Metrologia 45 (2008) 257–264 doi:10.1088/0026-1394/45/3/001

Classical and Bayesian interpretation of the Birge test of consistency and its generalized version for correlated results from interlaboratory evaluations

Raghu N Kacker[1], **Alistair Forbes**[2], **Rüdiger Kessel**[1] **and Klaus-Dieter Sommer**[3]

[1] National Institute of Standards and Technology, Gaithersburg, MD 20899-8910, USA
[2] National Physical Laboratories, Teddington, Middlesex, TW11 0LW, UK
[3] Physikalisch-Technische Bundesanstalt, D-38116 Braunschweig, Germany

E-mail: raghu.kacker@nist.gov

Received 31 August 2007
Published 10 April 2008
Online at stacks.iop.org/Met/45/257

Abstract

A well-known test of consistency in the results from an interlaboratory evaluation is the Birge test, named after its developer Raymond T Birge, a physicist. We show that the Birge test of consistency may be interpreted as a classical test of the null hypothesis that the variances of the results are less than or equal to their stated values against the alternative hypothesis that the variances of the results are greater than their stated values. A modern protocol for hypothesis testing is to calculate the classical p-value of the test statistic. The$p$-value is the maximum probability under the null hypothesis of realizing in conceptual replications a value of the test statistic equal to or larger than the realized (observed) value of the test statistic. The null hypothesis is rejected when the p-value is too small. We show that, interestingly, the classical p-value of the Birge test statistic is equal to the Bayesian posterior probability of the null hypothesis based on suitably chosen non-informative improper prior distributions for the unknown statistical parameters. Thus the Birge test may be interpreted also as a Bayesian test of the null hypothesis. The Birge test of consistency was developed for those interlaboratory evaluations where the results are uncorrelated. We present a general test of consistency for both correlated and uncorrelated results. Then we show that the classical p-value of the general test statistic is equal to the Bayesian posterior probability of the null hypothesis based on non-informative prior distributions. The general test makes it possible to check the consistency of correlated results from interlaboratory evaluations. The Birge test is a special case of the general test.

1. Introduction

Suppose n laboratories submit the following paired results of measurement and standard uncertainties $[x_1, u(x_1)], \ldots, [x_n, u(x_n)]$ for a common reference. Suppose the sampling probability distributions of the results $x_1, \ldots, x_n$ are mutually independent. Seventy-five years ago a physicist named Raymond T Birge [1] proposed that to check for consistency in the results $x_1, \ldots, x_n$ relative to the stated standard uncertainties $u(x_1), \ldots, u(x_n)$, calculate the following test statistic[4] from the realized data $[x_1, u(x_1)], \ldots, [x_n, u(x_n)]$:

$$R^2 = \sum_{i=1}^{n} w_i (x_i - x_W)^2/(n-1), \tag{1}$$

[4] In the Birge test statistic (1), $x_1, \ldots, x_n$ are random variables with sampling distributions and the squared uncertainties $u^2(x_1), \ldots, u^2(x_n)$ are regarded as the known variances of the sampling distributions of $x_1, \ldots, x_n$, respectively. We use the same symbols $x_1, \ldots, x_n$ for the random variables as well as for their realized values.

where $w_i = 1/u^2(x_i)$ for $i = 1, 2, \ldots, n$, and $x_W = \sum_i w_i x_i / \sum_i w_i$ is the weighted mean of the results $x_1, \ldots, x_n$. If the calculated value of R^2 is substantially larger than one or equivalently the calculated value of $(n-1)R^2 = \sum_i w_i (x_i - x_W)^2$ is substantially larger than $(n-1)$, then declare the results $x_1, \ldots, x_n$ to be inconsistent. A modern interpretation of the original form of Birge test is discussed in appendix A.

We need the following notation. Symbol R_0^2 denotes a value of R^2 calculated from the realized results $x_1, \ldots, x_n$ and their associated variances $u^2(x_1), \ldots, u^2(x_n)$. The symbol χ_ν^2 denotes the chi-square probability distribution with degrees of freedom ν as well as a random variable having the χ_ν^2 distribution. The symbol $\chi_\nu^2[1-\alpha]$ denotes the $100 \times (1-\alpha)$th percentile value of χ_ν^2 distribution; that is, $\Pr\{\chi_\nu^2 \leqslant \chi_\nu^2[1-\alpha]\} = 1-\alpha$, for $0 \leqslant \alpha \leqslant 1$. The percentiles of a chi-square distribution can be found in published tables or determined from statistical software.

In the Birge test statistic R^2, the sampling distributions of $x_1, \ldots, x_n$ are assumed to be independent and normal (Gaussian) with known variances $u^2(x_1), \ldots, u^2(x_n)$, respectively. Based on this assumption, a traditional statistical protocol is to compare the calculated value $(n-1)R_0^2$ with the 95th percentile $\chi_{(n-1)}^2[0.95]$ of the chi-square distribution with degrees of freedom $n-1$. If the event $\{(n-1)R_0^2 > \chi_{(n-1)}^2[0.95]\}$ occurs, then the calculated value $(n-1)R_0^2$ is said, with 95% confidence, to be significantly large. In this case the dispersion of the results $x_1, \ldots, x_n$ appears larger than what can reasonably be expected from the stated variances $u^2(x_1), \ldots, u^2(x_n)$ with 95% confidence. Thus the results are declared to be inconsistent with 95% confidence.

Inconsistency implies that either the expected values $E(x_1), \ldots, E(x_n)$ are not equal or the stated variances $u^2(x_1), \ldots, u^2(x_n)$ are too small [2]. If the stated variances $u^2(x_1), \ldots, u^2(x_n)$ are believed to be reliable then inconsistency would imply that the expected values $E(x_1), \ldots, E(x_n)$ do not appear to be equal, the degree of trust in this judgment is indicated by the confidence level.

In the next section, we present a definition of statistical consistency in interlaboratory results motivated by the Birge test. The Birge test applies to uncorrelated results; however, the definition of consistency given in this paper applies to both uncorrelated and correlated results. We show that the Birge test may be interpreted as a classical (frequentist sampling theory) test of the null hypothesis that the variances of the results $x_1, \ldots, x_n$ are less than or equal to their stated values $u^2(x_1), \ldots, u^2(x_n)$ against the alternative hypothesis that the variances of $x_1, \ldots, x_n$ are greater than $u^2(x_1), \ldots, u^2(x_n)$. We show that the traditional statistical protocol of the Birge test of consistency is equivalent to checking whether the classical p-value of $(n-1)R_0^2$ is less than 0.05. The classical p-value is the maximum probability under the null hypothesis of realizing a value of $(n-1)R^2$ equal to or larger than $(n-1)R_0^2$ in contemplated replications of the interlaboratory evaluation.

Subsequently, we show that the classical p-value of the Birge test statistic is equal to the Bayesian posterior probability of the null hypothesis based on suitably chosen non-informative prior distributions for the unknown statistical parameters. The Birge test may therefore be interpreted also as a Bayesian test of the null hypothesis. The Bayesian interpretation makes it possible to use Bayesian statistics for checking consistency in interlaboratory evaluations. The use of Bayesian statistics is important in metrology because the *Guide to the Expression of Uncertainty in Measurement* (GUM) [3] agrees with Bayesian statistics [4].

