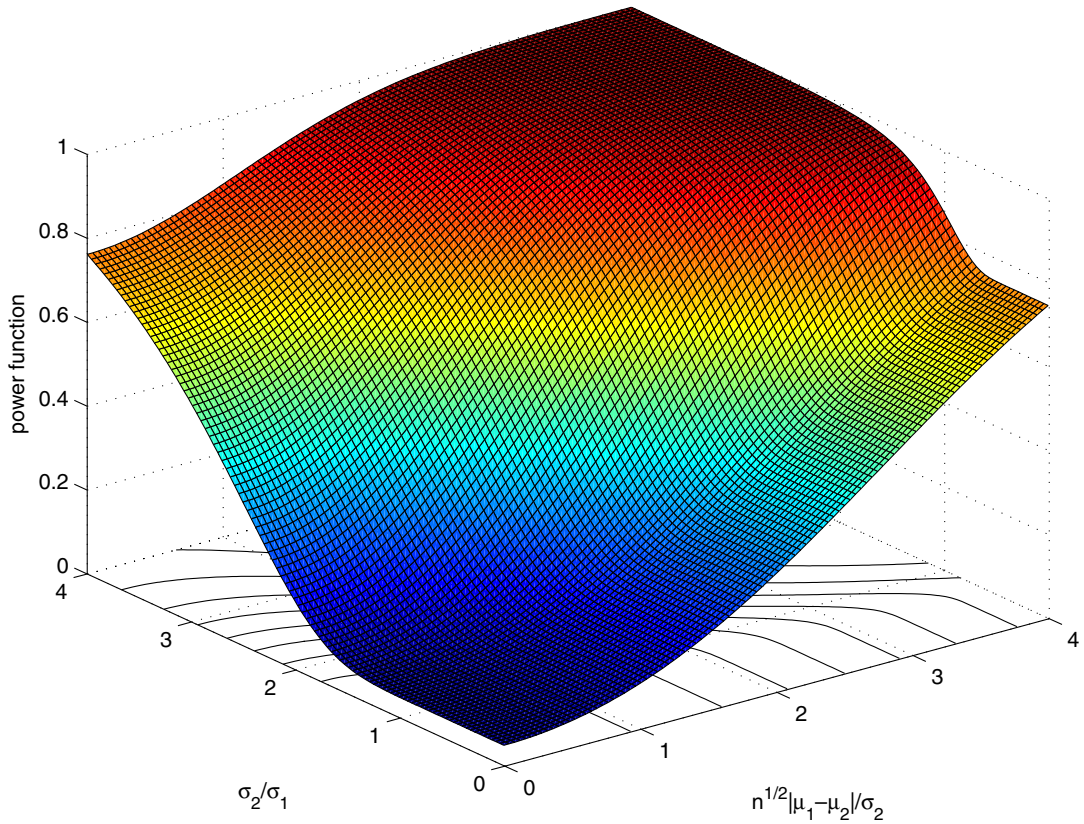

Figure 7. The power function of test (7) with $B = 1.5$.

