UNIVERSITA
DEGLI STUDI
DI TORINO

POLITECNICO
DI TORINO

Università degli studi di Torino and Politecnico di Torino

Department of Maths
Doctoral School of Mathematical Sciences
XXX cycle

Hierarchical Bayesian models in metrology:
a walk in the accuracy evaluation and in its surroundings

This dissertation is submitted for the degree of
Doctor of Philosophy in Pure and Applied Mathematics
by

Giuseppe Galizia

Tutors:	prof. Roberto Fontana prof. Giulio Barbato
Coordinator:	prof. Riccardo Adami
Vice-coordinator:	prof. Anna Fino

Academic years 2014 to 2017
SSD MAT/06

For Martina,
lovely wife and
all-weather friend

Contents

ABSTRACT

In this research study the application of Hierarchical Bayesian models to metrology is due to a real industrial need. Nowadays, the problem of evaluating accuracy for a whole product line of testing equipments has not been yet completely addressed by an overall international standard. The most useful standard, because its aim is close to the purpose of accuracy evaluation of a product line, is the ISO 5725. This standard, under the general title “Accuracy (trueness and precision) of measurement methods and results”, sets the general principles, definitions and basis methods for the determination of trueness and precision under repeatability and reproducibility conditions. The statistics set out by the ISO 5725 are drawn from the ANOVA general linear model. Nevertheless, these statistics work well as we need to establish the deviation from the general mean value of each specific group and the components of variability within each specific group and between groups. That means, ISO 5725 works well for inter-laboratory experiments. Whereas, if we need to describe a possible coverage interval for the systematic errors and the ones for random errors, the ANOVA general model does not work so well. In fact, the confidence interval of the general mean, according to the ANOVA scheme, is not able to properly catch a reasonable interval describing where the systematic errors for a whole population of test equipments are expected to lay. This is relevant because in the last years, pushed by technological improvements, the international standards for testing methods are setting smaller and smaller variability requirements. Therefore, the manufacturers of the relative testing equipments need to deeply understand the systematic measurement errors component that is the error percentage that can be eliminated by removing its root cause, or controlled or adjusted. In order to overcome the limits of the ANOVA general linear models in the accuracy description for a whole product line, the present research suggests the usage of the hierarchical Bayesian model. Besides, it is coined the term homogeneity conditions in order to distinguish the intermediate level of precision, in between repeatability and reproducibility, that occurs in the experiments for the accuracy evaluation of a product line. The usage of Bayesian approach to overcome some limits of ANOVA modeling is not so new in metrology literature. Actually, Bayesian model has been already presented for the evaluation of between-bottle homogeneity (again homogeneity with regard to production conditions) studies in the production of reference and proficiency test materials [1]. Besides, the working group 1 of the International Bureau of Weights and Measures (BIPM) is developing another supplement to the Guide to the Expression of Uncertainty in Measurement, named JCGM 108 [2], that definitely opens to the Bayesian models in metrology. The efforts of the International Bureau are oriented to fade the dichotomy of the guide to the expression of uncertainty in measurement [3], that it is currently based on a mixture of frequentist (Type A statistical uncertainty evaluation from repeated measurements) and Bayesian (Type B uncertainty evaluation, based on state-of-knowledge distributions) frameworks. So, this research is also on the same line of the ongoing works of the International Bureau, in the direction of a fully Bayesian formulation for the accuracy evaluation.

The goodness of fit of the proposed hierarchical Bayesian model for accuracy evaluation is tested on two datasets: the former is drawn out from the ISO 5725-2 in order to have a reliable and authoritative benchmark at which compare the Bayesian outcome, while the latter expresses the real industrial application case, drew from the field of the equipments for thermo-mechanical testing of polymer, as the ISO 306 standard prescribes. The former is under reproducibility conditions, the latter under homogeneity conditions. Overall, the point estimates used to describe the accuracy for both ISO 5725 and Bayesian models are closer than the magnitude of the product resolution and this is considered enough in order to validate the model. The greatest advantage of the Bayesian model is the way of describing trueness and precision in terms of credibility intervals from the posterior distributions of bias and standard deviations under both repeatability and homogeneity conditions. The credibility interval for bias is really a good description of the deviance of the product line's general mean respect to the reference accepted value, being able to encompass the systematic errors that we can expect from a product line because they have not already been adjusted through a specific calibration. Moreover, at the end of the product development project, before the official launch of the new Instron testing equipments' family for measuring the Vicat Softening Temperature (VST in the following, it reflects the point of softening to be expected when a polymer is used in an elevated temperature application), it was adopted a

validation criteria based on the proposed hierarchical Bayesian model. That is the first tangible acknowledgment.

The present research ends with a quick overview on the application of Hierarchical Bayesian models to proficiency testing scheme, data consistency techniques and outliers detection tests in order to show how easily the hierarchical Bayesian models can be extended as number of levels for describing more nested probability structures and as number of metrology application fields. Also in this case an International standard, the ISO 13528, is leveraged in order to have a reliable benchmark. The choice for the representative dataset was to enlarge the VST experimental one through a random number generator according to some accuracy assumptions describing two additional product lines. In this way is possible provide that even if we add an additional hierarchical level, we do not lose consistency in the description of the accuracy for the specific product line that are in the second hierarchical level.

This research project is within the continuous research flow. It is far from putting the finishing touch to the Bayesian formulation of the accuracy description. Instead, it was conceived as a first determined step toward the fully Bayesian accuracy evaluation. That is the spirit, and also the letter, of the present research. And always in this spirit, for future developments and further industrial applications the whole R script is shared at the end of this study.

1 INTRODUCTION

The intellectual assent to the idea of the mathematics as the first driver for innovation is the vision that inspired this research project. During my PhD experience, this vision was nested into my work-life, that was a mix between the academic world of the Polytechnic of Turin and the industrial environment of Instron Italy. So attending courses, seminars, and sharing ideas with professors and other students, I gradually reached the root idea for this study.

Before jumping in the core topic, just few words about Instron Italy: it is a manufacturer of test equipments, designed to evaluate the mechanical properties of materials and structures using tensile, compression, flexural, fatigue, impact, dynamic, torsional, multi-axial and thermal loading. The majority of these characterizations for the material behaviour under specific testing conditions is in accordance with international standards, such as ASTM and ISO, and the final output is always a measurement result, that is, the set of quantity values that are attributed to a quantity together with any other available relevant information [5, def. 2.9], such as the measurement uncertainty. Hence, the main domain of this paper is the mathematics applied to the science of measurement, in other words, the metrology.

In the life span of a manufacturer of test equipments, there comes always the time for launching a new product. For each of these events, the product development team has to evaluate some metrological performances of the new equipment. Although in the last 15 years the International Organization for Standardization (ISO) has produced several documents on measurement accuracy, in such events there are two possible scenarios.

In the first one the problem of metrological evaluation is well defined because there is an international standard aimed at the calibration and verification of the specific test equipment and providing accuracy classes that meet stated metrological requirements to which every single instrument has to comply. For example, ISO 7500-1 specifies the calibration and verification of tension/compression testing machine in order to confirm that the performance properties satisfy the limits given for a specified class [5]. So, there is a measurement procedure to be used in assessing the accuracy of measured quantity values with respect to some specified classes.

In the second scenario a similar procedure does not exist at all or the reference operating conditions [4, def. 4.11] are not so clear. For example ISO 306 specifies four methods for the determination of the Vicat softening temperature (VST) of thermoplastic materials [6, def.3.3], whereas the verification procedure is not well defined. In such a case, the manufacturer has also to define a way for describing and verifying the metrological performances of every influence quantity [4, def. 2.52] controlled by the instrument. Back to VST example, the measurand is the temperature at which a flat-ended needle penetrates the specimen to a depth of 1 mm under a specified load using a selected uniform rate of temperature rise, so, the VST measurement is affected by the following main influence quantities:

- length, that is the needle penetration depth;
- mass, that is the specified load;
- temperature rate, derived from time and temperature.

For each of this quantities, the manufacturer not only has to design a testing equipment able to satisfy the standard requirements but also has to define a procedure for providing the metrological evidence of the compliance respect to the acceptance criteria stated in the ISO 306. In order to be more clear, looking at the requirement for the penetration measuring device, the standard requires an accuracy of ±0,01 mm [6, Cpv. 5.2.4] but it not clarifies the procedure aimed to verify this acceptance criterion once the machine is on field. The verification procedure is on the shoulder of manufactures and, so, procedures can differ manufacturer by manufacturer. So, in this shady regulatory environment the degrees of freedom of the verification procedure, such as test conditions and statistical objects for accuracy description of the penetration measuring device, make no easy the performance comparison between testing equipments of different manufacturers.

The core business of my PhD work is mathematics, so, all my efforts were focused on finding a generic and flexible model aimed at the evaluation of the metrological performances for a whole new product line. Thus, in this study the design of the verification procedure and the choice for the references is negligible with respect to the effort lavished on analyzing the modelling problem.

The industrial case around which the idea of this project grew up is the evaluation of displacement measurement accuracy, that, as said, is one of the main reference conditions for VST method. To meet this scope, the most consolidated approach follows the statements of ISO-5725. This international standard, under the general title “Accuracy (trueness and precision) of measurement methods and results”, provides basic methods for the determination of trueness and precision of a standard measurement method and results. The standard states that the variability of a measurement method can range as function of two extremes conditions, termed repeatability and reproducibility. The former corresponds to the minimum and the latter to the maximum variability in results. Intermediate conditions between these two extreme conditions for precision evaluation are also conceivable, when one or more of the different factors contributing to the variability are allowed to vary [7]. Looking at the manufacturer point of view, the factor needing to be assessed is primarily the contribution of the equipments. Instead, to be consistent with the method point of view, what needs to be assessed is the maximum level of variability arising from collaborative inter-laboratory experiment, that is reproducibility for ISO 5725. It is now evident as the main focus of manufacturers is an intermediate level of precision between repeatability and reproducibility conditions. Hence, the general problem can be modelled as a fully nested experimental design with three different levels: the upper level is for laboratories, the bottom level is for residuals and the intermediate level is for the equipments. It is just the intermediate level which manufacturers take care of, that is about the homogeneity of manufacturing conditions.

The ISO 5725-3 states that the analysis of the results of an *n*-factor fully-nested experiment is carried out by the statistical technique “analysis of variance” (ANOVA). As an alternative, the approach proposed in this study is the hierarchical Bayesian model. In a certain way, it is “natural” to recognise that it is unlikely that all different manufacturers have the same underlying precision for their test equipment, due to employing different design solutions, but it is also reasonable to assume that knowing something about the equipments of other manufacturers tells some additional information about the test method. So, the general purpose of such study is to provide a basis for international comparison of test equipment performances. The project is rooted in the idea of generic reusable components that can be put together as desired, like a child’s construction set but not so colourful. A unique model where it is easy to add additional intermediate hierarchical levels. One of these intermediate levels is just useful for manufactures needs. In addition, the posterior distribution achieved with the Bayesian inference would be fully consistent with the definition of measurand, that is the quantity to be measured and it is supposed to be specified not by a value but only by a description [3, Cpv. D]. Although this definition leaves room for interpretation, the posterior distribution achieved with the Bayesian approach contains more information with respect to the interval estimate evaluated with the frequentist inference, because the final results is the probability density function after Bayesian updating instead of an interval that may be expected to encompass a fraction of the distribution of values that could be reasonably attributed to this quantity. In addition, it solves the criticism raised at §2.26 of VIM [4] and thoroughly analysed in the Guide to the Expression of Uncertainty in Measurement (GUM) [5], that is based on a mixture of frequentist and Bayesian thinking. In particular, the Type A (statistical) uncertainty evaluations are frequentist, whereas the Type B evaluations, using state-of-knowledge distributions, have a Bayesian attitude. In contrast, making the hierarchical model fully Bayesian implies, among other things, that a conventional objective Bayesian approach to Type A uncertainty evaluation for a number *n* of observations leads to the impractical consequence that *n* must be at least equal to 4, thus presenting a difficult for many metrologists [8]. In the interests of providing fuller information, the usage of Bayesian approach to overcome some limits of ANOVA modeling is not so new in metrology literature. As matter of fact, a Bayesian model has been already presented for the evaluation of between-bottle homogeneity studies in the production of reference and proficiency test materials [1].

The analysis starts in chapter §2 with a short introduction of the ISO 5725, staring at the relationships between the fundamental objects describing metrology accuracy, that are trueness and precision, under both repeatability and reproducibility conditions, and the parameters of an ANOVA general linear model. Then, the focus turns on how it was thought to use the ISO 5725 for the evaluation of the intermediate level of precision aimed at the metrological characterization of a whole product line. Intermediate means between repeatability and reproducibility conditions. So, these two words are stressed in order to clarify why the product line characterization needs a level

in between. The intermediate level does not describe the full variability conditions like reproducibility, but just the production-inherent variability, that is, under minimized room noise conditions. For this reason, it is said under homogeneity manufacturing conditions. Having clarified these concepts, the dissertation moves through the analysis of the ISO 5725 models and looks for which of them is the better option for the intermediate level of precision evaluation. Data examples are used to illustrate the approach and clarify the statistical models as ISO 5725 suggests. Two different datasets are used. The first is an example drawn from the ISO 5725-2 in order to have a reliable and authoritative benchmark against which to compare the Bayesian outcome. The second arises from a real industrial application case, drawn from the field of the equipments for thermo-mechanical testing of polymers, as the ISO 306 standard prescribes. The former is under reproducibility conditions, the latter under homogeneity conditions.