In addition, we present a general test of consistency for both uncorrelated and correlated results, of which the Birge test is a special case. Then we show that the classical p-value of the general test statistic is equal to the Bayesian posterior probability of the null hypothesis based on non-informative prior distributions. The general test makes it possible to check consistency in correlated interlaboratory results.

2. Classical interpretation of the Birge test and its generalized version

A definition of consistency in the results from an interlaboratory evaluation, motivated by the Birge test, is as follows.

Definition. The results $x_1, \ldots, x_n$ are said to be consistent relative to their stated variances $u^2(x_1), \ldots, u^2(x_n)$ and covariances $u(x_{1,}x_2), \ldots, u(x_{n-1}, x_n)$ if their dispersion is not greater than what can be expected from the following statistical model:

$$x_i = \mu + e_i, \qquad (2)$$

for $i = 1, 2, \ldots, n$, where (i) μ is an unknown constant statistical parameter; (ii) the errors $e_1, \ldots, e_n$ are random variables having sampling probability distributions with a common expected value zero, variances $\sigma_1^2, \ldots, \sigma_n^2$, and covariances $\sigma_{1,2}, \ldots, \sigma_{n-1,n}$; (iii) the variances $\sigma_1^2, \ldots, \sigma_n^2$ are known and equal to $u^2(x_1), \ldots, u^2(x_n)$, respectively, all assumed to be positive; (iv) the covariances $\sigma_{1,2}, \ldots, \sigma_{n-1,n}$ are either all zero or known and equal to certain stated values $u(x_{1,}x_2), \ldots, u(x_{n-1}, x_n)$, respectively; and (v) the joint distribution of $e_1, \ldots, e_n$ is an n-variate normal distribution.

The statistical model (2) postulates that the random variables $x_1, \ldots, x_n$, have a joint n-variate normal sampling distribution with a common unknown expected value μ, known variances $u^2(x_1), \ldots, u^2(x_n)$, and known covariances $u(x_1, x_2), \ldots, u(x_{n-1}, x_n)$. Occasionally it is necessary to check the consistency of correlated interlaboratory results. Therefore, we have defined consistency for both uncorrelated and correlated results. In the Birge test all covariances are zero. We will refer to the statistical model (2) as a model of consistency.

One-parameter statistical consistency model. In matrix form, the statistical consistency model (2) is that the random vector $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{t}}$ has an n-variate normal distribution, $N(\mu\mathbf{1}, \boldsymbol{D})$, with expected value $\mu\mathbf{1}$ and variance–covariance matrix (dispersion matrix) $\boldsymbol{D}$, where $\mathbf{1} = (1, \ldots, 1)^{\mathrm{t}}$, the variances $u^2(x_1), \ldots, u^2(x_n)$ are diagonal elements of $\boldsymbol{D}$, and the covariances $u(x_{1,}x_2), \ldots, u(x_{n-1}, x_n)$ are off-diagonal elements of $\boldsymbol{D}$; that is

$$\boldsymbol{x} \sim N(\mu\mathbf{1}, \boldsymbol{D}). \qquad (3)$$

The superscript t introduced in the definitions of $\boldsymbol{x}$ and $\mathbf{1}$ indicates transpose of a vector or of a matrix. By the relational symbol $\sim$ used in (3) we mean that the random vector $\boldsymbol{x}$ has the probability distribution $N(\mu\mathbf{1}, \boldsymbol{D})$. Using the notation $u(x_i, x_i) = u^2(x_i)$ for $i = 1, 2, \ldots, n$, we can express the variance–covariance matrix $\boldsymbol{D}$ as $[u(x_i, x_j)]$.

Following [5], the dispersion matrix $\boldsymbol{D}$ in model (3) is assumed to be *known* and *positive definite*. A square matrix $\boldsymbol{D}$ is said to be positive definite if $\boldsymbol{a}^{\mathrm{t}}\boldsymbol{D}\boldsymbol{a} \geqslant 0$ for all $\boldsymbol{a} = (a_1, \ldots, a_n)^{\mathrm{t}}$ and $\boldsymbol{a}^{\mathrm{t}}\boldsymbol{D}\boldsymbol{a} = 0$ only if $\boldsymbol{a} = \mathbf{0}$ [6]. If $\boldsymbol{D} = \mathrm{Diag}[u^2(x_1), \ldots, u^2(x_n)]$ then $\boldsymbol{a}^{\mathrm{t}}\boldsymbol{D}\boldsymbol{a} = \sum_i a_i^2 u^2(x_i)$; therefore, if the variances $u^2(x_1), \ldots, u^2(x_n)$ are positive, then the matrix $\boldsymbol{D}$ is positive definite. For typical covariances between interlaboratory results, the variance–covariance matrices are positive definite. If the dispersion matrix $\boldsymbol{D}$ is not known then the problems of defining and assessing consistency in interlaboratory results are considerably more difficult.

More general two-parameter statistical model. The only unknown parameter in the statistical consistency model (3) is μ. One way of interpreting the Birge test of consistency is to consider the following more general statistical model: $\boldsymbol{x}$ has the normal distribution, $N(\mu\mathbf{1}, \tau^2\boldsymbol{D})$, with expected value $\mu\mathbf{1}$ and dispersion matrix $\tau^2\boldsymbol{D}$, for some positive parameter τ^2; that is

$$\boldsymbol{x} \sim N(\mu\mathbf{1}, \tau^2\boldsymbol{D}). \tag{4}$$

Model (4) has two unknown parameters μ and τ^2. Model (4) has the virtue that it can fit for some value of τ^2 any degree of dispersion, large or small, between the results $x_1, \ldots, x_n$. This property of model (4) makes it suitable for statistical interpretation of the Birge test of consistency. The statistical consistency model (3) is a special case of (4) in which τ^2 is assumed to be one.

Estimation of the parameters of model (4). We show in appendix B that the classical (frequentist sampling theory) estimates of μ and τ^2 and their sampling distributions are as follows.

(i) An unbiased estimate of the parameter μ is $m = \boldsymbol{B}^{\mathrm{t}}\boldsymbol{x}$, where $\boldsymbol{B}^{\mathrm{t}} = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}$.
(ii) The sampling distribution of m is normal with expected value $E(m) = \mu$ and variance $V(m) = \tau^2 \times (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}$.
(iii) An unbiased estimate of the parameter τ^2 is $Q^2 = \boldsymbol{x}^{\mathrm{t}}\boldsymbol{C}\boldsymbol{x}/(n-1)$, where $\boldsymbol{C} = [\boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}]$.
(iv) The sampling distribution of Q^2 is $\tau^2/(n-1)$ times the $\chi^2_{(n-1)}$ distribution.
(v) The sampling distributions of the estimates m and Q^2 are independent.

The estimate $m = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}$ is free of the value of τ^2. Thus m is an estimate of μ in model (3) also. We will use the symbol Q_0^2 to denote a value of Q^2 calculated from the realized results $x_1, \ldots, x_n$. A benefit of the statistical independence of the estimates m and Q^2 is indicated in appendix F.

In model (4), the estimate $m = \boldsymbol{B}^{\mathrm{t}}\boldsymbol{x}$ is the *minimum variance unbiased estimate* of μ in the sense that it is unbiased and it has the smallest variance among all unbiased estimates of μ [8, section 5a.2]. The estimate $Q^2 = \boldsymbol{x}^{\mathrm{t}}\boldsymbol{C}\boldsymbol{x}/(n-1)$ is also the *minimum variance unbiased estimate* of τ^2 [8, section 5a.2]. Thus m and Q^2 are statistically optimal classical (frequentist sampling theory) estimates of μ and τ^2, respectively.