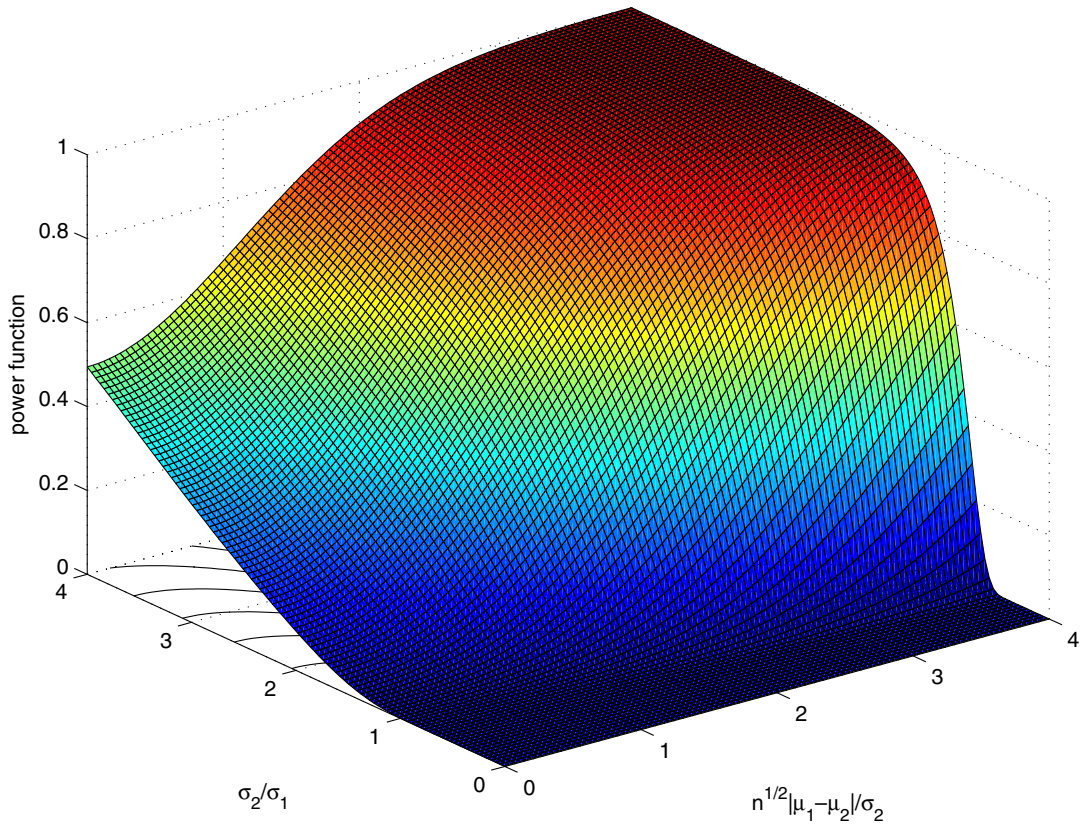

Figure 8. The power function of the Mood test (8) with $B = 1.5$.