The core of the study is in chapter §3. Here, guided by metrology considerations and not by pure statistical criteria, it is drawn up the full Hierarchical Bayesian model. The first test bench for the model is the sulfur example illustrated in the ISO 5725-2. Then, the dissertation moves to the industrial application case of the accuracy evaluation for the LVDT displacement measurements of the ISO 306 testing equipment family. The hierarchical Bayesian model is gradually sophisticated in order to guarantee the coherence with the metrology mindset and a good fit with the ANOVA results. A couple of reasonable prior distributions are tested in order to provide for the lack of unintended sensitivity to apparently innocuous “non-informative” assumptions. In other words, the sensitivity of the Hierarchical Bayesian model is tested respect to different prior distribution assumptions, both proper and improper. In greater detail, for the sulphur accuracy experiment the sensitivity test is performed comparing the results achieved by two different improper prior assumptions, specifically, an uniform having infinite lower and upper bounds against a gamma having null shape and scale parameters. Instead, for the ISO 306 accuracy experiment, the sensitivity test is performed with improper against proper priors, specifically, uniform distributions having infinite against finite lower and upper bounds. Both sensitivity analyses showed that the prior assumptions do not make any difference, and this finding was welcomed. Looking at the final outcomes in terms of expected values of the hyperparameter posteriors for both cases we have a good fit with the ANOVA ones. The differences are of the same order of magnitude of the equipment resolution and this also was welcome. Overall, the Bayesian framework offers the great advantage of a better description of the accuracy. The manufacturers can achieve the maximum of the information from the experimental data and, in addition, they may leverage some prior information. Then, solving the Bayesian framework, the combination of the two sources of information, experimental data and prior knowledge, is split between the bias component and the precision components for both repeatability and reproducibility or homogeneity conditions. The former component is correlated to the concepts of trueness and systematic errors, whereas the other components are correlated to the concepts of precision and random errors. Bearing in mind, that we have the maximum of information, as combination of experimental data and prior knowledge, the full posterior probability density functions mostly confers a "natural" benefit upon the descriptions of trueness and precision, respect to those achieved with the ISO 5725-2 model. This is more evident as we turn from point to interval estimates. As a matter of fact, in the Bayesian framework we have a better decomposition of variability between the parameters describing accuracy and, so, the credibility interval for trueness is really able to encompass every laboratory or equipment expected value and the precision parameters are really engaged to describe only the sources of random variability. Lastly, exploiting exchangeability assumption, it was possible to design a unique and flexible mathematical model able to evaluate the accuracy of a whole new product line of testing equipments, according to the concepts expressed in the International Vocabulary of Metrology [4]. On the other hand, this descriptive coherence and the data information maximization are paid in a more complex accuracy model, that needs higher statistics skills. This chapter ends with some considerations about the validation metrics adopted in Instron, that are the first concrete acknowledgement for this study.

The last chapter §4 seeks to further extend the application scope of the Hierarchical Bayesian models in metrology, even if its deep reason is to show how easily it is possible to extend the number of hierarchical levels and to consider more complex and nested probability structures. It considers the proficiency testing scheme that is a standardized way of conducting inter-laboratory tests in order to determine the performance of participants for specific tests or measurements, and

to monitor their continuing performance independently several times [9]. So, once again, the benchmarking is provided by an international standard, the ISO 13528, that this time is applied to a simulated dataset for a proficiency inter-laboratory experiment. Then, the statistics, scoring participants performances according to the ISO 13528, are compared with ones achieved with the proposed Hierarchical Bayesian model. The test bench is the displacement accuracy experiment on the VST test equipments. Nevertheless, in order to be more consistent with a real proficiency test, where not all participants have the same model of testing equipment, the size of the experiment was extended through numerical random generation of data according to two different hypothesis for describing the performances of other product lines, that is, other manufacturer designs. Overall, the patterns between scores is still guaranteed. In other words, if we sort the participants as an increasing or decreasing function of their score statistics, we will achieve the same sorting for both approaches, ISO 13528 and the proposed Bayesian. Nevertheless, due to the Bayesian intrinsic shrinkage to the mean attitude (let's refer to [10, Cpv. 10.1]), the Bayesian scores are a bit more conservative with respect to the ISO 13528 ones. The Bayesian scores are closer each other, due to the force of attraction with respect to their center of gravity (overall mean hyperparameter), that is induced by the Hierarchical Bayesian framework.

Mandel's graphical consistency and Grubb outliers tests, suggested by the standard, are used also in the Bayesian Framework in order to check laboratory results. The working hypotheses of the conceptual framework behind the Mandel's *h* statistic and the Grubb test are still met also for the Bayesian outcome. Hence, for consistency techniques and outlier-detection tests based on relative deviation for grouped sets of each laboratory's observations, we keep on having a good consistency between the two approaches, frequentist and Bayesian. Then, the last flexibility proof of the hierarchical Bayesian framework brings to the end this applicative research: an additional hierarchical level is set for the proficiency Bayesian model in order to distinguish the accuracy performances of the three different product lines. This last probability structure provides a top level for the hyperparameter of the test method trueness, a second level for the hyperparameters of the specific-product lines trueness and precision under both repeatability and reproducibility conditions and a third level for the specific-laboratory bias. Thus the core purpose of the research is satisfied. We have the searched Hierarchical Bayesian model: a model able to describe the accuracy for more than one product line at the same time, without losing consistency with the accuracy description achieved by analyzing the subset specific for each product line separately. This single model is able to both catch the single laboratory proficiency scores and describe the accuracy performances of each of the product lines involved in the proficiency experiment. That means, the proficiency testing aims can be enlarged: not only evaluation of the laboratory performances but also rate the manufacturers designs performances. We would like to emphasize that the capability of guaranteeing the consistency of the accuracy results was just qualitatively proved and only for a balanced experiment plan. Further analysis need to be carried out before extending the conclusions as a general principle. Then, special attention should be paid to these sorts of cases, especially because between participants it could be possible to have someone neglectful with respect to maintenance plans or in general to the testing conditions. Even if these cases are extremely rare, because participants goal is to demonstrate the goodness of the results achieved in their laboratories for marketing purpose, the consistency techniques and the outlier detection test should ensure an adequate protection of the final rating outcome for manufacturers. The manufacturer ratings can be stated as the point estimates for their accuracy or as normalized statistics. We suggest the second option and we provide for two statistics normalized over the accepted range for displacement accuracy according the ISO 306 standard: the former able to rate the precision whereas the latter able to rate the combined effect of trueness and precision.

In conclusion, let me remark once again that the hierarchical Bayesian models can be used as an alternative to the frequentist ones, that are currently recommended in the international standards for the accuracy evaluation. The Bayesian framework keeps the outcome consistent with respect to the established methods. In addition, it offers an enhanced descriptive capability for accuracy evaluation, as the result is not just a point estimator or a confidence interval but the full probability density for each accuracy parameter, that are trueness, precision under repeatability, homogeneity and reproducibility conditions. This ought to really be appreciated by people who are not deep inside the metrology environment because they can clearly distinguish all the accuracy components.

2 ISO 5725 AS STANDARD FOR THE EVALUATION OF TEST EQUIPMENTS PERFORMANCES

2.1 INTRODUCTION TO ISO 5725

Just as the nearly universal use of the International System of Units (SI) has brought coherence to all scientific and technological measurements [3], international standards on the evaluation and expression of accuracy permit more clarity in the rating of the metrological performance. This may also be summarized as the wish to guarantee more clearness to the customers in order to provide a result to be readily understood and properly interpreted.

The most important standards on accuracy topic are the “International Vocabulary of Metrology” (VIM) [4] and the ISO 5725:1994 (Part 1:6) [11], [12], [7], [13], [14], [15]. The vocabulary sets the definitions and the associated terms for the system of basic and general concepts used in metrology. It is meant to be a common reference for people involved in planning and performing measurements, irrespective of the field of application. It is intended to promote global harmonization of terminology used in metrology [4]. At the upper level, it can be seen as a set of propositions that are mutually consistent. It serves as a premise or starting point for further reasoning and arguments, such as really ISO-5725 does, dealing with accuracy of measurements methods and results. Before going on, it is opportune to remark some of the definitions drawn from the VIM in order to set the steering compass for the next steps of this paper.

Measurement **accuracy** refers to the closeness of agreement between a **measured** quantity value and the **true** quantity value. So, the concept accuracy is not a quantity and is not given a numerical value. In addition, accuracy is related to the concepts of measurement trueness and precision.

Measurement **trueness** refers to the closeness of agreement between the average of **infinite** number of replicate **measured** quantity values and the accepted **reference** value, while **precision** refers to the closeness agreement between indications or measured quantity values obtained by **replicate measurements** on the same or similar objects under specified conditions. The need of taking into account the precision arises from the inner variability of the measurement procedure in a test methods, even if every covariate is assumed to be identical. This is due to unavoidable occurrence of **random errors** in each measurement process. As a matter of fact, the inner variability of test method results reflects the influence of many different variables, for example:

a. operator;
b. test equipment;
c. equipment calibration;
d. environment (temperature, humidity, air pollution, etc.);
e. time between measurements.

Instead, trueness is the expression of the component of measurement error that in replicate measurements remains constant or varies in a predictable manner, that is the **systematic error**. The causes of systematic measurement errors not always are known but a correction can be applied to compensate for them.

Two measure of trueness may be of interest: bias of the measurement method, that is the difference between the average of replicated indications obtained from all laboratories given a certain method and the accepted reference value, and the laboratory bias, that is the difference between the expectation of the replicated test results for a particular laboratory and the accepted reference value. In both cases a reference quantity value is required. On the other hand, precision does not require a reference value, as it involves only comparisons between replicate measurements of a quantity under specified conditions of **repeatability** or **reproducibility**. The former exists when independent test results are obtained with the same measurement procedure, same operators, same measuring system, same operating conditions and same location, and with replicate measurements on the same or similar objects over a short period of time. Reproducibility conditions exist when test results are obtained with the same measurement procedure in different laboratories with different operators using different equipments. Thus repeatability and reproducibility are the two extremes of precision, the former describing the minimum and the latter

the maximum variability of the results. Precision is normally expressed in terms of standard deviation.

Figure 2.1 shows a legible concept diagram of the relations between the accuracy and the other definitions termed in the preceding clauses. The accuracy is the “tympanon” standing on the two columns trueness and precision. So, the accuracy encompasses the concepts of trueness and precision but cannot be strictly numerically expressed. In order to get our hands dirty with numbers, we have to go down the steps of the temple. Actually, trueness and precision are related to systematic and random components of a measurement result, respectively. So, the former is expressed numerically by the bias of a finite number of replicated measurements with respect to a reference quantity value, whereas the latter is expressed by measures of imprecision, such as standard deviation. Precision also depends on the test conditions that can range from repeatability to reproducibility ones.

Figure 2.1: Visual presentation of the relations between the concepts about accuracy

As expected by the VIM [4, Par. §2.13, NOTE 2], ISO-5725 uses the two terms trueness and precision in order to assess the accuracy. It should be evaluated from a series of test results obtained through a collaborative study, that is an inter-laboratory experiment run under reproducibility conditions [11]. Such an inter-laboratory experiment is called an “accuracy experiment”.

Following ISO-5725, the statistical model explaining the measurand Y depends on the experimental design used to obtain the measurement. In general, each observed value y_{ik} of the measurand Y can be expressed as a realization of a general linear model, such as

$$y_{ik} = m + B_i + e_{ik} \qquad 2.1$$

where m is the overall mean, B_i (the notation is the same of ISO-5725 and denotes an observed value even if the capital letter can lead to think differently) is the laboratory component, e_{ik} is the residual random error. As a consequence of the hypotheses for the linear model, these random errors e_{ik} are the realizations of independent and normally distributed variables (with same null mean and same unknown variance) variables, supposed to incorporate all variation in the result variable due to factors other than laboratories. The overall mean m is the particular reference value tested and it is not necessarily equal to the true value or its accepted reference value μ. B_i is a quantity value depending on the categorical variable ”laboratory”, with $i = 1, \dots, p$ levels. Besides, for the *i*-th laboratory it is required a sample of $k = 1, \dots, n_i$ replicated measurements, so, k subscript denotes the position of the observation within i laboratory sample.

Generally, mathematicians follow the convention of denoting random variables by upper case italic letters and observed values by the corresponding lower case letters. Greek letters are used to denote parameters and corresponding lower case letters are used to denote estimators and estimates; occasionally the symbol ^ is also used for estimators and estimates. Therefore, according to the convention, the general linear model for the generic measurand Y stated in the equation 2.1, as it is in the ISO 5725-2, can be also expressed in terms of random variables as:

$$Y_{ik} = \mu + \beta_i + \varepsilon_{ik} \quad where \quad \varepsilon_{ik} \sim Norm(0, \sigma^2) \tag{2.2}$$

or in terms of its realizations and least square estimates as:

$$y_{ik} = m + b_i + e_{ik} \tag{2.3}$$

Careful readers will have already recognized in 2.1, 2.2 or 2.3 the general linear model of 1-way ANOVA [16]. Indeed, it can be easily proved that the estimate of the repeatability variance is the mean square error. As a matter of fact, the repeatability variance, according to ISO 5725-2 [12, Par. 7.4.5.1] is

$$s_r^2 = \frac{\sum_{i=1}^{p}(n_i - 1)s_i^2}{\sum_{i=1}^{p}(n_i - 1)} \tag{2.4}$$

or, if the number of replicated measurement is the same for each laboratory

$$s_r^2 = \frac{\sum_{i=1}^{p} s_i^2}{p}, \tag{2.5}$$

where $s_i^2 = \frac{\sum_{k=1}^{n}(y_{ik} - \bar{y}_{i\cdot})^2}{n-1}$ represents the sample variance for the measurements of each laboratory. Finally, we have

$$s_r^2 = \frac{\sum_{i=1}^{p}\sum_{k=1}^{n}(y_{ik} - \bar{y}_{i\cdot})^2}{p(n-1)} = \frac{SSE}{p(n-1)} = MSE = s_e^2 \tag{2.6}$$

SSE is the sum of square errors and MSE is the mean square error, as they are commonly presented in classical ANOVA literature [17], [18]. They both measure the variation that would be present within samples.

