Journal of Statistical Planning and Inference ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Fusion learning for inter-laboratory comparisons

Jan Hannig [a,*], Qing Feng [a], Hari Iyer [b], C.M. Wang [b], Xuhua Liu [c]

[a] *Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, United States*
[b] *National Institute of Standards and Technology, United States*
[c] *Department of Mathematics, China Agricultural University, China*

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Confidence distributions
Generalized fiducial inference
Model averaging
Inter-laboratory trials
Key comparison experiments

ABSTRACT

In this paper we propose a Generalized Fiducial Inference inspired method for finding a robust consensus of several independently derived collection of confidence distributions (CDs) for a quantity of interest. The resulting fused CD is robust to the existence of potentially discrepant CDs in the collection. The method uses computationally efficient fiducial model averaging to obtain a robust consensus distribution without the need to eliminate discrepant CDs from the analysis. This work is motivated by a commonly occurring problem in inter-laboratory trials, where different national laboratories all measure the same unknown true value of a quantity and report their CDs. These CDs need to be fused to obtain a consensus CD for the quantity of interest. When some of the CDs appear to be discrepant, simply eliminating them from the analysis is often not an acceptable approach, particularly so in view of the fact that the *true value* being measured is not known and a discrepant result from a lab may be closer to the true value than the rest of the results. Additionally, eliminating one or more labs from the analysis can lead to political complications since all labs are regarded as equally competent. These considerations make the proposed method well suited for the task since no laboratory is explicitly eliminated from consideration. We report results of three simulation experiments showing that the proposed fiducial approach has better small sample properties than the currently used naive approaches. Finally, we apply the proposed method to obtain consensus CDs for gauge block calibration inter-laboratory trials and measurements of Newton's constant of gravitation (G) by several laboratories.

1. Introduction

Inter-laboratory trials are often conducted by leading metrology laboratories in the world to compare each others' capabilities for measuring various fundamental properties of substances. Such a trial typically involves two or more participants each of whom measures the (nominally) same unknown value (called *measurand*) and provides the result along with an assessment of the uncertainty in the result. The results are meant to be the best estimates of the measurand the participating laboratories are able to provide. Often the same or very similar protocols are used by the participating laboratories. In some cases different subsets of participants use different methods for measuring the same unknown quantity. This is particularly so when specific laboratories have special expertise in particular measurement methods. The results from such experiments are used to gauge how comparable the measurement capabilities are across the participating laboratories. In some cases such experiments lead to the creation of certified reference materials (CRMs) and a consensus value for the

* Corresponding author.
E-mail address: jan.hannig@unc.edu (J. Hannig).

https://doi.org/10.1016/j.jspi.2017.09.011
0378-3758/

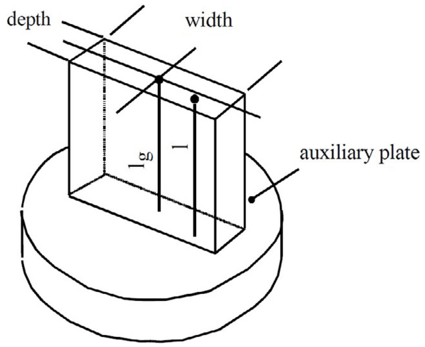

Fig. 1. The length of a gauge block is the distance from the gauging point on the top surface to the plane of the platen adjacent to the wrung gauge block.

measurand is arrived at by combining the results from the participating laboratories. This consensus value is used as the certified value for the CRM. The uncertainty associated with this certified value is used to provide an interval estimate of the value for the CRM.

Key comparisons

There is a particular class of inter-laboratory trials which takes on international significance. With the signing of the Mutual Recognition Arrangement (MRA) CIPM (1999), National Metrology Institutes (NMI's) and Regional Metrology Organizations (RMO's) around the world have undertaken the task of examining the *degree of equivalence* of their measurement standards. The CIPM (*Comité international des poids et mesures* — The International Committee on Weights and Measures), an entity whose principal task is to promote world-wide uniformity in units of measurement, works with member countries on issues related to the creation of measurement standards and comparisons of measurement capabilities of various national metrological laboratories (such as the National Institute of Standards and Technology (NIST) in the U.S, the National Physical Laboratory (NPL) in Great Britain, and Physikalisch-Technische Bundesanstalt (PTB) in Germany), and oversees the conduct of inter-laboratory experiments by participating NMIs to evaluate the relative measurement capabilities of each other and also to establish standard reference values (called Key Comparison Reference Value(s) or KCRV) for many important fundamental measurements and standards. The results obtained by the different laboratories are combined to arrive at the consensus KCRV value. Such comparisons *provide for the mutual recognition of calibration and measurement certificates issued by NMIs and thereby to provide governments and other parties with a secure technical foundation for wider agreements related to international trade, commerce and regulatory affairs.*

During any inter-laboratory trial it is generally the case that the results from one or a few laboratories differ noticeably from the rest even though all participating laboratories are considered to be more or less equally competent. It is natural to think that these apparently nonconforming values should perhaps be excluded from the calculation of a consensus value. There are at least two problems with this thinking. First, since the true value of the measurand is not known, one cannot say, based on any objective evidence, that one result is more believable than another. Second, there are political overtones associated with leaving out measured results of a laboratory since all participating laboratories are considered to be competent in their own right. Although discrepant results are subjected to further scrutiny to make sure such discrepancies are not the result of identifiable errors, when no errors are identified, each laboratory stands behind its result and the associated statements of uncertainty. Hence the problem of arriving at a consensus value takes on a greater level of significance when it comes to International Key Comparison Studies.

Gauge blocks

A gauge block (Thalmann, 2002) is a length standard having flat and parallel opposing surfaces. The cross-sectional shape is not very important, although the standard does give suggested dimensions for rectangular, square and circular cross-sections. Gauge blocks have nominal lengths defined in either the metric system (millimetres) or in the English system (1 inch = 25.4 mm). The length of the gauge block is defined at standard reference conditions:

temperature = 20 °C (68 °F)
barometric pressure = 101,325 Pa (1 atmosphere)
water vapour pressure = 1333 Pa (10 mm of mercury)
CO2 content of air = 0.03%.

The length of a gauge block is defined as the perpendicular distance from a gauging point on one end of the block to an auxiliary true plane wrung to the other end of the block, as shown in Fig. 1.

Please cite this article in press as: Hannig J., et al., Fusion learning for inter-laboratory comparisons. J. Statist. Plann. Inference (2017), https://doi.org/10.1016/j.jspi.2017.09.011.