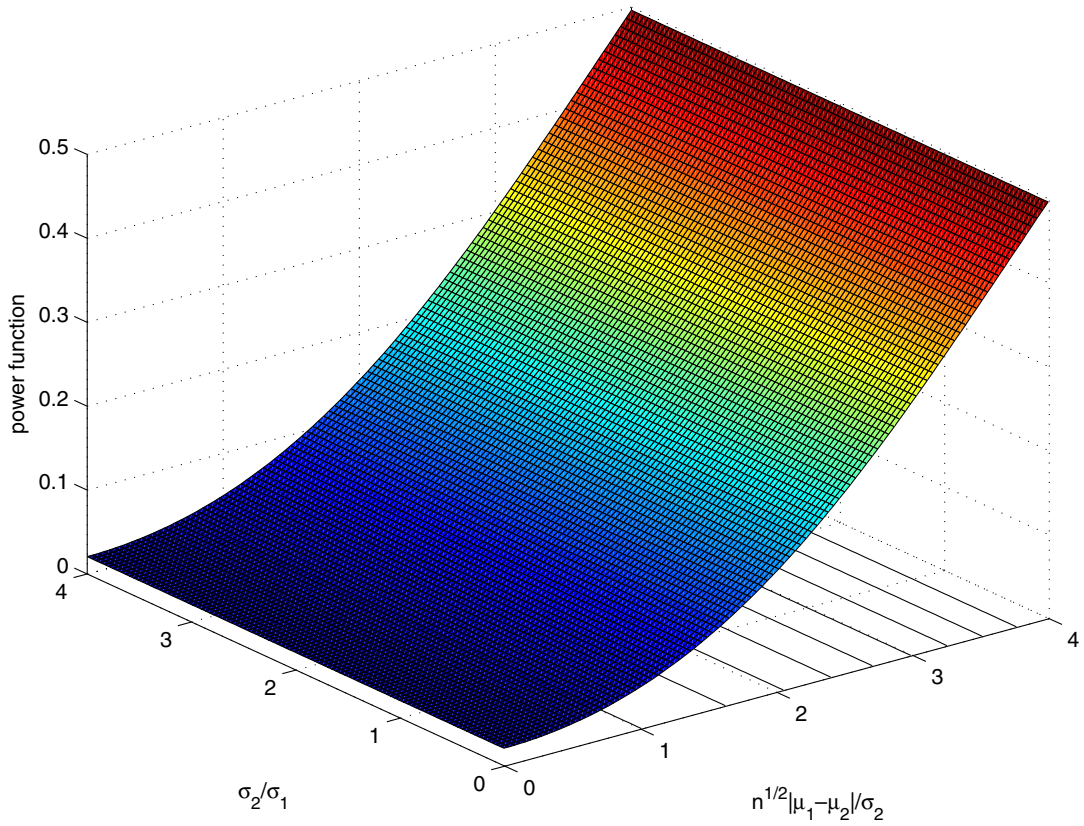

Figure 9. The power function of test (10) with $c_0 = 2$.

A mathematical difficulty related to testing of null non-conformity hypotheses is that the Type I error is to be controlled over a large subset of the parameter space. Indeed, by definition, the alternative then is a small subset of this space. The complicated form of test statistics (if such are available) makes evaluation of p-values much more difficult.

Because of the presumption of innocence principle, such a null hypothesis cannot be used in legal metrology. Arguably, asserting incompatibility to be a null hypothesis is less natural in other metrology applications where both labs are supposed to measure the same measurand. Instead of non-conformity hypothesis, one can put forward the conformance null hypotheses, $\sqrt{n}|\mu_1 - \mu_2|/\sigma_2 \leqslant c_0$ or $|\mu_1 - \mu_2|/\sigma_1 \leqslant c_1$. The first of these hypotheses is rejected when

$$\frac{(\bar{x} - \mu_1)^2}{u^2} \geqslant t_\alpha^2(\nu, c_0, \nu), \tag{10}$$

where $t = t_\alpha(\nu, c_0, \nu)$ is the solution of the equation

$$1 - nctcdf(t, c_0, \nu) + nctcdf(-t, c_0, \nu) = \alpha.$$

Here $nctcdf(t, c_0, \nu)$ denotes the distribution function of the non-central t-distribution with the non-centrality parameter c_0 and ν degrees of freedom [7, section 6.4]. The shape of this region is similar to that in (1), but the threshold constant is larger, so this test rejects less frequently than (1).

The rejection regions of the second hypothesis, $(\bar{x} - \mu_1)^2 \geqslant g_1 u^2$ or $(\bar{x} - \mu_1)^2 \geqslant g_2 \sigma_1^2$, cannot have a guaranteed type I error. None of these two hypotheses uses all parameters, and they may be less attractive than our H_0 described in section 5 while sharing with it the task of specifying the needed constant (c_0 or c_1).

With X_1 denoting the measurand of lab 1, Willink [18] suggested to use $E(X - X_1)^2$ as a symmetric measure of compatibility of labs 1 and 2. Wang and Iyer [16] addressed the same issue and proposed a directional (asymmetric) interchangeability characteristic A_{12} of two labs as Willink's coefficient normalized by $2\sigma_1^2$. When these variables have normal distributions,

$$A_{12} = \frac{1}{2}\left[\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2} + \frac{\sigma_2^2}{\sigma_1^2} + 1\right], \tag{11}$$

which looks similar to (5) except that the important log-ratio term is missing. Because of that, the smallest value of A_{12}, which is equal to 0.5, is attained when $\sigma_1^2 \to \infty$, no matter how different μ_1 and μ_2 are. Thus, one can argue that the classical by now Kullback–Leibler information number better serves the purpose of assessing asymmetric interchangeability in the context of interlaboratory comparisons. This statement is supported by the examples considered in the next section.