It is a bit more complicated to show the link between reproducibility variance according to ISO 5725-2 and the variance component model according to Fisher [18]. Let the reproducibility variance as ISO 5725-2 be the sum of the between-laboratory variance s_L^2 and the repeatability variance, that is

$$s_R^2 = s_r^2 + s_L^2 \tag{2.7}$$

Fisher's book (1925) shows that the mean square among groups is equal to [18]

$$MSA = \frac{SSA}{p-1} = ns_a^2 + s_e^2 \tag{2.8}$$

where the sum of squares among groups is $SSA = n\sum_{i=1}^{p}(\bar{y}_{i\cdot} - \bar{\bar{y}}_{\cdot\cdot})^2$ and s_a^2 is the variance component for the variability among groups. In a consistent way the ISO 5725-2 uses s_a^2 to evaluate the between-laboratory variance s_L^2 and consequently

$$s_L^2 = s_a^2 = \frac{MSA - s_e^2}{n} = \frac{\frac{SSA}{p-1} - s_e^2}{n} = \frac{\frac{n\sum_{i=1}^{p}(\bar{y}_{i\cdot} - \bar{\bar{y}}_{\cdot\cdot})^2}{p-1} - s_e^2}{n} = \frac{s_d^2 - s_e^2}{n}, \tag{2.9}$$

where s_d^2 is the mean square error among groups.

If n is not constant for each group, equation 2.9 becomes

$$s_L^2 = \frac{s_d^2 - s_r^2}{\bar{n}} \quad where \quad s_d^2 = \frac{\sum_{i=1}^{p} n_i\left(\bar{y}_{i\cdot} - \bar{\bar{y}}_{\cdot\cdot}\right)^2}{p-1} \quad and \quad \bar{n} = \frac{1}{p-1}\left[\sum_{i=1}^{p} n_i - \frac{\sum_{i=1}^{p} n_i^2}{\sum_{i=1}^{p} n_i}\right] \tag{2.10}$$

As a general rule [12, Par. 5.1.1], in the layout used in the accuracy experiment, samples from q batches of materials, representing q different levels of the test, are sent to p laboratories which each obtain exactly n replicate test results under repeatability condition. This type of experiment is called balanced uniform-level.

Hence, model 2.1 becomes:

$$y_{ijk} = m_j + B_{ij} + e_{ijk} \qquad 2.11$$

where $j = 1, \dots, q$ is the index for the level of the batch or, in general, the level of the reference value. Nevertheless, ISO 5725-2 prefers to treat each batch as different populations in order to not push for the hypothesis of same variance between different batches. It follows that for each *j*-batch an individual linear model is given, so we prefer the notation:

$$y_{ik}^{(j)} = m^{(j)} + B_i^{(j)} + e_{ik}^{(j)} \qquad 2.12$$

Hence, equations 2.6, 2.7 and 2.10 become respectively:

$$s_r^{2(j)} = \frac{\sum_{i=1}^{p}\sum_{k=1}^{n}\left(y_{ik}^{(j)} - \bar{y}_{i\cdot}^{(j)}\right)^2}{\sum_{i=1}^{p}\left(n_i^{(j)} - 1\right)} \qquad 2.13$$

$$s_R^{2(j)} = s_r^{2(j)} + s_L^{2(j)} \qquad 2.14$$

$$s_L^{2(j)} = \frac{s_d^{2(j)} - s_r^{2(j)}}{\bar{n}^{(j)}}$$

$$s_d^{2(j)} = \frac{\sum_{i=1}^{p} n_i^{(j)}\left(\bar{y}_{i\cdot}^{(j)} - \bar{\bar{y}}_{\cdot\cdot}^{(j)}\right)^2}{p-1} \qquad 2.15$$

where

$$\bar{n}^{(j)} = \frac{1}{p-1}\left[\sum_{i=1}^{p} n_i^{(j)} - \frac{\sum_{i=1}^{p} n_i^{2(j)}}{\sum_{i=1}^{p} n_i^{(j)}}\right]$$

We shall try to make this dissertation clearer by solving an example drawn from Annex B of ISO 5725-2 with R, the most powerful and flexible statistical software [19], enabling users to apply many statistical techniques such as ANOVA.

Laboratory i	Level j			
	1	2	3	4
1	0,71	1,20	1,68	3,26
	0,71	1,18	1,70	3,26
	0,70	1,23	1,68	3,20
	0,71	1,21	1,69	3,24
2	0,69	1,22	1,64	3,20
	0,67	1,21	1,64	3,20
	0,68	1,22	1,65	3,20
3	0,66	1,28	1,61	3,37
	0,65	1,31	1,61	3,36
	0,69	1,30	1,62	3,38
4	0,67	1,23	1,68	3,16
	0,65	1,18	1,66	3,22
	0,66	1,20	1,66	3,23
5	0,70	1,31	1,64	3,20
	0,69	1,22	1,67	3,19
	0,66	1,22	1,60	3,18
	0,71	1,24	1,66	3,27
	0,69	—	1,68	3,24
6	0,73	1,39	1,70	3,27
	0,74	1,36	1,73	3,31
	0,73	1,37	1,73	3,29
7	0,71	1,20	1,69	3,27
	0,71	1,26	1,70	3,24
	0,69	1,26	1,68	3,23
8	0,70	1,24	1,67	3,25
	0,65	1,22	1,68	3,26
	0,68	1,30	1,67	3,26

Table 2-1: Original data for sulfur content of coal experiment (Annex B ISO 5725-2)

Eight laboratories participated in the experiment for the determination of the sulphur content in coal, carrying out the analysis according to a standardized measurement method. The original data are given in Table 2-1, as percentage by mass. Looking at these data, it can be noticed that the experiment is not balanced and there are four different batches ("Levels *j*") of material.

The original data have several inadequacies for processing in R as are collected in Table 2-1, so the first step for data analysis with R is to arrange the dataset as shown in Table 2-2: a data frame of three variables, of which two are categorical and one is numerical.

	Laboratory	Batch	Sulfur
1	Lab1	LV1	0.71
2	Lab1	LV1	0.71
3	Lab1	LV1	0.70
4	Lab1	LV1	0.71
5	Lab1	LV2	1.20
6	Lab1	LV2	1.18
7	Lab1	LV2	1.23
8	Lab1	LV2	1.21
...	...	...	...
100	Lab8	LV2	1.22
101	Lab8	LV2	1.30
102	Lab8	LV3	1.67
103	Lab8	LV3	1.68
104	Lab8	LV3	1.67
105	Lab8	LV4	3.25
106	Lab8	LV4	3.26
107	Lab8	LV4	3.26

Table 2-2: R data.frame for sulfur content of coal observed values, corresponding to the R object named `data2`

The numerical variable sulfur contains the observed quantity values $y_{ik}^{(j)}$, where i is the index for the level of the categorical variable laboratory, k is the index for replicated measurements and j refers to the batch level. The batch variable has four levels and, so, according to ISO 5725-2 four ANOVA 1-way models have to be performed, one for each level of batch.

Once the batch level j is selected, the ANOVA 1-way for the model 2.1 can be performed in R using the `aov` function [20]. Table 2-3 summarizes the "aov" outcome for the first level of the factor Batch. Who wish can found all the R script details for the sulfur content experiment in the annex B §7.

```
#                            Df   Sum Sq     Mean Sq     F value   Pr(>F)
#Laboratory[Batch == "LV1"]  7    0.012555   0.0017935   7.849     0.000163 ***
#Residuals                   19   0.004342   0.0002285
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 2-3: Anova 1-way `summary(AOV)` output for the first level of Sulfur content of coal batches

The summary in Table 2-3 has all the values needed to estimate the precision variance under repeatability conditions or, if you prefer, the mean square error for residuals s_e^2, that is the value in the position identified by column "Mean Sq" and row "Residuals". So, the repeatability variance for the first batch of coal is:

$$s_r^{2(1)} = 0.0002285 \tag{2.16}$$

Then, in the column "Mean Sq" and row "Laboratory[Batch == "LV1"]", we have the mean square error among groups estimate:

$$s_d^{2(1)} = MSA = 0.0017935 \tag{2.17}$$

Another way in R for reading the output of the "aov" function is the usage of the "model.tables" function in order to report the means and the number of each i-laboratory group. For the first level of the batch factor this result is given in Table 2-4.