If the results $x_1, \ldots, x_n$ were mutually independent then, as discussed in appendix B, we have the following simplifications.

(i) The estimate $m = \boldsymbol{B}^{\mathrm{t}}\boldsymbol{x}$ of μ reduces to the weighted mean $x_{\mathrm{W}} = \sum_i w_i x_i / \sum_i w_i$.
(ii) The sampling distribution of x_{W} is $N(\mu, \tau^2/\sum_i w_i)$.
(iii) The estimate $Q^2 = \boldsymbol{x}^{\mathrm{t}}\boldsymbol{C}\boldsymbol{x}/(n-1)$ of τ^2 reduces to the Birge statistic $R^2 = \sum_i w_i(x_i - x_{\mathrm{W}})^2/(n-1)$.
(iv) The sampling distribution of R^2 is $\tau^2/(n-1)$ times the $\chi^2_{(n-1)}$ distribution.
(v) The sampling distributions of the estimates x_{W} and R^2 are independent.

Testing hypothesis of consistency. A test of consistency in the results $x_1, \ldots, x_n$ relative to the variance–covariance matrix $\boldsymbol{D}$ may be thought of as a test of the null hypothesis $\mathrm{H}_0 : \tau^2 \leqslant 1$ against the alternative hypothesis $\mathrm{H}_1 : \tau^2 > 1$ in model (4). Since the variance–covariance matrix of $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{t}}$ in model (4) is $\tau^2\boldsymbol{D}$, the null hypothesis H_0 is that the variance–covariance matrix of $\boldsymbol{x}$ is less than or equal to $\boldsymbol{D}$, and the alternative hypothesis H_1 is that the variance–covariance matrix of $\boldsymbol{x}$ is greater than $\boldsymbol{D}$. If the results $x_1, \ldots, x_n$ fit the model (4) for an estimate Q_0^2 of τ^2 that is greater than 1, then the dispersion of $x_1, \ldots, x_n$ is greater than what can be expected from the smaller variance–covariance matrix $\boldsymbol{D}$; then the results $x_1, \ldots, x_n$ would be inconsistent relative to the variance–covariance matrix $\boldsymbol{D}$. If the results $x_1, \ldots, x_n$ fit the model (4) for an estimate Q_0^2 of τ^2 that is less than or equal to 1, then the dispersion of $x_1, \ldots, x_n$ is in agreement with the estimated variance–covariance matrix $Q_0^2\boldsymbol{D}$, which is less than or equal to $\boldsymbol{D}$. In this case either the dispersion of the results $x_1, \ldots, x_n$ fits model (3) or it is less than the dispersion that may be expected from model (3). In either case the hypothesis of consistency relative to the variance–covariance matrix $\boldsymbol{D}$ cannot be rejected.

Birge test of consistency for uncorrelated results. In this subsection all covariances $u(x_1, x_2), \ldots, u(x_{n-1}, x_n)$ are zero. Thus $\boldsymbol{D}$ is $\mathrm{Diag}[u^2(x_1), \ldots, u^2(x_n)]$, a known positive definite matrix. In model (4), a statistically optimal classical (frequentist sampling theory) estimate of τ^2 is $R^2 = \sum_i w_i(x_i - x_{\mathrm{W}})^2/(n-1)$. The sampling distribution of $(n-1)R^2$ is known to be τ^2 times the chi-square distribution with degrees of freedom $n-1$. Therefore, a suitable test statistic for testing the null hypothesis $\mathrm{H}_0 : \tau^2 \leqslant 1$ against its alternative hypothesis $\mathrm{H}_1 : \tau^2 > 1$ is $(n-1)R^2$. Small values of $(n-1)R^2$ favour the null hypothesis, $\mathrm{H}_0 : \tau^2 \leqslant 1$, and large values of $(n-1)R^2$ favour the alternative hypothesis, $\mathrm{H}_1 : \tau^2 > 1$.

A traditional statistical protocol to test the null hypothesis $\mathrm{H}_0 : \tau^2 \leqslant 1$ against the alternative hypothesis $\mathrm{H}_1 : \tau^2 > 1$ is to compare the calculated value $(n-1)R_0^2$ with the 95th percentile $\chi^2_{(n-1)}[0.95]$. If the event

$$(n-1)R_0^2 > \chi^2_{(n-1)}[0.95] \tag{5}$$

occurs then the hypothesis $H_0 : \tau^2 \leqslant 1$ is rejected and the results $x_1, \ldots, x_n$ are declared to be inconsistent with at least 95% confidence; that is, the probability of committing a Type I error is 0.05 or less. Thus the Birge test as described in section 1 may be interpreted as a classical test of the null hypothesis $H_0 : \tau^2 \leqslant 1$ versus the alternative hypothesis $H_1 : \tau^2 > 1$ in model (4).

The Birge test is the *uniformly most powerful statistical test* of the null hypothesis $H_0 : \tau^2 \leqslant 1$ against the alternative hypothesis $H_1 : \tau^2 > 1$ [9, Theorem 8.3.17, p 391]. This means that the Birge test has a larger statistical *power* than any other statistical test of $H_0 : \tau^2 \leqslant 1$ against $H_1 : \tau^2 > 1$ for every value of $\tau^2 > 1$. The statistical power of a test of hypothesis is one minus the probability of committing a Type II error.

Classical p-value of the null hypothesis. A modern statistical protocol to test the null hypothesis $H_0 : \tau^2 \leqslant 1$ against the alternative hypothesis $H_1 : \tau^2 > 1$ is to determine the classical p-value of $(n-1)R_0^2$ under the null hypothesis. The classical p-value is the maximum probability under the null hypothesis of realizing a value of $(n-1)R^2$ equal to or larger than $(n-1)R_0^2$ in contemplated replications of the interlaboratory evaluation according to model (4) [9, section 8.3.4, p 397]. We use the symbol p_C for the p-value. As discussed in appendix C,

$$p_C = \Pr\{\chi^2_{(n-1)} \geqslant (n-1)R_0^2\}. \tag{6}$$

If the p-value p_C is very small then the computed value $(n-1)R_0^2$ does not favour the null hypothesis $H_0 : \tau^2 \leqslant 1$. In this case, the results are judged to be inconsistent relative to the given variances. A traditional benchmark for assessing the classical p-value is 0.05. As noted in appendix D, the event $\{p_C < 0.05\}$ is equivalent to the event $\{(n-1)R_0^2 > \chi^2_{(n-1)}[0.95]\}$. Therefore the traditional protocol of the Birge test of consistency is equivalent to computing the p-value and checking whether it is less than 0.05.