7. Examples

The environmental standard reference material (SRM), SRM 1944 New York/New Jersey Waterway Sediment [19] was

issued by the National Institute of Standards and Technology (NIST) in 1995 as a fresh frozen homogenate to meet the needs of laboratories analysing bivalve tissues. The focus of the study was to assess the interlaboratory and temporal comparability of data and to improve methods for the monitoring of organic contaminants. It was characterized at NIST using multiple analytical methods.

The data for this SRM came from 16 laboratories participating in a performance-based study over a period of several years. The SRM interval for 14 selected compounds is given by the first two columns of table 1. The test data (number of rejections) for tests (1), (2), (3), (4), (7) and (8) when $\alpha = 0.05$ are provided by the six right-hand columns of table 1. Overall out of 215 data points, there are 65 instances of non-overlapping CRM and labs intervals (which is the total of the column corresponding to (2)). As an example, lab 15 reported in the PCB 118 a value of $50.23\,\mu g\,kg^{-1}$ with the standard deviation $u = 1.39\,\mu g\,kg^{-1}$. The corresponding CRM interval is $58.0\,\mu g\,kg^{-1} \pm 4.3\,\mu g\,kg^{-1}$, and the intervals do not overlap.

Test (1) rejected 108 times, test (4) 93 times, while the results of (3) were quite similar to those of (2) with 66 rejections. Test (7) is more stringent rejecting 132 times, while (8) rejected 105 times ($B = 4$).

Some of these rejections seem to be due to clerical or registration errors, but most demonstrate the reported uncertainties which are unrealistically small and/or sample means which are too far from the certificate value. For example, in the case of *cis*-chlordane whose SRM interval is 16.51 ± 0.83 in $\mu g\,kg^{-1}$ units, lab 9 states an unrealistically small standard deviation of 0.0577; lab 14's reported value $\bar{x} = 1.38$ probably should have been 13.8.

As an example of interchangeability index evaluation, consider lab 1 reporting in the fluoranthene $\bar{x} = 8924$, $u = 348$ (in $\mu g\,kg^{-1}$ units). Then all conformance tests considered so far accept the equality of means (and even the equality of uncertainties), with $\hat{K}(Q, P) = 0.0075$. However, the estimated interchangeability index, $A_{12} = 1.0914$ looks to be too large, especially when compared with smaller values of (11) for other analytes. As another example, lab 7 in the PCB 153 reported a value $\bar{x} = 73.433\,\mu g\,kg^{-1}$ with standard deviation $u = 5.312\,\mu g\,kg^{-1}$. Then $\hat{K}(Q, P) = 0.5917$, but $A_{12} = 1.917$ is more than three times larger.

8. Power comparisons: numerical results

The invariance property shows that the power of all considered tests (i.e. the probability to reject the null hypothesis) is a function of $|\mu_2 - \mu_1|/\sigma_2$ and σ_2/σ_1. For a good test, the power function is about equal to α, the type I error (significance level), when the null hypothesis holds, and this function assumes large (close to one) values on the alternative.

Figure 2 shows the plot of significance level (expected rejection rate) of tests (1), (2), (3) and (4) when $\mu_2 = \mu_1$, i.e. when their null hypothesis is true. Clearly (1) has its type I error, chosen to be 0.05, exactly at this level, but other tests are too conservative with the rejection rate well below 0.05. A similar plot for tests (7) and (8) demonstrates superiority of (7) in this regard.

Table 2. The maximum power of tests (1), (2), (3), (4), (7), (8) and (10) when $n = 5$, $\alpha = 0.05$ in the region $\sqrt{n}|\mu_2 - \mu_1|/\sigma_2 \leqslant 4$, $\sigma_2/\sigma_1 \leqslant 4$.