```
#Tables of means
#Grand mean
#
#0.6903704
#
# Laboratory[Batch == j[1]]
#        Lab1     Lab2     Lab3     Lab4     Lab5     Lab6     Lab7     Lab8
#        0.708    0.680    0.667    0.660    0.690    0.733    0.703    0.677
#rep     4        3        3        3        5        3        3        3
```

Table 2-4: Means and replications' number of each i-laboratory group for the first level of Sulfur content of coal batches as output of the function `model.tables(AOV, "means")`

Looking at the output in Table 2-4, the grand mean is the general mean as it is called in ISO 5725-2, that is:

$$\hat{m}^{(j)} = \bar{\bar{y}}_{..}^{(j)} = \frac{\sum_{i=1}^{p} n_i^{(j)} \bar{y}_{i\cdot}^{(j)}}{\sum_{i=1}^{p} n_i^{(j)}} \qquad 2.18$$

where $\bar{y}_{i\cdot}^{(j)}$ is the specific i-laboratory mean or cell mean according to ISO 5725-2, in formula:

$$\bar{y}_{i\cdot}^{(j)} = \frac{\sum_{k=1}^{n_i^{(j)}} y_{ik}^{(j)}}{n_i^{(j)}} \qquad 2.19$$

Having clarified the links between the ISO 5725-2 and the ANOVA output in R, for the first batch of sulfur content experiment, we have the grand mean equal to

$$\hat{m}^{(j)} = \bar{\bar{y}}_{..}^{(j)} = 0.6904 \qquad 2.20$$

and the cell means are equal to:

i	1	2	3	4	5	6	7	8
$\bar{y}_{i\cdot}^{(1)}$	0.708	0.680	0.667	0.660	0.690	0.733	0.703	0.677

Table 2-5: Specific *i-th* laboratory means for the first batch of sulfur content experiment

The last information in the Table 2-4 is the row called "rep" reporting the number of replicates for each i-laboratory, that means:

i	1	2	3	4	5	6	7	8
$n_i^{(1)}$	4	3	3	3	5	3	3	3

Table 2-6: number of replicates for each *i-th* laboratory of the first batch of sulfur content experiment

Due to the unbalanced experiment, the next step is about the evaluation of the average of replicates number $\bar{n}$, as stated in the equation 2.15. At the moment we are interested just to the level 1 of the factor batch, so with the input data in Table 2-6 the average number of replicates is:

$$\bar{n}^{(1)} = 3.3545 \qquad 2.21$$

Hence, according to the equation 2.15, it is possible calculate the between-laboratory variance s_L^2 as combination of the mean square error among group in 2.17, the repeatability variance in 2.16 and the average number of replicates in 2.21. It results:

$$s_L^{2\,(1)} = 0.0004665 \qquad 2.22$$

Finally, we can achieve the reproducibility variance s_R^2, as stated in the equation 2.14, achieving:

$$s_R^{2\,(1)} = 0.0006950 \qquad 2.23$$

The same operating flow can be followed for the other batches of materials. As a matter of fact, general mean and reproducibility and repeatability variances for the other batches can be easily achieved with the proper routine in R. The final result is summarized in Table 2-7, having

reproducibility and repeatability expressed as standard deviations. Even if they were achieved through the ANOVA model, the final outcome is equal to one in Table B.5 of ISO 5727-2 [12].

$Level_j$	$\hat{m}^{(j)}$	$s_r^{(j)}$	$s_R^{(j)}$
1	0.690	0.015	0.026
2	1.252	0.029	0.061
3	1.667	0.017	0.035
4	3.250	0.026	0.058

Table 2-7: Computed values [%] of $\hat{m}^{(j)}$, $s_r^{(j)}$ and $s_R^{(j)}$ for sulfur content of coal

Just in case of balanced experiment the replicates number n can be straight obtained from the degrees of freedom of residuals df_e and from the number p of laboratories involved. Actually, it has been proved [18] that the degree of freedom for residuals are equal to

$$df_e = p(n-1) \tag{2.24}$$

Now that the analogy between ISO 5725-2 and ANOVA 1-way was revealed and, so, it was clarified what means variance repeatability and reproducibility from general linear model point of view, let's see basic methods for the determination of the trueness drawn from ISO 5725-4 [13]. We remind that this part of ISO 5725 can be applied only if the accepted reference value can be established as conventional true value. In these cases bias values give quantitative estimates of the ability of a measurement method to give the correct (true) result. According to this definition in the basic model 2.1 the general mean m may be replaced by

$$m = \mu + \delta \tag{2.25}$$

where μ is the accepted reference value of the property being measured and δ is the bias of the measurement method. Even if it is by definition not straight correlated with the concept of closeness between the average of replicated measurements and the accepted reference value, trueness of the measurement method is normally expressed in terms of bias δ. So, the model becomes

$$y_{ik} = \mu + \delta + B_i + e_{ik} \tag{2.26}$$

Hence, the laboratory bias is given by:

$$\Delta_i = \delta + B_i \tag{2.27}$$

Summarizing, in this paragraph it has been shown the values describing, in quantitative terms, the ability of the method to give a correct result (trueness) and to replicate a given result (precision), both obtained by a collaborative study, that is an inter-laboratory experiment run under reproducibility condition.

2.2 PERFORMANCES EVALUATION OF A PRODUCT FAMILY WITH ISO 5725-2

Testing equipments performing measurements of physical quantities are expected to provide not only the measured quantity value but also some quantitative indications of the quality of the results. Actually, a measurement result is generally expressed as a single measured quantity value and an indication of its accuracy. Without such an indication, measurement equipments among different manufacturers and, in general, measurements cannot be compared.

In this scenario a slice of competition among manufacturers of test equipments is hardly played around the will of having on their own test equipments the best performances of accuracy for every influence quantity affecting the test method. As a matter of fact, accuracy is for the majority of the customers the criterion on which they decide from which manufacturer to buy a test equipment. For this reason, it is convenient to standardize the procedures describing the method for evaluating accuracy of each equipment family in order to ensure the maximum transparency. In other words, it shall exist a common language for the evaluation of the metrological performances of each test equipments family. Here, family means the whole product line of a specific testing equipment manufacturer for a specific test method, such as Instron Vicat test equipments (HV), TESA Vernier calipers, Mitutoyo gauge blocks, *et similia*. The will to control every single variable affecting the measurement result is unrealistic. So, the unavoidable occurrence of random errors can affect the measurement response of the test equipment with different magnitude as a function of the manufacturer design, mainly due to the engineering solutions on uncontrolled variables. Moreover, in some cases, systematic errors arise due to the engineering design or to the manufacturer calibration procedure. The differences over the quality of the metrological performances between test equipments of different manufacturer can fade without a unique metric for accuracy evaluation. For some product lines the problem of the metrological performances evaluation is well defined because there is an international standard aimed just at the calibration and verification of the specific test equipment. In these cases the standard provides accuracy classes, as function of stated metrological requirements, in which every single instrument has to belong to (e.g. the calibration and verification of static uniaxial testing machines follows the ISO 7500-1:2015 [5]). In other words, there is a measurement procedure to assess the measurement accuracy of the measured quantity values with respect to some specified classes. Otherwise, if the calibration and verification procedure there is no at all or the reference operating conditions [4, Par. 4.11] are not so clear, the manufacturer has to define its own procedure to describe the metrological performances of every influence quantity controlled by the instrument. So, in this shady regulatory environment the main manufacturer's needs are the choice of reference measurement standards, conditions and models in order to evaluate the metrological performances for the whole family of a new measuring instrument. This is not completely true because the ISO 5725 comes to rescue whenever there is the need to deal with accuracy.

Actually, the standard in its third part [7] also provides for **intermediate precision conditions** of measurement, out of the set of reproducibility conditions that include the same measurement procedure, same location, and replicate measurements on the same or similar objects over a limited period of time, but may include other conditions involving changes, for example different instruments of the same product line instead of different laboratories. In other words, at the end of the development project of a new product the engineering team can set up an experimental plan as the one provided for inter-laboratory experiment. The plan scope, rather than the evaluation of the test method accuracy, is the accuracy evaluation for the production homogeneity, that means just every influence quantity arising due to the tolerance design choices. So, the repeatability conditions are the same of the inter-laboratory experiment but with different instruments of the same manufacturer instead of different laboratories as in the reproducibility conditions. Moreover, in homogeneity conditions the location is always the same, the repetitions of measurements are carried out in a short time span and the instruments are all new. That is, the room noise and obsolescence effect are minimized with respect to the reproducibility condition in order to focus the analysis only on the production variability. So, this intermediate level sets the experimental conditions for the evaluation of the product line precision. As said, it takes into account just the variability introduced by the production processes. Different equipments of the same manufacturer are only nominally the same. In fact, due to the design tolerance approach, each single physical quantity of every single component of the equipment can vary within the limits set by the tolerance

design. The infinite number of combinations for these dimensions allows the inherent variability of the production process and, so, that affects homogeneity precision of the product family.

The repeatability, homogeneity and reproducibility conditions are summarized in Table 2-8. It is clear that the homogeneity conditions, those in which the manufacturers are interested, are intermediate between repeatability and test method reproducibility.

Repeatability	**Homogeneity**	**Reproducibility**
Same test method	Same test method	Same test method
Same test equipment	Different test equipments but of the same manufacturer	Different test equipments
Same operator	Different operators	Different operators
Same location	Same location	Different locations
Same condition of use	Same condition of use (new)	Different conditions of use
Repetition over a short period of time	Repetition over a short period of time	Repetition over a long period of time

Table 2-8: Experiment conditions

The problem of the accuracy description, as set, can be solved with the same models introduced in chapter §2.1 by simply changing the design of the experiment.

In conclusion, the experimental layout for the evaluation of the metrological performances of a product family provides for the use of samples from q batches of materials, representing q different levels of the test methods, measured with $\boldsymbol{p}$ **test equipments** of the same manufacturer which each obtains exactly n replicate test results under repeatability condition in a short time span. This type of experiment is still balanced uniform-level and the repeatability and homogeneity variances can be calculated using equations 2.13 and 2.14 respectively.

2.3 INDUSTRIAL APPLICATION CASE: VST TEST EQUIPMENTS PERFORMANCES

ISO 306:2013, or the equivalent ASTM D1525 – 17, deserves to be mentioned among the standards lacking a well defined procedure for the calibration and verification of the reference operating conditions [4, Par. 4.11]. This International Standard specifies four methods for the determination of the Vicat softening temperature (VST) of thermoplastic materials. Furthermore, this kind of test is becoming increasingly important as plastics continue to replace more traditional materials in many applications. Even if the test does not provide results for the usage in design calculations, it is very useful as a quality control or development tool.

The result is the temperature value at which thermoplastics begin to rapidly soften. In detail, these test methods are used in gauging the ability of polymers to retain their mechanical (in this case surface) properties at high temperatures. The test itself is performed by slowly raising the temperature of the medium, where the specimen stands, while applying a point load on the specimen surface. The combination of two possible values for both the control variables temperature rate (50 K/h or 120 K/h) and load (10 N or 50 N) gives the four conditions for the four test methods of ISO 306:2013. As function of one over four conditions of the test method, when the point load has penetrated 1 mm into the material, the test is ended and the temperature recorded. So, the final outcome of VST measurement method is the temperature at which the needle penetrates 1 mm under one of the four temperature rate and load combinations allowed by the standard.

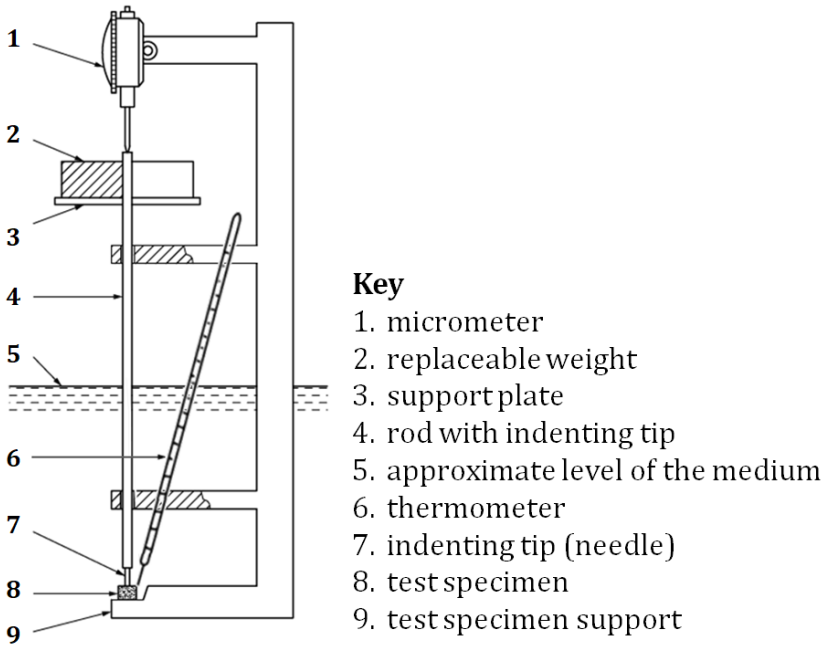

Figure 2.2: Schematic view of the apparatus for Vicat Softening Temperature determination

Of course, the VST standard is documented in sufficient detail to enable the manufacturer to design the measurement equipment. Briefly ([6] for more details), the test frame assembly (see Figure 2.2), consists of a heating bath, containing a medium (liquid or fluidized bed) in which the test specimen can be immersed to a depth of at least 35 mm. An efficient stirring mechanism shall be used to achieve the medium temperature homogeneity in the specimen area. The base of the frame (see point 9 in Figure 2.2) supports the test specimen under the indenting tip at the end of the rod, provided with a support plate or other suitable load-application device. The rod shall be able to move freely, with minimum friction, in a vertical direction. The surface of the indenting tip in contact with the specimen shall be flat and perpendicular to the axis of the rod, and free from burrs circular cross-section, having a diameter of 1,128 ± 0,008 mm. Weights are applied to the rod centrally, so that the total load, (10 ± 0,2) N or (50 ± 1 N), is applied to the test specimen. The heating equipment raises the temperature at a uniform rate of (50 ± 5) K/h or (120 ± 10) K/h, through a temperature control system (the temperature transducer shall be accurate to at least ± 0,5 K). A penetration-measuring device (calibrated micrometer dial gauge, LVDT «linear variable

differential transformer» or other suitable instrument to measure the penetration of the indenting tip into the test specimen to an accuracy of ± 0,01 mm) completes the VST measuring system.

The test specimens shall be between 3 mm and 6,5 mm thick and at least 10 mm square or of 10 mm diameter. Considering that the test ends when the needle is penetrated 1 mm, Instron design uses of LVDTs having 15 mm as range of the stroke nominal indication interval [4, Par. 4.5]. Being the stroke more than 4,5 mm, it is surely possible to cover the penetration measurement for every specimen thickness without mechanical adjustment.

The measurement procedure consists of the following steps:

1. position the specimen in the support under the indenting tip of the unloaded rod;
2. wait until the medium has a uniform temperature of 25 °C;
3. with the indenting tip still in position, lower the assembly with at least 2 specimens into the bath;
4. add a sufficient weight to the support plate, so that the total load on the test specimen is (10 ± 0,2) N for methods A50 and A120 or (50 ± 1) N for methods B50 and B120;
5. after 5 min, set to zero the reading of the penetration-measuring transducer;
6. increase the temperature at a uniform rate of (50 ± 5) K/h or (120 ± 10) K/h;
7. record the temperature of the bath when the needle has penetrated 1 mm;
8. express the VST of the material under test as the arithmetic mean of the VSTs of the specimens tested, unless the range of individual results exceeds 2 K. If the range is greater than 2 K, record the individual results and repeat the test a second time using an additional set of at least two specimens.

So, the control factors of the VST method are temperature, time (temperature rate), penetration and load measurements. For each of these four quantities the standard sets specific validation requirements. Nevertheless, their verification procedure is well defined just for temperature rate.