General test of consistency for both correlated and uncorrelated results. In this subsection $\boldsymbol{V} = \tau^2\boldsymbol{D}$, where $\boldsymbol{D} = [u(x_i, x_j)]$, a known positive definite matrix. A statistically optimal classical (frequentist sampling theory) estimate of τ^2 is $Q^2 = \boldsymbol{x}^{\rm t}\boldsymbol{C}\boldsymbol{x}/(n-1)$, where $\boldsymbol{C} = [\boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\boldsymbol{1}(\boldsymbol{1}^{\rm t}\boldsymbol{D}^{-1}\boldsymbol{1})^{-1}\boldsymbol{1}^{\rm t}\boldsymbol{D}^{-1}]$. The sampling distribution of $(n-1)Q^2$ is known to be τ^2 times the chi-square distribution with degrees of freedom $n-1$. Therefore, a suitable test statistic for testing the consistency hypothesis $H_0 : \tau^2 \leqslant 1$ against its alternative $H_1 : \tau^2 > 1$ is $(n-1)Q^2 = \boldsymbol{x}^{\rm t}\boldsymbol{C}\boldsymbol{x}$. Small values of $(n-1)Q^2$ favour the consistency hypothesis, $H_0 : \tau^2 \leqslant 1$, and large values of $(n-1)Q^2$ favour its alternative, $H_1 : \tau^2 > 1$. The realized value of the test statistic $(n-1)Q^2$ is denoted by $(n-1)Q_0^2$.

As discussed in appendix C, the classical p-value under the null hypothesis of realizing a value of $(n-1)Q^2$ equal to or larger than $(n-1)Q_0^2$ in contemplated replications of the interlaboratory evaluation according to model (4) is

$$p_C = \Pr\{\chi^2_{(n-1)} \geqslant (n-1)Q_0^2\}. \tag{7}$$

Thus a general test of consistency for both correlated and uncorrelated interlaboratory results is to compute the p-value from expression (7). If the p-value is very small then the realized value $(n-1)Q_0^2$ of the test statistic does not favour the null hypothesis $H_0 : \tau^2 \leqslant 1$. In this case, the results $x_1, \ldots, x_n$ are judged to be inconsistent relative to the given variance–covariance matrix $\boldsymbol{D} = [u(x_i, x_j)]$. A traditional benchmark for assessing the classical p-value is 0.05. The corresponding protocol is to declare the results to be inconsistent with a confidence level of at least 95% if $p_C < 0.05$.

The general test of consistency is the *uniformly most powerful statistical test* of the null hypothesis $H_0 : \tau^2 \leqslant 1$ against the alternative hypothesis $H_1 : \tau^2 > 1$ [9, theorem 8.3.17, p 391]. When the results $x_1, \ldots, x_n$ are uncorrelated and hence $\boldsymbol{D} = \mathrm{Diag}[u^2(x_1), \ldots, u^2(x_n)]$ the general test of consistency reduces to the Birge test of consistency discussed in the previous subsection.

3. Bayesian interpretation of the Birge test and its generalized version

With reference to model (4), Bayesian statistical inference deals with state-of-knowledge probability density functions (pdfs) about the statistical parameters μ and τ^2. A state-of-knowledge pdf represents belief probabilities about the possible values of a parameter based on all available information. The realized results $x_1, \ldots, x_n$ and their functions, such as $m = \boldsymbol{B}^{\rm t}\boldsymbol{x}$ and $Q_0^2 = \boldsymbol{x}^{\rm t}\boldsymbol{C}\boldsymbol{x}/(n-1)$, are regarded as given quantities (constants) in Bayesian statistics (as well as in classical statistics).

Bayesian inference. A Bayesian analysis starts with a prior distribution $p(\mu, \tau^2)$ which represents *a priori* state of knowledge about μ and τ^2 before the results $x_1, \ldots, x_n$ are seen. An output of Bayesian analysis is a posterior distribution $p(\mu, \tau^2|\boldsymbol{x})$ which represents *a posteriori* state of knowledge about μ and τ^2 conditional on the given results $x_1, \ldots, x_n$. The relationship between the results $x_1, \ldots, x_n$ and the parameters μ and τ^2 is described by *a likelihood function* $l(\mu, \tau^2|\boldsymbol{x})$ conditional on the results $\boldsymbol{x}$. A likelihood function is the sampling pdf $f(\boldsymbol{x}|\mu, \tau^2)$ regarded as a function of the parameters μ and τ^2 rather than of $\boldsymbol{x}$. A sampling pdf is a property of the data generation process with one or more unknown parameters. The posterior distribution $p(\mu, \tau^2|\boldsymbol{x})$ is obtained using Bayes's theorem [10, p 34] which states that the posterior distribution $p(\mu, \tau^2|\boldsymbol{x})$ is proportional to the product of the likelihood function $l(\mu, \tau^2|\boldsymbol{x})$ and the prior distribution $p(\mu, \tau^2)$. In symbols Bayes's theorem states that

$$p(\mu, \tau^2|\boldsymbol{x}) \propto l(\mu, \tau^2|\boldsymbol{x}) \times p(\mu, \tau^2). \tag{8}$$

A prior distribution need not be a proper probability distribution. Prior distributions that are not probability distributions are called improper distributions. A valid posterior distribution is always a proper probability distribution.

Improper prior distributions. Absence of prior knowledge (complete ignorance) is represented by using non-informative prior distributions, which are improper distributions. Bayesian inferences based on suitably chosen non-informative prior distributions are often numerically similar or identical to classical (frequentist sampling theory) inferences albeit with Bayesian interpretation.

We will use non-informative improper prior distributions for μ and τ^2. The choice of non-informative prior distributions keeps the basis of statistical inference about μ and τ^2 identical in the classical and the Bayesian analyses. Then Bayes's theorem yields a joint posterior distribution for μ and τ^2 conditional on the given results $x_1, \ldots, x_n$. From the joint posterior distribution for μ and τ^2, we will determine the marginal posterior distribution for τ^2. We will then use the posterior distribution for τ^2 to determine the probability of the interval $0 < \tau^2 \leqslant 1$ corresponding to the null hypothesis $H_0 : \tau^2 \leqslant 1$.

Bayesian posterior pdf for τ^2. The pdf of $\boldsymbol{x}$ according to model (4) is

$$f(\boldsymbol{x}|\mu, \tau^2) = (2\pi)^{-n/2}(\tau^2)^{-n/2}|\boldsymbol{D}|^{-1/2} \times \exp\left\{-\frac{1}{2\tau^2}(\boldsymbol{x} - \mu\mathbf{1})^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x} - \mu\mathbf{1})\right\}. \tag{9}$$

As discussed in appendix E, the quadratic form in (9) can be expressed as

$$(\boldsymbol{x} - \mu\mathbf{1})^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x} - \mu\mathbf{1}) = (n-1)Q_0^2 + \frac{(m-\mu)^2}{(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}}. \tag{10}$$

Thus the likelihood function of μ and τ^2 conditional on the results $\boldsymbol{x}$ is

$$l(\mu, \tau^2|\boldsymbol{x}) \propto (\tau^2)^{-n/2} \times \exp\left\{-\frac{1}{2\tau^2}\left[(n-1)Q_0^2 + \frac{(m-\mu)^2}{(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}}\right]\right\}. \tag{11}$$