(1)	(2)	(3)	(4)	(7)	(8)	(10)
0.8443	0.5866	0.7685	0.8428	0.9999	0.9997	0.5972

Figures 3–9 depict the plots of power functions for the tests (1), (2), (3), (4), (7), (8) and (10) (along with their contours when seen) for $\alpha = 0.05$, $\nu = 4$ in the region $\sqrt{n}|\mu_2 - \mu_1|/\sigma_2 \leqslant 4$, $\sigma_2/\sigma_1 \leqslant 4$. These functions were evaluated via numerical integration involving the standard normal distribution and $\chi^2(\nu)$ distribution. Actually, calculations for (1), (8) and (10) do not even require numerical integration as the power functions of these tests can be expressed through the classical distribution functions (including the non-central t-distribution).

Clearly, tests (1) and (4) outperform (2) and (3) (with $f = t^2_{\alpha/2}(\nu)$) but all of these tests have insufficient power if the ratio σ_2/σ_1 is small even when $|\mu_2 - \mu_1|$ is large. Their power is close to 0.05 when σ_2/σ_1 is large while a perfect test should have it close to one. These facts alone are a good reason to employ test (7) which demonstrates a fairly good performance helped by the fact that large values of σ_2/σ_1 now belong to the alternative. The Mood's test is competitive, but (10) does not gain enough power as $|\mu_2 - \mu_1|/\sigma_2$ increases beyond c_0.

The situation does not change for the better when the null hypothesis is that of non-conformity as the power function remains small on the alternative. The power function of the test of $|\mu_1 - \mu_2|/\sigma_2 \geqslant c_0 = 2.35$ according to [16] does not exceed 0.16 even when $\nu = \infty$.

The maximum power of the considered tests is summarized in table 2.

9. Conclusions

The exact mathematical formulation of a conformity hypothesis may not be so important in metrology applications. However, without such a formulation it is impossible to evaluate the test performance or to compare two different tests.

Table 3 presents a summary of available compatibility testing methods. This table along with the results of section 8 amply demonstrate the difficulties of testing the classical 'equality of means' hypotheses. Out of these, tests (1) and (4) have the best power function and test (2) based on the intersecting intervals is the worst. It should not be used by metrologists indeed.

There is no test whose power would depend only on $|\mu_2 - \mu_1|/\sigma_2$. Equally absent are tests whose significance level is equal to α when $\mu_2 = \mu_1$, and whose rejection region has the form $|\bar{x} - \mu_1|/\sqrt{\sigma_1^2 + u^2} \geqslant g(u/\sigma_1)$ with a smooth function g [20]. The power of all tests of the

Table 3. Summary of existing compatibility testing procedures.

Hypothesis	Rejection region	Advantages	Disadvantages
$\mu_2 = \mu_1$ for some σ_2	$\lvert\bar{x} - \mu_1\rvert > t_{\alpha/2}(\nu)u$	Simplicity	σ_1 not used poor power for large σ_2/σ_1
$\mu_2 = \mu_1$ for some σ_1, σ_2	$\lvert\bar{x} - \mu_1\rvert > 2\sigma_1 + t_{\alpha/2}(\nu)u$	Geometric appeal	Poor power
Behrens–Fisher problem	$\lvert\bar{x} - \mu_1\rvert > \sqrt{f(u^2 + \sigma_1^2)}$	Classics	Labs exchangeable f non-unique
Bayes setting $\mu_2 = \mu_1, \sigma_2 \leqslant B\sigma_1$ $\lvert\mu_2 - \mu_1\rvert \leqslant c_0\sigma_2$	$\lvert\bar{x} - \mu_1\rvert > t_{\alpha/2}(\nu)\sqrt{\max(u^2, \sigma_1^2)}$ (7) or (8) See [16]	Better uncertainties Good power	Poor power for large σ_2/σ_1 B assessment c_0 difficult

hypothesis $\mu_2 = \mu_1$ is close to the significance level α when σ_2/σ_1 is large while a perfect test should have it close to one.