In the Instron design solution of the VICAT instrument system, the most tricky accuracy requirement to proof is that concerning the penetration transducer. This measurement is affected by many noise variables such as the quadrature error and non-linearity introduced by the LVDT [21], the digital processing errors, and the rod positioning errors due to the mechanical clearances aimed to ensure that the system is frictionless. In order to manage all these noise variables in the accuracy verification of the penetration measurement requirement, an experimental method was designed. The designed experiment consists in using traceable [4, Par. 2.41] gauge blocks instead of the specimen in order to refer the LVDT reading to an accepted reference value. The first step is zeroing the LVDT reading with respect to a first class 0 gauge block (according to ISO 3650:1998), thick t_{j0}. Then, the first gauge block is substituted with another one of the same class 1 mm less thick t_{j-1} and the LVDT reading $y_{ik}^{(j)}$ is registered. $y_{ik}^{(j)}$ is the observed quantity value of the accuracy experiment with respect to the reference accepted value:

$$\mu_{\text{set}}^{(j)} = t_{j0} - t_{j-1} = 1 \text{ mm} \qquad 2.28$$

where, even if t_{j0} shall be any value in the range $[3 \quad 6{,}5]$ mm, it was chosen to zero the LVDT in the positions corresponding to gauge block thicknesses of 6 mm, 5 mm, 4 mm and 3 mm. So,

$$t_{j0} \in \{6{,}5{,}4{,}3\} \text{ mm} \qquad 2.29$$

and the relative reference accepted values are summarized in Table 2-9.

j	μ_j [mm]	t_{j0} [mm]	t_{j-1} [mm]
1	1	6	5
2	1	5	4
3	1	4	3
4	1	3	2

Table 2-9: Reference accepted values

Therefore, the experiment plan has four ($q = 4$) reference values (as shown in Figure 2.3), called “jump”, having the same magnitude of 1 mm but differing for the zeroing point of the differential displacement measurement.

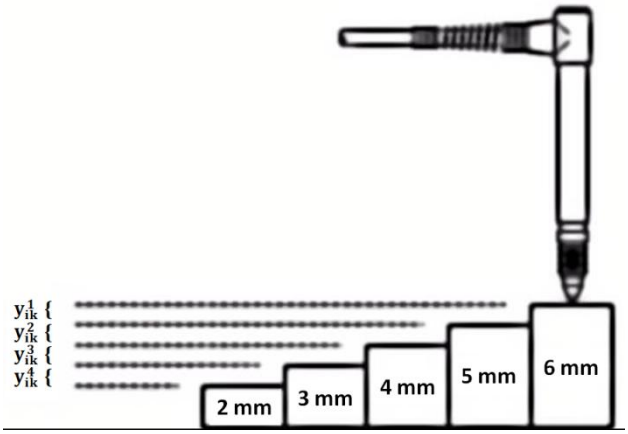

Figure 2.3: Schematic view of the verification method with blocks from 6 mm to 2 mm in order to have 4 jumps of 1 mm

Besides, the layout of this accuracy experiment implies a balanced plan, having four replicated measurement ($n = 4$) to explore repeatability conditions and eighteen ($p = 18$) different VST measuring systems (as the one in Figure 2.2) to explore the homogeneity conditions of the product family, as explained in Table 2-8. The eighteen apparatuses are identified with the LVDT serial number mounted in each. The zeroing is repeated for each measurement in order to avoid autocorrelation in the results due to the same zero. The results of the VST accuracy experiment are summarized in Table 2-10 and extensively reported in annex A 6.

	yijk	Jump	LVDT
1	0.996	6»5	155534
2	1.006	6»5	155534
3	0.997	6»5	155534
4	0.998	6»5	155534
5	0.999	5»4	155534
6	0.998	5»4	155534
7	1.012	5»4	155534
8	1.006	5»4	155534
9	0.999	4»3	155534
10	0.999	4»3	155534
11	0.998	4»3	155534
12	1.011	4»3	155534
13	1.001	3»2	155534
14	1.011	3»2	155534
15	1.013	3»2	155534
16	1.013	3»2	155534
...	.....	- -	.....
285	0.991	3»2	155601
286	0.990	3»2	155601
287	0.999	3»2	155601
288	0.995	3»2	155601

Table 2-10: R data.frame of the VST accuracy experiment ($y_{ik}^{(j)}$ are in millimetres), having the R object named `data3`

On this data frame we can apply the model for accuracy evaluation discussed in chapter 2.1, with the interpretation of the reproducibility variance as discussed in chapter 2.2, that is, what we called production homogeneity variance. In this accuracy experiment the VST testing equipment are all made by Instron. Each of them was randomly drawn from the mass production population and the time span for the whole experiment was kept as short as possible. For ease of discussion no further variables affecting the measurand (for example room temperature) are considered. Therefore, for each subset of dataset corresponding to one over four levels for the reference accepted value, that in this case plays the role of the factor batch of material, we can solve the general linear model of equation 2.12 through the "aov" R function. Beside, due to the fact that the design matrix associated to the equation linking the response and explanatory variables in the ANOVA model is not full rank, its least square solution needs an additional constraint equation. As default the "aov" R function uses the constraint $\mathrm{B}_1^{(\mathrm{j})} = 0$. Nevertheless, this choice is not easy to interpret. Whereas, it can be proved that, in the balanced experiment, forcing the constrain $\sum_{i=1}^{p} \mathrm{B}_{\mathrm{i}}^{(\mathrm{j})} = 0$ in the least square solution of the linear problem 2.12 (instead of $\mathrm{B}_1^{(\mathrm{j})} = 0$ used by default in R) the estimate $\hat{m}^{(j)}$ is straightly the average of the group means [22]:

$$\hat{m}^{(j)} = \frac{1}{np} \sum_{i=1}^{p} \sum_{k=1}^{n} y_{ik}^{(j)}. \qquad 2.30$$

This is exactly the general mean as provided by the ISO 5725-2 [12, Par. 7.4.4] and reported in equation 2.18. In this regard, we just specify into the "aov" R function the kind of contrast as "contr.sum".

As we learnt in section 2.1, the purpose of the accuracy description according to ISO 5725-2 can be achieved by running the ANOVA model for each subset of the dataset in Table 2-10 grouped by a specific level of the reference accepted value parameter. Since the experimental plan for the accuracy evaluation of the displacement measurement in VST apparatuses is balanced, we can also leverage the equation 2.24 and the considerations behind equation 2.30. The R script details can be found in annex C 8. After rounding, the final accuracy outcome is summarized in Table 2-11 with repeatability and homogeneity expressed as standard deviations.

j	Level$^{(j)}$	p$^{(j)}$	$\hat{m}^{(j)}$/mm	$s_r^{(j)}$/mm	$s_H^{(j)}$/mm
1	3»2	18	1.0017	0.0021	0.0052
2	4»3	18	0.9987	0.0033	0.0049
3	5»4	18	0.9994	0.0019	0.0051
4	6»5	18	1.0003	0.0029	0.0051

Table 2-11: estimates in millimeter of $\hat{m}^{(j)}$, $s_r^{(j)}$ and $s_H^{(j)}$ for VST accuracy experiment based on ANOVA 1-way model

For the purpose of this research it is better to replace the point estimates of the general means $\hat{m}^{(j)}$ with the lower and upper limits of their 95 % confidence intervals. Actually, it is possible to proof that:

$$\frac{\hat{m}^{(j)} - \mu^{(j)}}{s_r^{(j)} / \sqrt{np}} \sim t_{df=(n-1)p} \qquad 2.31$$

So, through the quantile function $Q(z)$ of equation 2.31, or if you prefer its inverse cumulative distribution function $\Phi^{-1}(z)$, we can specify the value at which the probability of the random variable is less than or equal to the given probability. That means we are looking for:

$$\left[\hat{m}_{lower}^{(j)} \quad \hat{m}_{upper}^{(j)}\right] = \hat{m}^{(j)} \pm c \frac{s_r^{(j)}}{\sqrt{np}} \quad : \quad Pr\left(\hat{m}_{lower}^{(j)} \le \mu^{(j)} \le \hat{m}_{upper}^{(j)}\right) = 0.95 \qquad 2.32$$

where $\Phi^{-1}_{t-student_{df=(n-1)p}}(0.975) = 2.005$.

Hence, replacing the column $\hat{m}^{(j)}/\mathrm{mm}$ with the upper and lower limits of the 95 % confidence interval, as stated in 2.32, we have the outcome shown in Table 2-12.

j	Level$^{(j)}$	p$^{(j)}$	$\hat{m}^{(j)}_{lower}/\mathrm{mm}$	$\hat{m}^{(j)}_{upper}/\mathrm{mm}$	$s_r^{(j)}/\mathrm{mm}$	$s_H^{(j)}/\mathrm{mm}$
1	3»2	18	1.0012	1.0022	0.0021	0.0052
2	4»3	18	0.9979	0.9995	0.0033	0.0049
3	5»4	18	0.9990	0.9998	0.0019	0.0051
4	6»5	18	0.9996	1.0010	0.0029	0.0051

Table 2-12: estimates of $\hat{m}^{(j)}_{lower}$, $\hat{m}^{(j)}_{upper}$, $s_r^{(j)}$ and $s_H^{(j)}$ for VST accuracy experiment based on ANOVA 1-way model

Lastly, we can arrange this final outcome in terms of bias, as expressed in 2.25, and expanded uncertainties for both repeatability $U_{\mathrm{r}}^{(\mathrm{j})}$ and homogeneity $U_{\mathrm{H}}^{(\mathrm{j})}$ conditions. A view of the Uncertainty approach is detailed in the *Guide to the expression of uncertainty in measurement (GUM)* that focuses on the mathematical treatment of measurement uncertainty through an explicit measurement model under the assumption that the measurand can be characterized by an essentially unique value [3]. Briefly, the guide requires to describe the combined standard uncertainty [4, Par. 2.31] as a proper linear combination of the individual standard uncertainties associated with the input quantities in the measurement model. From the combined standard uncertainty it is possible to obtain the expanded uncertainty as the product of the combined standard uncertainty and a coverage factor depending on the type of probability distribution of the output quantity in the measurement model and on the selected coverage probability.

A possible choice for the measurement model associated with this application is the following:

$$D = Y - \left(T_{j+2} - T_{j+1}\right) \cdot (1 + \alpha \cdot \Delta T) \qquad 2.33$$

where:

- all input quantities are assumed to be independent;
- D is the difference between the measured value Y and the reference accepted value $\left(T_{j+2} - T_{j+1}\right) \cdot \alpha \cdot \Delta T$;
- T_{j+2} is the thickness of the first gauge block where we zeroing the transducer;
- T_{j+1} is the thickness of the second gauge block with respect to which we observe the displacement;
- α is the linear expansion coefficient of the gauge blocks material;
- ΔT is the temperature difference respect to the room temperature, conventionally assumed 20°C.

An additional common assumption provides for considering the measured value Y as affected by two uncertainty components: the apparatus resolution $u(Y_{\mathrm{d}})$ and the amount of all noise factor $u(Y_r)$, so that $u^2(Y) = u^2(Y_{\mathrm{d}}) + u^2(Y_{\mathrm{r}})$.

Hence, the uncertainty budget comprises many components. Out of all these components just the amount of all noise factors Y_d may be evaluated by **Type A evaluation of measurement uncertainty** from the statistical distribution of the quantity values from the series of measurements in Table 2-10 and it can be characterized by their precision standard deviations under conditions either repeatability or homogeneity. The other components may be evaluated by **Type B evaluation of measurement uncertainty**, and so, they also can be characterized by standard deviations, evaluated from probability density functions based on experience or other information. This inharmonious mixture of frequentist statistics for the type A uncertainty evaluation and Bayesian attitude for type B uncertainty evaluation could be fixed working in a fully Bayesian framework as we propose in the next chapters.

Going back to the accuracy experiment, during the measurements the test temperature was under control and monitored with resolution $0{,}2°C$. The reference accepted values were provided by certified gauge blocks, with thickness stated as having expanded uncertainty equal to $0{,}1\ \mu\mathrm{m}$.

The LVDT was zeroed at the beginning of each measurement in order to avoid autocorrelation between the replications. The resolution of the displacement transducer is equal to $1\ \mu\text{m}$ and is provided in digital way. Besides, for the purpose of this research we can consider the linear expansion coefficient $\alpha = 10.8 \cdot 10^{-6}\ \text{K}^{-1}$ as a constant. Combining all these inputs through the measurement model, we can achieve the uncertainty budget summarized in Table 2-13. This outcome is just for the level corresponding to the reference accepted value of the "jump 5»4", that is the level with the minimum repeatability precision (see Table 2-12), where the column labels follow the glossary indication of JCGM:100 [3] Annex J and the row labels follow the symbols used in the measurement model of equation 2.33. We need to pay attention to the meaning of the *i* subscript that in the uncertainty budget table does not indicate the specific laboratory but the source of variability.

	i	Source	pdf	$\overline{X}_i$	$u(x_i)$	ν_i	c_i	$u_i^2(y)$
$Y_r^{(3)}$	1	Measurements	T-student	0.9994 mm	0.0019 mm	54	1	3.61e-06 mm^2
$Y_d^{(3)}$	2	Resolution	Uniform	0 mm	3e-04 mm	8	1	9e-08 mm^2
$T_4^{(3)}$	3	Block[4]	Normal	4 mm	5e-05 mm	60	1	2.5e-09 mm^2
$T_5^{(3)}$	4	Block[5]	Normal	5 mm	5e-05 mm	60	1	2.5e-09 mm^2
$T^{(3)}$	5	Temperature	Uniform	0 °C	0.0577 °C	8	1.08e-05 mm/°C	3.88e-13 mm^2

Table 2-13: Uncertainty budget for the 3rd level of the reference accepted value factor, that corresponds to "jump 5»4"

The combined (squared) uncertainty $u_c^2(y)$, associated with the output estimate y of the response variable D of the measurement model 2.33, is achieved as the sum of the variance components for all the variability source terms. For the components in Table 2-13 the combined standard uncertainty is

$$u_c(y)^{(3)} = \sqrt{\sum_{i=1}^{5} u_i^2(y)} = 0.00192\ \text{mm} \qquad 2.34$$

Even if we chose the (minimum) repeatability condition, the additional sources of variability are negligible with respect to those inherent in the test equipments, that are expressed in terms of their repeatability precision standard deviation.

It follows that the attempt to provide explanation in terms of ever smaller entities leads to neglect the additional sources of variability and characterize the dispersion of the values that could reasonably be attributed to the displacement measurand Y just in terms of its precision standard deviation under either repeatability conditions or homogeneity conditions. Once again, we remark that this assumption is well supported by the outcome in Table 2-13.

So, the expanded precisions are achieved as the product of a coverage factor $k = 2$, in order to have 95 % confidence level, and repeatability or homogeneity standard deviations. Of course, the underlying assumption is that the PDF for the measurand is a Gaussian (which is reasonable, since precision deals with random effects). We named the expanded precisions $U_{\text{r}}^{(\text{j})}$ and $U_{\text{H}}^{(\text{j})}$ respectively, in order to highlight the degree closeness to the expanded uncertainty, even if they are not exactly the same quantity [23]. Then, according to 2.25, the trueness can be described by the interval limits for the general means estimate minus the reference accepted values. Consistently we can apply the same considerations to the outcome in Table 2-12 in order to describe the accuracy in terms of its bias intervals and extended precisions under both condition of repeatability and homogeneity, as it is shown in Table 2-14.

j	Level$^{(j)}$	p$^{(j)}$	$\hat{\delta}^{(j)}_{lower}$/mm	$\hat{\delta}^{(j)}_{upper}$/mm	$U_r^{(j)}$/mm	$U_H^{(j)}$/mm
1	3»2	18	0.0012	0.0022	0.0042	0.0104
2	4»3	18	-0.0021	-0.0005	0.0066	0.0098
3	5»4	18	-0.0010	-0.0002	0.0038	0.0102
4	6»5	18	-0.0004	0.0010	0.0058	0.0102

Table 2-14: evaluated values of $\hat{d}^{(j)}_{lower}$, $\hat{d}^{(j)}_{upper}$, $U_r^{(j)}$ and $U_H^{(j)}$ for VST accuracy experiment based on ANOVA model

From the manufacturer point of view the homogeneity expanded precision $U_{\mathrm{H}}^{(\mathrm{j})}$ gives a good confidence that the production of VST measurement systems meets the ISO 306:2013 requirement about penetration measurement accuracy. Actually, rounding to the most significant digit gives $U_{\mathrm{H}}^{(\mathrm{j})} = 0{,}01\ \mathrm{mm}$. On the contrary, the results in terms of bias intervals that do not encompass the null bias leave serious doubts, so that there is enough evidence for rejecting the null hypothesis of null bias. This is an unexpected result from the manufacturer's point of view.

For the scope of this research, the outcome in Table 2-11 is the benchmark against which to compare the results obtained with the hierarchical Bayesian model defined in the next chapters. In this regard, we would like to remind the readers that the repeatability variance is the mean square error of the ANOVA 1-way linear model (see equation 2.6) whereas the reproducibility/homogeneity variance is the sum of repeatability variance and between-laboratory variance, that describes among groups variability. Let's keep in mind this relation for better understanding the probability structure assumed for the Bayesian formulation of the accuracy description problem.

3 HIERARCHICAL BAYESIAN MODEL FOR THE EVALUATION OF TEST EQUIPMENTS PERFORMANCES

3.1 INTRODUCTION TO HIERARCHICAL BAYESIAN MODELS

Thomas Bayes (1702-1761), from whom Bayes' Theorem takes its name, is believed to have been the father of several results in probability stimulated by his seminal paper [24], in which he wrote: "*Given the number of times in which an unknown event has happened and failed: Required the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named*". In this statement the usage of the two words "chance" and "probability" is not just a matter of insisting on a point but the conscious intention to describe uncertainty as the result of two nested processes: the "probability" defines the underlying risk of the event occurring, and it is affected by the "chance", as the riskiness related to the proper event description. In other words, Bayes used "probability" to refer to uncertainty concerning an observable event and "chance" to refer to uncertainty concerning that probability [10]. Thus, the first crucial step taken in Bayesian Statistics is to consider any parameter θ, describing a generic probability distribution $p(\theta)$, as a random variable too (therefore, theoretically we should use capital and lower case letters, but this is not generally done in this study just for convenience and practicality). Second, as a consequence of the Bayes' Theorem, the probability distributions in the Bayesian framework apparently express opinions rather than being solely based on formal data analysis like in the frequentist approach. Actually, Bayes' Theorem tells us how to learn from new evidence, but inevitably requires the specification of what we thought before that evidence was taken into account. So, specification of one or more such "priors" is an additional responsibility of the analyst. The choice of the appropriate prior distribution depends on the degree of knowledge of the specific subject matter and a strong degree of judgment. For these reason, often a Bayesian model is said a full probability model in which both observed realizations x and parameters θ are random variables, having the joint probability distribution:

$$f_{X,\Theta}(x,\theta) = \pi(\theta)\mathcal{L}(\theta;x). \qquad 3.1$$

In expression 3.1, $\pi(\theta)$ is the prior distribution for the parameters, representing the state of knowledge or personal belief about the values of θ before taking into account the observed data, and $\mathcal{L}(\theta;x)$ is the likelihood function, providing a credible description of the model parameters given the observed data. From another point of view, the likelihood gives an indication of how much the data contribute to the probability of the parameter value or of the model. Per Bayes theorem, the likelihood can be multiplied by a prior probability and then normalized, to give a posterior probability. Therefore, as a consequence of the Bayes theorem, for random variables described by probability density function, we can achieve the conditional probability $\pi(\theta|x)$, that is assigned to the parameters θ after the relevant evidence or background of the observed data $X = x$ is taken into account. This is said called posterior probability and is given by

$$\pi(\theta|x) = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)} = C(x)\pi(\theta)\mathcal{L}(\theta;x). \qquad 3.2$$

So, this conditional distribution for θ is a function of the selected prior, updated in the light of new relevant data (evidence) expressed by the likelihood $\mathcal{L}(\theta;x)$. The constant $C(x)$ normalizes to one the total probability.

Leaving aside the philosophical crusades between the Fisherian and Bayesian schools, distinguished by a greater "purely" objectivist view of the former against a subjectivist attitude of the latter, we need to acknowledge that Bayesian methods tend to be inherently more complex than classical analyses, and the numerical data summarization is even worst. It is obvious that the objectivistic climate of the late 19th century, combined with the complexity of calculation, initially slowed the spread of Bayesian techniques. In the last decade of the 20th century the tones of religious war, that have characterized the dawn of the frequentist / Bayesian debate have faded. This is also due to the increased flexibility and computability of the Bayesian models, thanks to the development of new computational methods and to the IT progress. Among the several mathematicians which promoted the Bayesian paradigm, the Italian Bruno de Finetti (1906-1985) deserves a special mention. Today, where no closed-form algebraic formula for posterior

probability distribution is available, the quantities of interest can be calculated using computer simulation techniques, known as Markov Chain Monte Carlo (MCMC) methods [16]. And even the most complex hierarchical Bayesian models can be explored.

What does Hierarchical Bayesian model mean? These models are invariably used for fitting data from multiple "units", for example, different patients, geographical areas, schools, hospitals, etc. [10], and provide a formal framework for analysis with a complexity of structure that matches the system being studied. They are flexible, that is, all sources of correlation and heterogeneity can be incorporated in a modular fashion, in particular by the introduction of unit-specific parameters. The posterior distribution for such a model is still proportional to the likelihood multiplied by the prior, but the prior distribution, of all unknown parameters, is decomposed into an exchangeability assumption for the unit-specific parameters, for example laboratories and equipments, and a prior for the population parameters [10].

The flexibility of MCMC techniques enables the application of Bayesian models in areas that Bayes had never dreamed about, such as metrology. Actually, since the release in 2008 of the supplement 1 of the GUM [25] the Monte Carlo method to approximate a probability density function has found widespread use in the evaluation of measurement uncertainty [26]. Besides, GUM-S1 appears to move the guidelines towards the Bayesian point of view, in the sense that produces a probability distribution that shall encode one's state of knowledge about the measurand. In contrast to a Bayesian uncertainty analysis, however, Bayes' Theorem is not applied explicitly. Instead, a distribution is assigned for the input quantities which is then 'propagated' through a model that relates the input quantities to the measurand. The relation between the Bayesian uncertainty analysis and the application of the GUM and its supplements has been already investigated [27] and it has been shown that under certain assumptions both analyses yield the same results but this is not true in general [26], [28], [29]. Current GUM-S1 recommendations for evaluating uncertainty of measurement are based on the Bayesian interpretation of probability distributions as encoding the state of knowledge about the quantities to which those distributions refer. Conceptually, this allows to overcome the criticism often levelled at the GUM, that it is based on a mixture of frequentist and Bayesian thinking. In particular, the GUM Type A uncertainty evaluations are frequentist, whereas the Type B evaluations, using state-of-knowledge distributions, have a Bayesian attitude. In contrast, making the GUM fully Bayesian, as shown in [8], leads to the impractical consequence that the number n of observations must be at least equal to 4, which introduces a difficulty in many measurements. Moreover, it still remains a bit of more complexity in the model formulation and in the following calculation.

As declared in the introduction, §1, this study carries on the research flow about the Bayesian models applied to metrology. The main objective of this paper is to show an alternative formulation of the ISO 5725 models for accuracy evaluation, based on Hierarchical Bayesian models in order to provide more consistent use of measurement data and prior information. The proposed Hierarchical Bayes model is then applied to the datasets introduced in chapter 2, demonstrating that there are no evident discrepancies between the two formulations, but rather improvement accuracy description thanks to the better consistency with the concepts of trueness and precision as declared in the VIM.

3.2 HIERARCHICAL BAYESIAN MODEL FOR ACCURACY EVALUATION

The present research arises from the desire to design a unique and flexible mathematical model able to evaluate the accuracy of a whole new product line of test equipment, according to the concepts expressed in the International Vocabulary of Metrology [4]. The experimental plan is the same introduced in section §2.2, that provides for q different levels of the test methods, measured with p test equipments of the same manufacturer which each obtains exactly n replicate test results under repeatability condition. Using a medical analogy, the replications made in repeatability conditions with the same test equipment can be thought as technical replicates, while the measurements on different instruments under reproducibility conditions such as biological replicates. The problem thus framed lends itself well to representation through hierarchical Bayesian models. In this framework, the first step is to express the qualitative structure of the model, that is, all the assumptions concerning the joint relationship between all known and unknown quantities. Graphs are an effective aid to communicate qualitative conditional independence structure and, actually according to BUGS syntax [10], it is possible to arrange variables in order to reduce "globally" complex models into a set of fairly simple "local" components through direct acyclic graph representation. So, graphical representation of the hierarchical models can be helpful in understanding the structure of the proposed metrological model. Nevertheless, before going ahead, some preliminary simplifying assumptions are needed. First of all, we will take into account separately the different levels of the test methods in order not to introduce dependence effects as for ISO 5725-2 (see section §2.2). The assumption of "exchangeability", which is shown to be equivalent to assuming the observations were independent and identically distributed from a distribution with unknown parameters, where those parameters are given a prior distribution[10], is made about the unit-specific parameters (in this case the laboratory-specific parameters). This is equivalent to assuming that they arise from a common "population" distribution whose parameters are unknown and assigned proper prior distributions [30]. Therefore, the laboratory-specific parameters are similar but not identical.

As concerns notation, the parameter variables are denoted with the Greek letters and the corresponding lower case letters denote their estimates, the sole exception being the estimates of the specific-laboratory means where we still use the capital letter in order to be consistent with the ISO 5725-2 glossary.

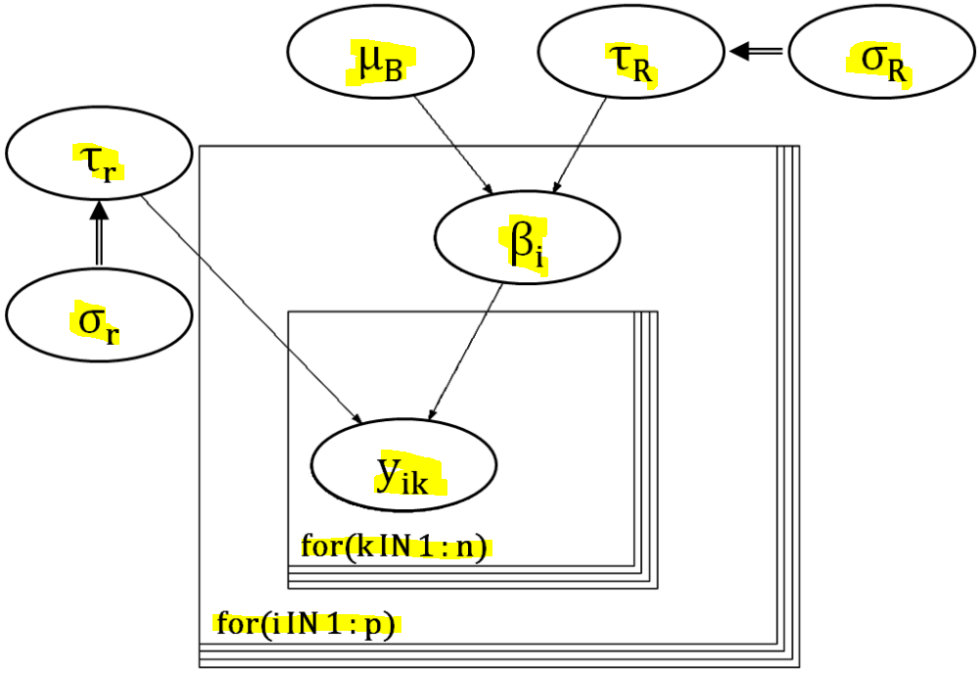

Figure 3.1: Doodle graph of the Hierarchical Bayesian model for the accuracy evaluation

Under the previous assumptions, the proposed hierarchical Bayesian model for accuracy evaluation is graphically represented in Figure 3.1. It is worth noting the lack of the subscript j in the observed values y_{ik}. This is due to the first assumption of not having any dependence on the

levels of the reference accepted value. So, the model is going to be applied q-time, for each of data subsets grouped by reference accepted value level. Then, each y_{ik} is assumed to be independent and identically distributed from a distribution of parameters β_i and σ_r. β_i are the laboratory-specific parameters and, in turn, are exchangeable from a distribution of parameters μ_B and σ_R. Looking at the taxonomy of μ_B, σ_r and σ_R the meaning behind is clear enough:

- μ_B was thought in order to describe the method trueness;
- σ_r was thought in order to describe the method precision under repeatability conditions and it represents the mean standard error by analogy with ANOVA models;
- σ_R was thought in order to describe the method precision under reproducibility conditions.

A proper prior distribution has to be assigned to these last three parameters. In the GUM-S1 context, the prior distribution expresses the state of knowledge on the method. If such information exists, it is expressed in the form of a prior, otherwise a non-informative prior is needed and it depends on the statistical models encoded by the likelihood function. The posterior distribution corresponding to them is then obtained after the Bayesian updating by analytical or numerical marginalization. It is now evident the metrological completeness of this model: we have three distinct probability distributions, each of them to describe the concepts of trueness, precision under repeatability conditions and precision under reproducibility conditions respectively.

The assumptions behind the accuracy model in Figure 3.1 can be written as follows:

$$\begin{aligned} Y_{ik}^{(j)}|\beta_i^{(j)},\sigma_r^{(j)} &\sim i.i.d.\ \ \pi\left(\beta_i^{(j)},\sigma_r^{(j)}\right) \\ \beta_i^{(j)}|\mu_B^{(j)},\sigma_R^{(j)} &\sim i.i.d.\ \ \pi\left(\mu_B^{(j)},\sigma_R^{(j)}\right) \\ \mu_B^{(j)} &\sim \pi(\theta_u) \\ \sigma_r^{(j)} &\sim \pi(\theta_w) \\ \sigma_R^{(j)} &\sim \pi(\theta_z) \end{aligned} \qquad 3.3$$

Nevertheless, whichever probability distributions are assigned to these quantities, it is not easy to calculate in closed form the joint probability distribution that encodes the final state of knowledge about the measurand and its input quantities according to the independence structure expressed in Figure 3.1. Besides, the search for some specific prior allowing a closed form for the joint probability distribution lies beyond the scope of this applied research. So, at the moment, the main pain point of this formulation is the need to use “advanced” software (not Microsoft Excel for example) in order to perform MCMC simulations for the numerical solution. As explained in the previous section §3.1, we also know that the international bureau of metrology has already open to the usage of Monte Carlo methods with the GUM-S1 [25]. Therefore, we do not think that the following discussion may be affected in terms of authority, although we will keep on moving just in the MCMC methods’ plan.

In the following calculations we use OpenBUGS [31], a popular software providing MCMC methods to analyze complex statistical Bayesian models. This software uses Gibbs sampling [32], [33] and the Metropolis algorithm [34] to generate a Markov chain by sampling from full conditional distributions. Using OpenBUGS, the users must just specify the model to be run and to load data and initial values for a specified number of Markov chains. Furthermore, the R2OpenBUGS package provides the tools to call OpenBUGS [35] directly in R, where it is possible to work with the results, for example to create posterior predictive simulations or graphical displays of data. Embedding all the routines in R was extremely useful to process the bunch of data sets at each level of the reference accepted value factor, thus saving a lot of time and efforts.

3.3 MODEL CHECKING AND COMPARISON WITH ISO 5725-2 RESULTS

The strength of the Bayesian graphical modelling techniques is the way they can represent the typical complexity of real data. Nevertheless, the conclusions of a Bayesian analysis are always conditional on the assumed probability model, so we need to be reassured that our assumptions are a reasonable approximation to reality, even though we do not generally believe any model is actually "true". Many aspects of an assumed model might be questioned: observations that don't fit, the distributional assumptions, qualitative structure, link functions, which covariates to include, and so on. Hence, we straightaway test the accuracy of our Bayesian model with the level 1 for Sulfur content of coal experiment (dataset in Table 2-2). The result in Table 2-7 is the benchmark against which to compare the Bayesian outcome.

As said in the previous section 3.2, the accuracy model in Figure 3.1 was implemented using the R library "R2OpenBUGS" [35], whereas all calculations were executed with the calculus engine of "OpenBUGS". The full R script is reported in the ANNEX B §7 in order to give the possibility of checking the reliability of the results or leveraging it for further analysis.

Using R2OpenBUGS the first step to solve is to arrange the dataset according to the OpenBUGS required format. So, we first extract the subset for the level one of the reference accepted value factor. This subset of data includes missing values. Actually, ISO 5725 already takes into account the possibilities that sometimes a laboratory may carry out and report more than n test results officially specified and, in other cases, some of the test results may be missing, for example because of loss of sample or mistake in performing the measurement [12]. Looking at Table 2-1, according to Rubin classification [36], we are dealing with missing responses distributed at random. Under this assumption, it is not necessary to specify a model for the missing data mechanism in order to make valid inference about the parameters [10]. The missing data in BUGS are denoted by Na and from a Bayesian perspective are treated as additional unknown quantities for which a posterior distribution can be estimated. If we simply denote the value as missing (Na) in the dataset, then BUGS will automatically generate values from its posterior predictive distribution and inferences on the parameters will be as if we had deleted that response [10]. So, we have to complete the vector of the measured quantity values $y_{\mathrm{ik}}^{(1)}$ with the information about the missing values, in order to have a balanced number of replicates, missing data included, for each laboratory. Finally, we can prepare the data input list for the BUGS engine. The list includes the two scalar quantities $p = 8$ and $a = 5$. The former is the number of laboratory-specific parameters and implies the total number of exchangeable cycles. The latter represents the number of replicates inside each cycle. This is the result of the two hierarchical level of the model 3.3 in Figure 3.1. Besides, the inputs list provides the $p \times q$ data matrix for the measured values $y_{\mathrm{ik}}^{(1)}$. The structure of the input list is shown in Table 3-1 where we can see that for most laboratories there are two missing values. This suggests that the minimum requirement of the inter-laboratory accuracy experiment of sulfur content was just 3 replicated measurements for each level of the materials batches.