A suitably chosen non-informative improper prior distribution for (μ, τ^2) is the product of improper prior distributions for μ and τ^2 which are uniform in μ and in $\log \tau^2$, respectively. Thus $p(\mu) \propto 1$, $p(\tau^2) \propto 1/\tau^2$ [10, p 53], and

$$p(\mu, \tau^2) = p(\mu) \times p(\tau^2) \propto \frac{1}{\tau^2}. \tag{12}$$

According to Bayes's theorem (8) the posterior distribution, $p(\mu, \tau^2|\boldsymbol{x})$, of μ and τ^2 given $\boldsymbol{x}$ is proportional to the product of the likelihood function $l(\mu, \tau^2|\boldsymbol{x})$ and the prior distribution $p(\mu, \tau^2)$. Thus,

$$p(\mu, \tau^2|\boldsymbol{x}) \propto (\tau^2)^{-n/2-1} \times \exp\left\{-\frac{1}{2\tau^2}\left[(n-1)Q_0^2 + \frac{(\mu-m)^2}{(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}}\right]\right\}. \tag{13}$$

The Bayesian posterior pdf for τ^2 given $\boldsymbol{x}$, $p(\tau^2|\boldsymbol{x})$, is obtained by integrating out μ from the joint posterior distribution $p(\mu, \tau^2|\boldsymbol{x})$; thus $p(\tau^2|\boldsymbol{x}) = \int p(\mu, \tau^2|\boldsymbol{x})\mathrm{d}\mu$. As discussed in appendix F, $p(\tau^2|\boldsymbol{x})$ is

$$p(\tau^2|\boldsymbol{x}) \propto (\tau^2)^{-(n-1)/2-1} \exp\left\{-\frac{(n-1)Q_0^2}{2\tau^2}\right\}. \tag{14}$$

Bayesian interpretation of the general test of consistency. Substituting $(n-1)Q_0^2/\tau^2 = \xi$, by the change of variables technique, the pdf of the distribution of ξ given $\boldsymbol{x}$ is

$$p(\xi|\boldsymbol{x}) \propto \xi^{(n-2)/2} \exp\left\{-\frac{\xi}{2}\right\}. \tag{15}$$

The expression (15) is the kernel of the pdf of a chi-square distribution with degrees of freedom $n-1$ [11, p 52]; therefore, $\xi = (n-1)Q_0^2/\tau^2 \sim \chi^2_{(n-1)}$. Consequently, the Bayesian posterior probability p_{B} of the interval $0 < \tau^2 \leqslant 1$ is

$$p_{\mathrm{B}} = \Pr\{\tau^2 \leqslant 1\} = \Pr\left\{\frac{(n-1)Q_0^2}{\tau^2} \geqslant (n-1)Q_0^2\right\} = \Pr\{\chi^2_{(n-1)} \geqslant (n-1)Q_0^2\}. \tag{16}$$

Thus a Bayesian interpretation of the general test of consistency is to determine the posterior probability p_{B} of the null hypothesis $H_0 : \tau^2 \leqslant 1$ from (16). If p_{B} is less than some benchmark such as 0.05, then the hypothesis that $\tau^2 \leqslant 1$ is rejected and the results are declared to be inconsistent.

From (7) and (16), we note that the Bayesian posterior probability p_{B} and the classical p-value p_{C} are identical. This relation between the classical p-value and Bayesian posterior probability is possible only when non-informative improper prior distributions are used for the unknown parameters [10, section 4.3]. This interpretation of the general test of consistency as a Bayesian hypothesis testing method is based on the methodology developed by Lindley [12].

The Birge test as a special case of the general test of consistency. When the results $x_1, \ldots, x_n$ are independent and hence $\boldsymbol{D} = \mathrm{Diag}[u^2(x_1), \ldots, u^2(x_n)]$, the quadratic form $Q_0^2 = \boldsymbol{x}^{\mathrm{t}}\,\boldsymbol{C}\,\boldsymbol{x}/(n-1)$, where $\boldsymbol{C} = [\boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}]$, reduces to the computed value of the Birge test statistic $R_0^2 = \sum_i w_i(x_i - x_{\mathrm{W}})^2/(n-1)$. Then the classical p-value of the Birge test under the null hypothesis $H_0 : \tau^2 \leqslant 1$ is $p_{\mathrm{C}} = \Pr\{\chi^2_{(n-1)} \geqslant (n-1)R_0^2\}$ as given in (6). When the results $x_1, \ldots, x_n$ are independent, the Bayesian posterior probability of the null hypothesis $\tau^2 \leqslant 1$ given in (16) reduces to $p_{\mathrm{B}} = \Pr\{\chi^2_{(n-1)} \geqslant (n-1)R_0^2\}$. Thus the classical p-value p_{C} of the Birge test and the corresponding Bayesian posterior probability p_{B} are identical.

4. Summary

The Birge test is a well known and widely used method to check for the consistency in the interlaboratory results $x_1, \ldots, x_n$ that are uncorrelated [5]. We interpreted the Birge test as a classical test of the null hypothesis that the variances of the results $x_1, \ldots, x_n$ are less than or equal to their stated values $u^2(x_1), \ldots, u^2(x_n)$ against the alternative hypothesis that the variances of $x_1, \ldots, x_n$ are greater than $u^2(x_1), \ldots, u^2(x_n)$. A modern statistical protocol for hypothesis testing is to calculate the classical p-value. The p-value is the maximum probability under the null hypothesis of realizing in conceptual replications a value of the test statistic equal to or larger than its realized (observed) value. The null hypothesis is rejected when the p-value is too small. We determined the p-value of the Birge test statistic. Then we showed that the classical p-value is equal to the Bayesian posterior probability of the null hypothesis based on non-informative prior distributions for the unknown statistical parameters. The Bayesian interpretation of the Birge test makes it possible to use Bayesian statistics for checking the consistency in uncorrelated interlaboratory results. This is important because the GUM agrees with Bayesian statistics and it is an international standard for expressing uncertainty.

Occasionally the interlaboratory results are correlated and it is necessary to check their consistency. We presented a general test of consistency for both uncorrelated and correlated results. We showed that the classical p-value of the general test statistic is equal to the Bayesian posterior probability of the null hypothesis based on non-informative prior distributions. The general test makes it possible to check the consistency of correlated results from interlaboratory evaluations. The Birge test is a special case of the general test of consistency.

Acknowledgments

An anonymous statistician, Tyler Estler, Bert Rust and Peter Huang provided useful comments on an earlier draft.

Appendix A. A modern interpretation of the original form of Birge test

The least squares estimate of the common expected value μ is the weighted mean x_{W} in both the statistical model (4) as well as its special case (3). In Birge's [1] terminology, the phrase 'probable error based on *internal consistency*' refers to the theoretical variance of the least squares estimate x_{W} based on model (3) and the phrase 'probable error based on *external consistency*' refers to the empirical (estimated) variance of x_{W} commensurate with the actual dispersion of the realized results $x_1, \ldots, x_n$. The internal consistency variance (theoretical variance) of x_{W} based on model (3) is $\sigma_{\mathrm{I}}^2 = 1/\sum_i w_i$. The external consistency variance (empirical variance) of x_{W} commensurate with the actual dispersion of $x_1, \ldots, x_n$ is given by model (4) as $\sigma_{\mathrm{E}}^2 = R_0^2 \times 1/\sum_i w_i$. Birge argued that if the results $x_1, \ldots, x_n$ were consistent then the ratio $\sigma_{\mathrm{E}}^2/\sigma_{\mathrm{I}}^2 = R_0^2$ of the external consistency variance σ_{E}^2 to the internal consistency variance σ_{I}^2 would be one except for statistical fluctuations. Therefore the values of the ratio $R_0^2 = \sigma_{\mathrm{E}}^2/\sigma_{\mathrm{I}}^2$ that are substantially larger than one indicate that the results are inconsistent.