The logical implication is to remove the troublesome values of σ_2/σ_1 from the null hypothesis and to formulate a compatibility hypothesis as the joint statement H_0 suggested here with the rejection region in (7). The paper argues that this null hypothesis H_0: $\mu_2 = \mu_1$, $\sigma_2 \leqslant B\sigma_1$ is a reasonable statistical expression for testing compliance, conformity or compatibility. The presence of the constant B (which can be determined from a desired measurement capability index) is an advantage inasmuch as it allows greater flexibility in applications.

Acknowledgments

The author thanks Dr Michelle Schantz for the complete data set in section 7. Many helpful comments by Dr David Duewer and by the referee are also acknowledged.

References

[1] Desimoni E and Brunetti B 2011 Uncertainty of measurement and conformity assessment: a review *Anal. Bioanal. Chem.* **400** 1729–41
[2] ISO 2003 *ISO 10576-1: Statistical methods—Guidelines for the evaluation of conformity with specified requirements, Part 1: General principles* (Geneva, Switzerland: International Organization for Standardization (ISO)/International Electrotechnical Commission)
[3] Ellison S L R and Williams A (ed) 2007 *EURACHEM/CITAC Guide 'Use of Uncertainty Information in Compliance Assessment'* http://www.eurachem.org
[4] ASME B89.7.3.1-2001 2002 *Guidelines for Decision Rules: Considering Measurement Uncertainty in Determining Conformance to Specifications* (New York: ASME)
[5] Schenker N and Gentleman J 2001 On judging the significance of differences by examining the overlap between confidence intervals *Am. Statist.* **55** 182–6
[6] Holst E, Thyregod P and Willrich P 2001 On conformity testing and the use of two stage procedures *Int. Stat. Rev.* **69** 419–32
[7] Lehmann E and Romano J P 2003 *Testing Statistical Hypotheses* 3rd edn (Springer: New York)
[8] Ballico M 2000 Limitations of the Welch–Satterthwaite approximation for measurement uncertainty calculations *Metrologia* **37** 61–4
[9] Searle S, Casella G and McCulloch C 1992 *Variance Components* (New York: Wiley)
[10] Rukhin A L 2012 Estimating heterogeneity variance in meta-analysis studies *J. R. Statist. Soc.* B doi:10.1111/j.1467-9868.2012.01047.x
[11] Arnold B C and Shavelle R M 1998 Joint confidence sets for the mean and variance of a normal distribution *Am. Statist.* **52** 133–40
[12] Van Nuland Y 1992 ISO 9002 and the circle technique *Quality Eng.* **5** 269–91
[13] Bothe D R 1997 *Measuring Process Capability: Techniques and Calculations for Manufacturing Engineers* (New York: McGraw-Hill)
[14] Czaske M 2008 Usage of the uncertainty of measurement by accredited calibration laboratories when stating compliance *Accred. Qual. Assur.* **13** 645–51
[15] Casella G and Berger R 2002 *Statistical Inference* 2nd edn (Belmont, CA: Duxbury Press)
[16] Wang C M and Iyer H K 2010 On interchangeability of two laboratories *Metrologia* **47** 435–47
[17] Wellek S 2010 *Testing Statistical Hypotheses of Equivalence and Noninferiority* 2nd edn (Boca Raton, FL: CRC Press)
[18] Willink R 2003 On the interpretation and analysis of a degree-of-equivalence *Metrologia* **40** 9–17
[19] Wise S *et al* 2004 Two new marine sediment standard reference materials (SRMs) for the determination of organic contaminants *Anal. Bioanal. Chem* **378** 1251–64
[20] Linnik Yu V 1967 *Statistical Problems with Nuisance Parameters* (Providence, RI: American Mathematical Society)