```
> print(data)
#$p
#[1] 8

#$a
#[1] 5

#$y1ik
#     [,1] [,2] [,3] [,4] [,5]
#[1,] 0.71 0.71 0.70 0.71   NA
#[2,] 0.69 0.67 0.68   NA   NA
#[3,] 0.66 0.65 0.69   NA   NA
#[4,] 0.67 0.65 0.66   NA   NA
#[5,] 0.70 0.69 0.66 0.71 0.69
#[6,] 0.73 0.74 0.73   NA   NA
#[7,] 0.71 0.71 0.69   NA   NA
#[8,] 0.70 0.65 0.68   NA   NA
```

Table 3-1: inputs list of the Hierarchical Bayesian model 3.3

The hierarchical model for accuracy evaluation as proposed in Figure 3.1 needs to be completed with assumptions about the prior probability distributions. Historically, it is common to distinguish between “non-informative” (sometimes called “objective”) and “informative” priors. The former are preferred when there is not explicitly knowledge on the data source, as for dataset in Table 3-1, where measurements come from a study done by someone else for ISO 5725-2. The term “non-informative” is misleading: it should be better to refer to them as “improper” priors. Actually, all priors contain information at least about the prior family distribution (for example, normal, gamma or uniform). Leveraging the metrology mindset, the following assumption is arosen for the prior distributions of the of the accuracy probability model expressed in equation 3.3

$$\begin{aligned} Y_{ik}^{(j)}|\beta_i^{(j)},\sigma_r^{(j)} &\sim i.i.d.\quad Normal\left(\beta_i^{(j)},\sigma_r^{(j)}\right) \\ \beta_i^{(j)}|\mu_B^{(j)},\sigma_R^{(j)} &\sim i.i.d.\quad Normal\left(\mu_B^{(j)},\sigma_R^{(j)}\right) \\ \mu_B^{(j)} &\sim Normal(0,\infty) \\ \sigma_r^{(j)} &\sim Uniform(0,\infty) \\ \sigma_R^{(j)} &\sim Uniform(0,\infty) \end{aligned} \qquad 3.4$$

Going into details, all the observed quantity values $Y_{\mathrm{ik}}^{(\mathrm{j})}$ are supposed to be normally distributed according to their population parameters, the specific-laboratory mean $\beta_{\mathrm{i}}^{(\mathrm{j})}$, assumed as an expression of potentially systematic laboratory bias, and the residual standard deviation $\sigma_{\mathrm{r}}^{(\mathrm{j})}$, that it is expected to play the role of the precision under repeatability conditions. Similarly, the specific-laboratory parameters $\beta_{\mathrm{i}}^{(\mathrm{j})}$ are supposed to be normally distributed according to their population hyperparameters, the general mean parameter $\mu_{\mathrm{B}}^{(\mathrm{j})}$ and the reproducibility standard deviation $\sigma_{\mathrm{R}}^{(\mathrm{j})}$. In this case due to the lack of reliable prior information, a normal improper prior was assumed for the general mean hyperparameter $\mu_{\mathrm{B}}^{(\mathrm{j})}$. Instead, the standard deviation is always thought as a value in $\mathbb{R}^+$, so the easiest assumption is the improper uniform probability distribution over the whole real positive line. So, in terms of priors, the lack of knowledge on data is translated in flat distributions, having an infinite standard deviation in the case of the normal prior distribution for general bias, and with the upper limit approaching infinity for both the uniform distributions of the standard deviations. Now, it should be clear why “improper” is preferred to “non informative”. In many circumstances this is not a problem, as an improper prior can still lead to a proper posterior distribution [10].

The Bayesian reader can have a bit of confusion due to the use of the standard deviation as parameter for the normal distribution. As a matter of fact, Bayesian mathematicians prefer to parameterize the normal with the mean and the inverse of the variance

$$\tau = \frac{1}{\sigma^2}. \qquad 3.5$$

The τ parameter is called “precision” and here it was deliberately not used in order to avoid misunderstanding with the metrology concept of precision, that is described with a measurement of imprecision such as the mean square error in 2.6. Nevertheless, OpenBUGS, such as other software for Bayesian computations, uses the Bayesian convention for the normal distribution parameterization, and, so, in the following scripts the reader is going to find the transformation according to the equation 3.5 and the normal improper prior distribution is going to have the scale parameter 0 instead of ∞. In the doodle graph representation of Figure 3.1 the transformation of the equation 3.5 is depicted with the double arrow symbol.

Then, the MCMC hierarchical accuracy model was written using the OpenBUGS code, given in the script 3.6. Working with finite computational capability the improper prior distributions are managed using finite values that reasonably approximate the infinite condition for the scale parameters. In other words we represent these distributions using proper distributions with a “large” variance, where “large” means five orders of magnitude less than the magnitude of the observed values for approximating the zero and five orders of magnitude more for approximating the infinite. Lastly, before running the MCMC method to obtain the sequence of random samples

describing the posterior probability distribution, we need to initialize the Metropolis algorithm. In this respect, tuple of elements for the model parameters are randomly drawn according to the prior distribution assumptions in 3.4 and in order to avoid elements from the tails, that could increase the computation time, proper prior assumptions are taken for the selection of the starting points. “Proper” in this case means having a scale parameter of the same magnitude of the resolution of the sulfur content of coal measurements.

```
model {
for (i in 1:p)
{
for(k in 1:a) {
xik[i,k] ~ dnorm (B[i], tau.r)
}
B[i] ~ dnorm (mu.B, tau.R)
}
mu.B ~ dnorm (0.0, 1.0E-6)
tau.R <- pow(sigma.R, -2)
sigma.R ~ dunif (0, 1000)