Birge [1, p 219] had adopted the conservative policy of using for the variance of x_{W} the larger of the two expressions $\sigma_{\mathrm{I}}^2 = 1/\sum_i w_i$ and $\sigma_{\mathrm{E}}^2 = R_0^2 \times 1/\sum_i w_i$. If R_0^2 is substantially larger than one, then the results are inconsistent relative to the variances $u^2(x_1), \ldots, u^2(x_n)$ but consistent relative to the larger variances $R_0^2 u^2(x_1), \ldots, R_0^2 u^2(x_n)$; thus, $\sigma_{\mathrm{E}}^2 = R_0^2 \times 1/\sum_i w_i = R_0^2 \times 1/\sum_i (1/u^2(x_i)) = 1/\sum_i (1/R_0^2 u^2(x_i))$ is an appropriate estimate of the variance of x_{W}. If R_0^2 is substantially smaller than one, then Birge used the larger variance $\sigma_{\mathrm{I}}^2 = 1/\sum_i w_i$ as the variance of x_{W} corresponding to the model of consistency (3). When R_0^2 is close to one then σ_{I}^2 and σ_{E}^2 are not very different.

An important reference for the Birge test is [2, p 429]. The authors of [2] point out that if the value of R_0^2 is substantially less than one, then the stated variances $u^2(x_1), \ldots, u^2(x_n)$ may well be too large.

Appendix B. Estimates of μ and τ^2 and their sampling distributions

We use the symbol $\boldsymbol{V}$ for the variance of $\boldsymbol{x}$, assumed to be positive definite. Thus $\boldsymbol{x} \sim N(\mu\mathbf{1}, \boldsymbol{V})$. In model (4), $\boldsymbol{V} = \tau^2\boldsymbol{D}$, where $\boldsymbol{D} = [u(x_i, x_j)]$ and $\boldsymbol{V}^{-1} = (1/\tau^2)\boldsymbol{D}^{-1}$. The generalized least squares estimate (GLSE) of μ in model (4) is that value m of μ for which the quadratic form $(\boldsymbol{x} - \mu\mathbf{1})^{\mathrm{t}}\boldsymbol{V}^{-1}(\boldsymbol{x} - \mu\mathbf{1})$ is minimum [6, section 3.3]. Thus the GLSE of μ is a solution of the normal equation $(\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\mathbf{1})m = \mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\boldsymbol{x}$, which is $m = (\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\boldsymbol{x}$ [6, section 3.3]. Substituting $\boldsymbol{V}^{-1} = (1/\tau^2)\boldsymbol{D}^{-1}$, we have $m = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}$. Let $\boldsymbol{B}^{\mathrm{t}} = (\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1} = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}$, then $m = \boldsymbol{B}^{\mathrm{t}}\boldsymbol{x}$. Since in model (4), $\boldsymbol{x} \sim N(\mu\mathbf{1}, \boldsymbol{V})$, the distribution of the GLSE $m = \boldsymbol{B}^{\mathrm{t}}\boldsymbol{x}$ is normal with expected value $\mu\boldsymbol{B}^{\mathrm{t}}\mathbf{1} = \mu$ and variance $V(m) = \boldsymbol{B}^{\mathrm{t}}\boldsymbol{V}\boldsymbol{B} = (\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\mathbf{1})^{-1} = \tau^2 \times (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}$. Thus we have the following results.

Result 1. An unbiased estimate of μ in model (4) is $m = \boldsymbol{B}^{\mathrm{t}}\boldsymbol{x}$ where $\boldsymbol{B}^{\mathrm{t}} = (\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1} = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}$.

Result 2. The sampling distribution of $m = \boldsymbol{B}^{\mathrm{t}}\boldsymbol{x}$ in model (4) is normal with expected value μ and variance $V(m) = \boldsymbol{B}^{\mathrm{t}}\ \boldsymbol{V}\boldsymbol{B} = \tau^2 \times (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}$.

Since m is GLSE, the minimum value of the quadratic form $(\boldsymbol{x} - \mu\mathbf{1})^{\mathrm{t}}\boldsymbol{V}^{-1}(\boldsymbol{x} - \mu\mathbf{1})$ is $(\boldsymbol{x} - m\mathbf{1})^{\mathrm{t}}\boldsymbol{V}^{-1}(\boldsymbol{x} - m\mathbf{1})$. Since $(\boldsymbol{x} - m\mathbf{1}) = (\boldsymbol{x} - \mathbf{1}m) = (\boldsymbol{x} - \mathbf{1}\boldsymbol{B}^{\mathrm{t}}\ \boldsymbol{x}) = [\boldsymbol{x} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\boldsymbol{x}] = [\mathbf{I} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}]\boldsymbol{x}$, we have $(\boldsymbol{x} - m\mathbf{1})^{\mathrm{t}}\boldsymbol{V}^{-1}(\boldsymbol{x} - m\mathbf{1}) = \boldsymbol{x}^{\mathrm{t}}[\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}]\boldsymbol{x} = (1/\tau^2)\boldsymbol{x}^{\mathrm{t}}[\boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}]\boldsymbol{x}$. Let $\boldsymbol{A} = [\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}]$ and $\boldsymbol{C} = [\boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}]$, then $\boldsymbol{A} = (1/\tau^2) \times \boldsymbol{C}$ and the minimum value of $(\boldsymbol{x} - \mu\mathbf{1})^{\mathrm{t}}\boldsymbol{V}^{-1}(\boldsymbol{x} - \mu\mathbf{1})$ is $(\boldsymbol{x} - m\mathbf{1})^{\mathrm{t}}\boldsymbol{V}^{-1}(\boldsymbol{x} - m\mathbf{1}) = \boldsymbol{x}^{\mathrm{t}}\boldsymbol{A}\boldsymbol{x} = (1/\tau^2)\boldsymbol{x}^{\mathrm{t}}\boldsymbol{C}\boldsymbol{x}$. We use the symbol Q^2 for $\boldsymbol{x}^{\mathrm{t}}\boldsymbol{C}\boldsymbol{x}/(n-1)$ and Q_0^2 for its realized value.

Now we state three theorems about the distributions of quadratic forms from [6, section 2.5].

Theorem 1. *If the expected value and variance of* $\boldsymbol{x}$ *are* $E(\boldsymbol{x}) = \boldsymbol{\mu}$ *and* $V(\boldsymbol{x}) = \boldsymbol{V}$, *then* $E(\boldsymbol{x}^{\mathrm{t}}\boldsymbol{A}\boldsymbol{x}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{V}) + \boldsymbol{\mu}^{\mathrm{t}}\boldsymbol{A}\boldsymbol{\mu}$.

(The symbol tr(.) stands for the trace of a matrix which means the sum of diagonal elements. This theorem does not require $\boldsymbol{x}$ to have a normal distribution.)

Theorem 2. *If* $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ *then* $\boldsymbol{x}^{\mathrm{t}}\boldsymbol{A}\boldsymbol{x} \sim$ *non-central chi-square distribution with degrees of freedom equal to rank of* $\boldsymbol{A}$ *and non-centrality parameter* $\frac{1}{2}\boldsymbol{\mu}^{\mathrm{t}}\boldsymbol{A}\boldsymbol{\mu}$ *if and only if* $\boldsymbol{A}\boldsymbol{V}$ *is idempotent, that is* $\boldsymbol{A}\ \boldsymbol{V}\boldsymbol{A}\ \boldsymbol{V} = \boldsymbol{A}\ \boldsymbol{V}$.

Theorem 3. *If* $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ *then* $\boldsymbol{x}^{\mathrm{t}}\boldsymbol{A}\boldsymbol{x}$ *and* $\boldsymbol{B}^{\mathrm{t}}\boldsymbol{x}$ *are distributed independently if and only if* $\boldsymbol{B}^{\mathrm{t}}\boldsymbol{V}\boldsymbol{A} = \mathbf{0}^{\mathrm{t}}$.

If $\boldsymbol{A} = [\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}]$ then $\boldsymbol{A}\,\boldsymbol{V} = [\boldsymbol{I} - \boldsymbol{V}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}]$ and $\boldsymbol{A}\,\boldsymbol{V}\boldsymbol{A}\,\boldsymbol{V} = \boldsymbol{A}\,\boldsymbol{V}$, so $\boldsymbol{A}\,\boldsymbol{V}$ is an idempotent matrix. The rank of an idempotent matrix is equal to its trace [7, p 134]. The trace of $\boldsymbol{A}\,\boldsymbol{V}$ is $\mathrm{tr}(\boldsymbol{I} - \boldsymbol{V}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}) = \mathrm{tr}(\boldsymbol{I}) - \mathrm{tr}((\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\mathbf{1}) = n - 1$; therefore, its rank is also $n - 1$. Since $\boldsymbol{V}$ is positive definite and hence non-singular, the rank of $\boldsymbol{A}\,\boldsymbol{V}$ is the rank of $\boldsymbol{A}$. Thus the rank of $\boldsymbol{A}$ is $n - 1$. In addition, $\mathbf{1}^{\mathrm{t}}\boldsymbol{A}\mathbf{1}$ is zero; so $(\mu\mathbf{1})^{\mathrm{t}}\boldsymbol{A}(\mu\mathbf{1})$ is zero. If $\boldsymbol{B}^{\mathrm{t}} = (\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}$ and $\boldsymbol{A} = [\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}]$ then $\boldsymbol{B}^{\mathrm{t}}\boldsymbol{V}\boldsymbol{A}$ is zero. By applying the three theorems from [6] with $\boldsymbol{\mu} = \mu\mathbf{1}, \boldsymbol{V} = \tau^2\boldsymbol{D}, \boldsymbol{A} = [\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}]$, and $\boldsymbol{B}^{\mathrm{t}} = (\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{V}^{-1}$ we get the following results.

Result 3. $E[\boldsymbol{x}^{\mathrm{t}}\boldsymbol{A}\boldsymbol{x}] = (1/\tau^2)E[\boldsymbol{x}^{\mathrm{t}}\boldsymbol{C}\,\boldsymbol{x}] = (n-1)E(Q^2)/\tau^2 = (n-1)$; therefore, $E(Q^2) = \tau^2$. That is, $Q^2 = \boldsymbol{x}^{\mathrm{t}}\boldsymbol{C}\boldsymbol{x}/(n-1)$ is an unbiased estimate of τ^2, where $\boldsymbol{C} = [\boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}]$. This particular result does not require $\boldsymbol{x}$ to have a normal distribution.

Result 4. The sampling distribution of $(\boldsymbol{x} - m\mathbf{1})^{\mathrm{t}}\boldsymbol{V}^{-1}(\boldsymbol{x} - m\mathbf{1}) = \boldsymbol{x}^{\mathrm{t}}\boldsymbol{A}\boldsymbol{x} = (1/\tau^2) \times \boldsymbol{x}^{\mathrm{t}}\boldsymbol{C}\boldsymbol{x} = (n-1)Q^2/\tau^2$ is $\chi^2_{(n-1)}$ (central chi-square distribution with degrees of freedom $n-1$).

Result 5. The sampling distributions of $m = \boldsymbol{B}^{\mathrm{t}}\boldsymbol{x}$, and $(\boldsymbol{x}-m\mathbf{1})^{\mathrm{t}}\boldsymbol{V}^{-1}(\boldsymbol{x}-m\mathbf{1}) = \boldsymbol{x}^{\mathrm{t}}\boldsymbol{A}\boldsymbol{x} = (1/\tau^2)\times\boldsymbol{x}^{\mathrm{t}}\boldsymbol{C}\boldsymbol{x}$ are statistically independent. It follows that the sampling distributions of the estimates $m = \boldsymbol{B}^{\mathrm{t}}\boldsymbol{x}$ and $Q^2 = \boldsymbol{x}^{\mathrm{t}}\boldsymbol{C}\boldsymbol{x}/(n-1)$ are independent.

If we substitute $\tau^2 = 1$ in the above results, we get the corresponding results for model (3).

If the results $x_1, \ldots, x_n$ are mutually independent then all covariances $u(x_i, x_j)$, for $i \neq j$ and $i, j = 1, 2, \ldots, n$, are zero and $V(\boldsymbol{x}) = \boldsymbol{V} = \tau^2\boldsymbol{D}$, where $\boldsymbol{D} = \mathrm{Diag}[u^2(x_1), \ldots, u^2(x_n)]$. It follows that $\boldsymbol{V}^{-1} = (1/\tau^2)\boldsymbol{D}^{-1}$, where $\boldsymbol{D}^{-1} = \mathrm{Diag}[w_1, \ldots, w_n]$ and $w_i = 1/u^2(x_i)$, for $i = 1, 2, \ldots, n$. Further, $\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1} = \sum_i w_i$, $(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1} = 1/\sum_i w_i$, and $\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x} = \boldsymbol{x}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1} = \sum_i w_i x_i$, and $\boldsymbol{x}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x} = \sum_i w_i x_i^2$.

Then the GLSE of μ reduces to $m = \boldsymbol{B}^{\mathrm{t}}\boldsymbol{x} = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x} = \sum_i w_i x_i / \sum_i w_i = x_{\mathrm{W}}$, the weighted mean; the sampling distribution of x_{W} is normal with expected value μ and variance $V(x_{\mathrm{W}}) = \boldsymbol{B}^{\mathrm{t}}\boldsymbol{V}\boldsymbol{B} = \tau^2 \times (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1} = \tau^2 \times [1/\sum_i w_i]$ (see results 1 and 2).

If $x_1, \ldots, x_n$ are independent then Q^2 (an unbiased estimate of τ^2) reduces to $Q^2 = \boldsymbol{x}^{\mathrm{t}}\boldsymbol{C}\boldsymbol{x}/(n-1) = [\boldsymbol{x}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x} - \boldsymbol{x}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}]/(n-1) = [\sum_i w_i x_i^2 - (\sum_i w_i x_i)^2/(\sum_i w_i)]/(n-1) = \sum_i w_i(x_i - x_{\mathrm{W}})^2/(n-1) = R^2$, the Birge test statistic (see result 3). The sampling distribution of $(n-1)Q^2/\tau^2 = (n-1)R^2/\tau^2$ is $\chi^2_{(n-1)}$ distribution (see result 4). Thus the sampling distribution of R^2 is $\tau^2/(n-1)$ times the $\chi^2_{(n-1)}$ distribution. The sampling distributions of the estimates x_{W} and R^2 are mutually independent (see result 5).

Appendix C. Classical p-value p_{C}

With respect to the sampling distribution of $\boldsymbol{x}$ in model (4), the probability that $(n-1)Q^2$ is equal to or larger than $(n-1)Q_0^2$ is $\Pr\{(n-1)Q^2 \geqslant (n-1)Q_0^2\} = \Pr\{(n-1)Q^2/\tau^2 \geqslant (n-1)Q_0^2/\tau^2\} = \Pr\{\chi^2_{(n-1)} \geqslant (n-1)Q_0^2/\tau^2\}$.

If $\mathrm{H}_0 : \tau^2 \leqslant 1$ were true, then $\Pr\{\chi^2_{(n-1)} \geqslant (n-1)Q_0^2/\tau^2\} \leqslant \Pr\{\chi^2_{(n-1)} \geqslant (n-1)Q_0^2\}$.

Therefore, $\Pr\{\chi^2_{(n-1)} \geqslant (n-1)Q_0^2\}$ is the maximum of $\Pr\{\chi^2_{(n-1)} \geqslant (n-1)Q_0^2/\tau^2\}$ for $\tau^2 \leqslant 1$. Thus the p-value of the realized test statistic $(n-1)\ Q_0^2$ under the null hypothesis $\mathrm{H}_0 : \tau^2 \leqslant 1$ is $p_{\mathrm{C}} = \Pr\{\chi^2_{(n-1)} \geqslant (n-1)Q_0^2\}$. If the results $x_1, \ldots, x_n$ are mutually independent then $\boldsymbol{D} = \mathrm{Diag}[u^2(x_1), \ldots, u^2(x_n)]$ and Q_0^2 reduces to R_0^2. In this case the p-value reduces to $p_{\mathrm{C}} = \Pr\{\chi^2_{(n-1)} \geqslant (n-1)R_0^2\}$.

Appendix D. Events $\{p_{\mathrm{C}} < 0.05\}$ and $\{(n-1)R_0^2 > \chi^2_{(n-1)}[0.95]\}$ are equivalent

Event $\{p_{\mathrm{C}} < 0.05\} \Leftrightarrow Pr\{\chi^2_{(n-1)} \geqslant (n-1)R_0^2\} < 0.05 \Leftrightarrow \Pr\{\chi^2_{(n-1)} < (n-1)R_0^2\} \geqslant 0.95 \Leftrightarrow$ Event $\{(n-1)R_0^2 > \chi^2_{(n-1)}[0.95]\}$.

Appendix E. Quadratic form $(\boldsymbol{x} - \mu 1)^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x} - \mu 1)$

We can parse $(\boldsymbol{x} - \mu\mathbf{1})$ as $(\boldsymbol{x} - \mu\mathbf{1}) = (\boldsymbol{x} - m\mathbf{1}) + (m - \mu)\mathbf{1}$, where $m = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}$. Then the quadratic form $(\boldsymbol{x} - \mu\mathbf{1})^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x} - \mu\mathbf{1})$ can be parsed as

$$(\boldsymbol{x} - \mu\mathbf{1})^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x} - \mu\mathbf{1}) = (\boldsymbol{x} - m\mathbf{1})^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x} - m\mathbf{1}) \\ +(m - \mu)\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1}(m - \mu).$$

The cross product term $(m - \mu)\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x} - m\mathbf{1}) = (m - \mu)(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x} - \mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1}m)$ is zero because $\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1}m = \mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1}\,(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x} = \mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}$.

Since $(\boldsymbol{x} - m\mathbf{1}) = (\boldsymbol{x} - \mathbf{1}m) = [\boldsymbol{x} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}] = [\boldsymbol{I} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}]\boldsymbol{x}$, we have $(\boldsymbol{x} - m\mathbf{1})^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x} - m\mathbf{1}) = \boldsymbol{x}^{\mathrm{t}}\boldsymbol{C}\boldsymbol{x} = (n-1)Q_0^2$, where $\boldsymbol{C} = [\boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}]$. Also we can express $(m - \mu)\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1}\,(m - \mu)$ as $(m - \mu)^2/(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}$. Thus

$$(\boldsymbol{x} - \mu\mathbf{1})^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x} - \mu\mathbf{1}) = (n-1)Q_0^2 + \frac{(m - \mu)^2}{(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}}.$$

Appendix F. Bayesian posterior pdf $p(\tau^2|x)$ of τ^2 given $\boldsymbol{x}$

Since the matrix $\boldsymbol{D}$ is known, from (13) the integral $p(\tau^2|\boldsymbol{x}) = \int p(\mu, \tau^2|\boldsymbol{x})\mathrm{d}\mu$ is proportional to

$$(\tau^2)^{-(n-1)/2-1}\exp\left\{-\frac{(n-1)Q_0^2}{2\tau^2}\right\}\int[2\pi\tau^2 \times (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}]^{-1/2} \\ \times \exp\left\{-\frac{(\mu - m)^2}{2\tau^2 \times (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}}\right\}\mathrm{d}\mu.$$

The value of the integral in the above expression is one because its integrand is the pdf of a normal distribution for μ with expected value m and variance $\tau^2 \times (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}$. Thus the Bayesian posterior pdf $p(\tau^2|\boldsymbol{x})$ is the part given in front of the above integral. The marginal pdf $p(\tau^2|\boldsymbol{x})$ integrates out so easily from the joint pdf $p(\mu, \tau^2|\boldsymbol{x})$ because (i) the estimates m and Q^2 are sufficient statistics [9, section 6.2] for μ and τ^2, which is easy to see from the expressions (10) and (9), and (ii) the sampling distributions of m and Q^2 are independent.

References

[1] Birge R T 1932 The calculation of errors by the method of least squares *Phys. Rev.* **40** 207–27

[2] Taylor B N, Parker W H and Langenberg D N 1969 Determination of e/h, using macroscopic quantum phase coherence in superconductors: implications for quantum electrodynamics and the fundamental physical constants *Rev. Mod. Phys.* **41** 375–496

[3] BIPM, IEC, IFCC, ISO, IUPAC, IUPAP and OIML *Guide to the Expression of Uncertainty in Measurement* 2nd edn 1995 (Geneva: International Organization for Standardization) ISBN 92-67-10188-9

[4] Kacker R N and Jones A T 2003 On use of Bayesian statistics to make the *Guide to the Expression of Uncertainty in Measurement* consistent *Metrologia* **40** 235–48

[5] Cox M G 2002 The evaluation of key comparison data *Metrologia* **39** 589–95

[6] Searle S R 1971 *Linear Models* (New York: Wiley)

[7] Harville D A 1997 *Matrix Algebra from a Statistician's Perspective* (Berlin: Springer)

[8] Rao C R 1973 *Linear Statistical Inference and its Application* 2nd edn (New York: Wiley)

[9] Casella G and Berger R L 2002 *Statistical Inference* 2nd edn (Pacific Grove, CA: Duxbury)

[10] Lee P M 1997 *Bayesian Statistics, An Introduction* 2nd edn (Oxford: Oxford University Press)

[11] Evans M, Hastings N and Peacock B 2000 *Statistical Distributions* 3rd edn (New York: Wiley)

[12] Lindley D V 1965 *Introduction to Probability and Statistics from a Bayesian Viewpoint* (two volumes—Part I: *Probability* and Part II: *Inference*) (Cambridge: Cambridge University Press)