tau.r <- pow(sigma.r, -2)
sigma.r ~ dunif (0, 1000)
}
```

3.6

Now, the MCMC simulation can start putting together the data, the randomly selected starting points, the model describing the probability structure in 3.6 and other control parameters depending on the specific numerical framework like number of total iterations per each Markov Chain. This model must be stored in a separate file, e.g. “sulfur1.txt” as it is named in ANNEX B: R script for the sulfur content experiment, in an appropriate directory, say C:/PhD/Thesis, in order to be inserted into the MCMC routine. Then, through the R function `bugs()` of R2OpenBUGS package, that is able to run an entire OpenBUGS session, the Hierarchical Bayesian problem is solved and MCMC posterior samples, summary statistics and other related information are available for further analysis.

The posterior data for sulfur content $y_{ik}^{(1)}$ are shown in Figure 3.2. Arranged as boxplot grouped by laboratory index, the posterior data fit well the raw initial values. Actually, posterior boxplots overall provide a good encompass of the experimental measurements.

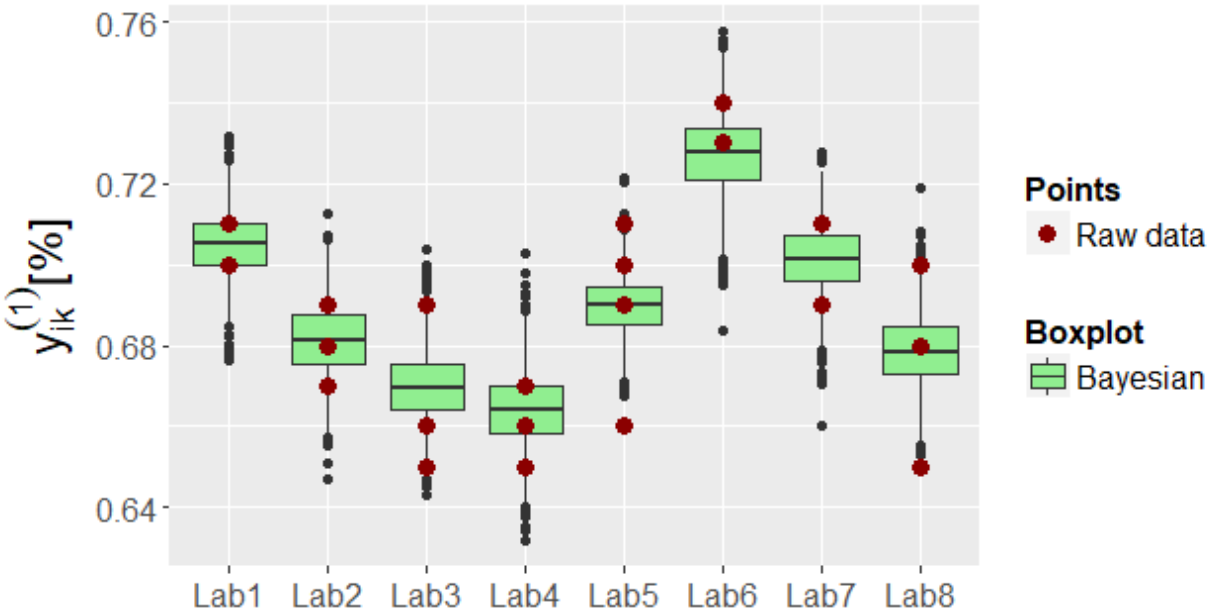

Figure 3.2: Red points representing specific-laboratory measurements (raw data) plotted with respect to green boxplots describing posterior data for each specific-laboratory parameter as a result of MCMC solution

Under the assumption of exchangeable laboratory-specific parameters, each posterior borrows strength (or variance) from the others, via their joint influence on the estimation of the underlying population parameters. There are not evident differences in the size of interquartile range of the boxplots for the posterior specific-laboratory samples, whereas it seems more evident a magnitude

of the variability between laboratories different from the raw data dispersion. Besides, in the Bayesian framework the extreme values tend to get pulled in towards the population mean. This is more clear in Figure 3.3 where the green points are the expected values $b_i^{(j)} = E(\beta_i^{(j)})$ for the posterior distributions of the laboratory-specific parameter, while the red ones are the sample means of the replicated values for each laboratory and the dash-dot line is the general mean. To improve description, it is used dodging that preserves the vertical position of the points while adjusting the horizontal position so that the paired points for each laboratory are more clear.

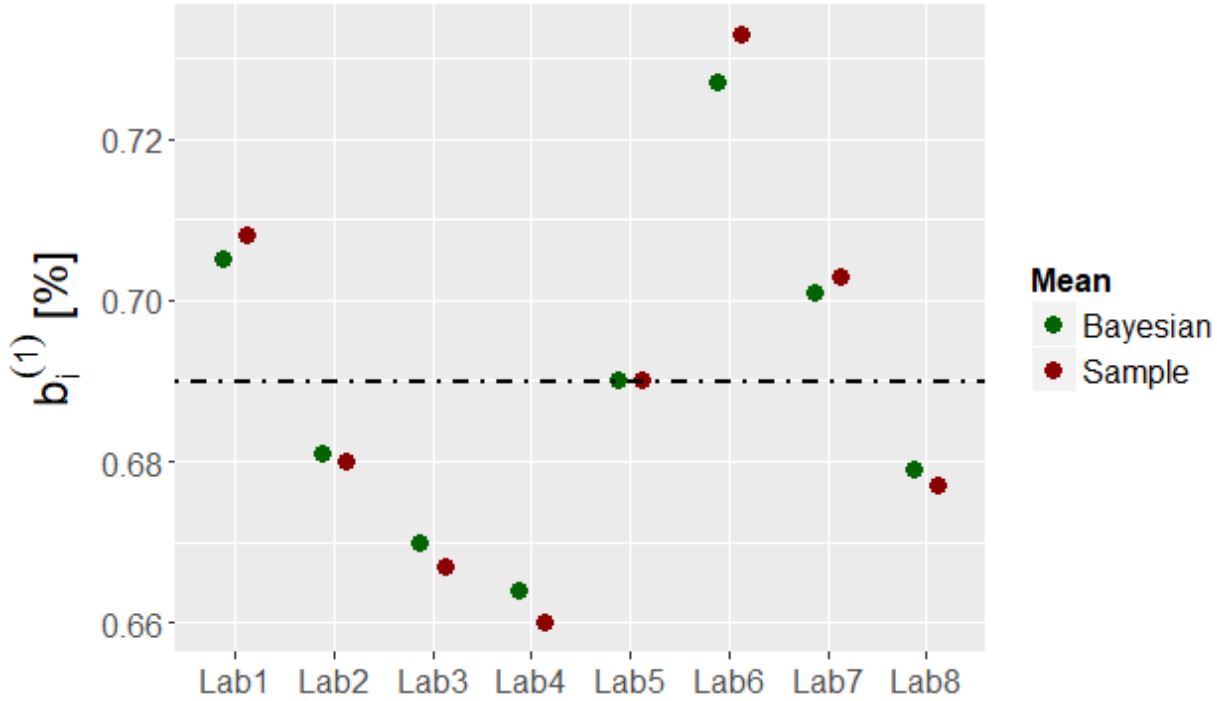

Figure 3.3: Comparison on the first batch of material between the expected values of the Bayesian posterior probability distributions for each laboratory-specific parameter and the corresponding sample means

After the overview of the posterior distributions for each laboratory-specific parameter B_i, in both terms of boxplot in Figure 3.2 and expected value in Figure 3.3, we can analyze the posterior description for accuracy parameters, that are $\mu_B^{(1)}$, $\sigma_r^{(1)}$ and $\sigma_R^{(1)}$, as explained in section §3.2. The results are shown in Figure 3.4 below.

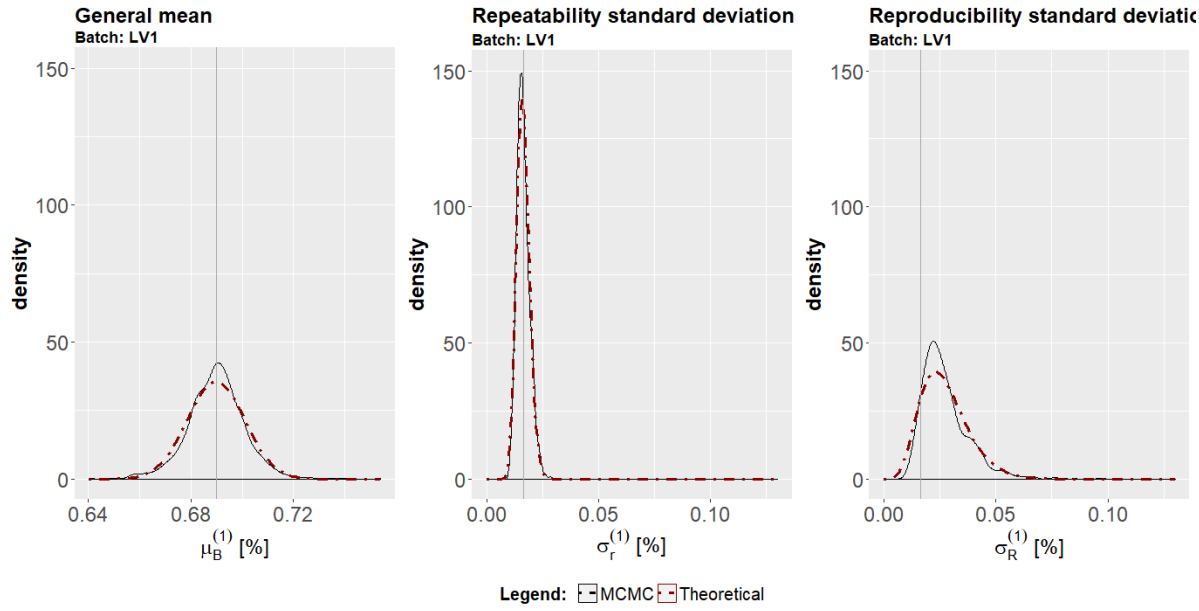

Figure 3.4: Marginal posterior distributions for general mean $\mu_B^{(1)}$, repeatability standard deviation $\sigma_r^{(1)}$ and reproducibility standard deviations $\sigma_R^{(1)}$

Hence, from the marginal posterior distributions of $\mu_{\mathrm{B}}^{(1)}$, $\sigma_{\mathrm{r}}^{(1)}$ and $\sigma_{\mathrm{R}}^{(1)}$ we can calculate the expected values in order to compare them with the ANOVA outcomes. Actually, comparing the Bayesian output in Table 3-2 with the first row of Table 2-7 we see a good fit between the two models. The difference, at most equal to 0.001 %, is small enough to be considered negligible, being an order of magnitude less than the resolution.

$\mathrm{Level_j}$	$\mathrm{m_B^{(j)}}$	$\mathrm{s_r^{(j)}}$	$\mathrm{s_R^{(j)}}$
1	0.690	0.016	0.027

Table 3-2: Expected values [%] of the marginal posterior probability distributions for the accuracy parameters according to probability structure formulated in the equation 3.4

The point estimators according to ISO 5725-2 yield single-valued results whereas the great advantage of the hierarchical Bayesian accuracy model according to the probability structure of the equation 3.4 is the possibility to approximate the whole marginal probability distribution for each parameter in Figure 3.1. This consistent with the metrology attitude of the last decade [37]. This attitude is fully satisfied in the sulfur content of coal experiments when we apply the hierarchical Bayesian accuracy model, as the final outcome is not just the expected value but the whole marginal posterior distribution for each of the parameter describing the accuracy (see Figure 3.4). With respect to the prior distribution assumptions, after the Bayesian updating, only the general mean $\mu_{\mathrm{B}}^{(1)}$ still has the same probability distribution family (Normal), even if it is no more improper but having mean and standard deviation as expressed in equation 3.7. Instead, the repeatability and reproducibility parameters are definitely no more uniform distributed: their marginal posterior distributions seem more like Gamma with shape and scale as reported in equation 3.7.

$$\begin{aligned} \mu_B^{(j)} &\sim Normal(0.690, 0.011) \\ \sigma_r^{(j)} &\sim Gamma(32.8, 2007) \\ \sigma_R^{(j)} &\sim Gamma(6.5, 237) \end{aligned} \qquad 3.7$$

Actually, Figure 3.4 shows a good agreement between the MCMC marginal posterior distributions (see black solid line) and their approximated distributions (see red dot-dashed line, called theoretical) as Normal for $\mu_B^{(1)}$ and Gamma for $\sigma_r^{(j)}$ and $\sigma_R^{(j)}$ having parameters as expressed in equation 3.7. Instead, after the likelihood updating, the marginal posterior distributions for both the standard deviations of repeatability and reproducibility are no more improper uniform distributions but they are effectively approximated (see Figure 3.4) by gamma distributions having parameters as reported in 3.7.

Thanks to the marginal posterior distributions, the related credibility interval can clearly be associated to each of the accuracy parameters. So, repeatability precision $u_{\mathrm{r}}^{(1)}$ we can construct a 95 % unilateral credibility interval derived from its marginal posterior, that is:

$$u_r^{(1)}: \Phi_{\sigma_r^{(1)}}\left(u_r^{(1)}\right) = P\left(\sigma_r^{(1)} \le u_r^{(1)}\right) = 0.95. \qquad 3.8$$

Similarly, a 95 % unilateral credibility interval for the reproducibility precision $u_{\mathrm{R}}^{(1)}$ can be derived from its marginal posterior distribution, that is:

$$u_R^{(1)}: \Phi_{\sigma_R^{(1)}}\left(u_R^{(1)}\right) = P\left(\sigma_R^{(1)} \le u_R^{(1)}\right) = 0.95 \qquad 3.9$$

The description of trueness is more complex, because for the sulfur content experiment we do not have a reference accepted value. Here we can assume that the reference value is equal to the expected value $m_B^{(1)} = E\left[\mu_B^{(1)}\right]$ of the general mean parameter and it does not introduce additional variability, that is equivalent to assume that bias is zero. So, we propose to express trueness with the 95 % bilateral credibility interval $\left[\hat{\delta}_{\mathrm{lower}}^{(1)} \quad \hat{\delta}_{\mathrm{upper}}^{(1)}\right]$ derived from the posterior marginal distribution of the general mean shifted to zero, that is

$$\hat{\delta}^{(1)}_{\text{lower}} \coloneqq \Phi_{\mu_B^{(1)} - E\left[\mu_B^{(1)}\right]}\left(\hat{\delta}^{(1)}_{\text{lower}}\right) = P\left(\mu_B^{(1)} - E\left(\mu_B^{(1)}\right) \leq \hat{\delta}^{(1)}_{\text{lower}}\right) = 0.025$$
$$\hat{\delta}^{(1)}_{\text{upper}} \coloneqq \Phi_{\mu_B^{(1)} - E\left[\mu_B^{(1)}\right]}\left(\hat{\delta}^{(1)}_{\text{upper}}\right) = P\left(\mu_B^{(1)} - E\left(\mu_B^{(1)}\right) \leq \hat{\delta}^{(1)}_{\text{upper}}\right) = 0.975 \quad . \qquad 3.10$$

In conclusion, the sulfur content measurement method can be expected to have a systematic error lying in the interval $\left[\hat{\delta}^{(1)}_{\text{lower}} \quad \hat{\delta}^{(1)}_{\text{upper}}\right]$ % and a random error encompassed between $\pm u_r^{(1)}$ % under repeatability conditions or $\pm u_R^{(1)}$ % under reproducibility conditions. The final outcome is reported in Table 3-3 below.

Level_j	$\hat{\delta}^{(j)}_{\text{lower}}$	$\hat{\delta}^{(j)}_{\text{upper}}$	$u_r^{(j)}$	$u_R^{(j)}$
1	-0.02	0.02	0.021	0.046

Table 3-3: Trueness limits and expanded precision under both repeatability and reproducibility conditions as mass percentage for Sulfur content of coal experiment according to probability structure formulated in the equation 3.4

Even if the conclusions of the Bayesian analysis, given the assumed probability model in 3.7, seem to be in agreement with the treatment given in ISO 5725-2, we would further assess the probability structure assumptions by carrying on a sensitivity analysis on the effect of the number of iterations for each Markov chain in the R function `bugs()` and on the choice for priors.

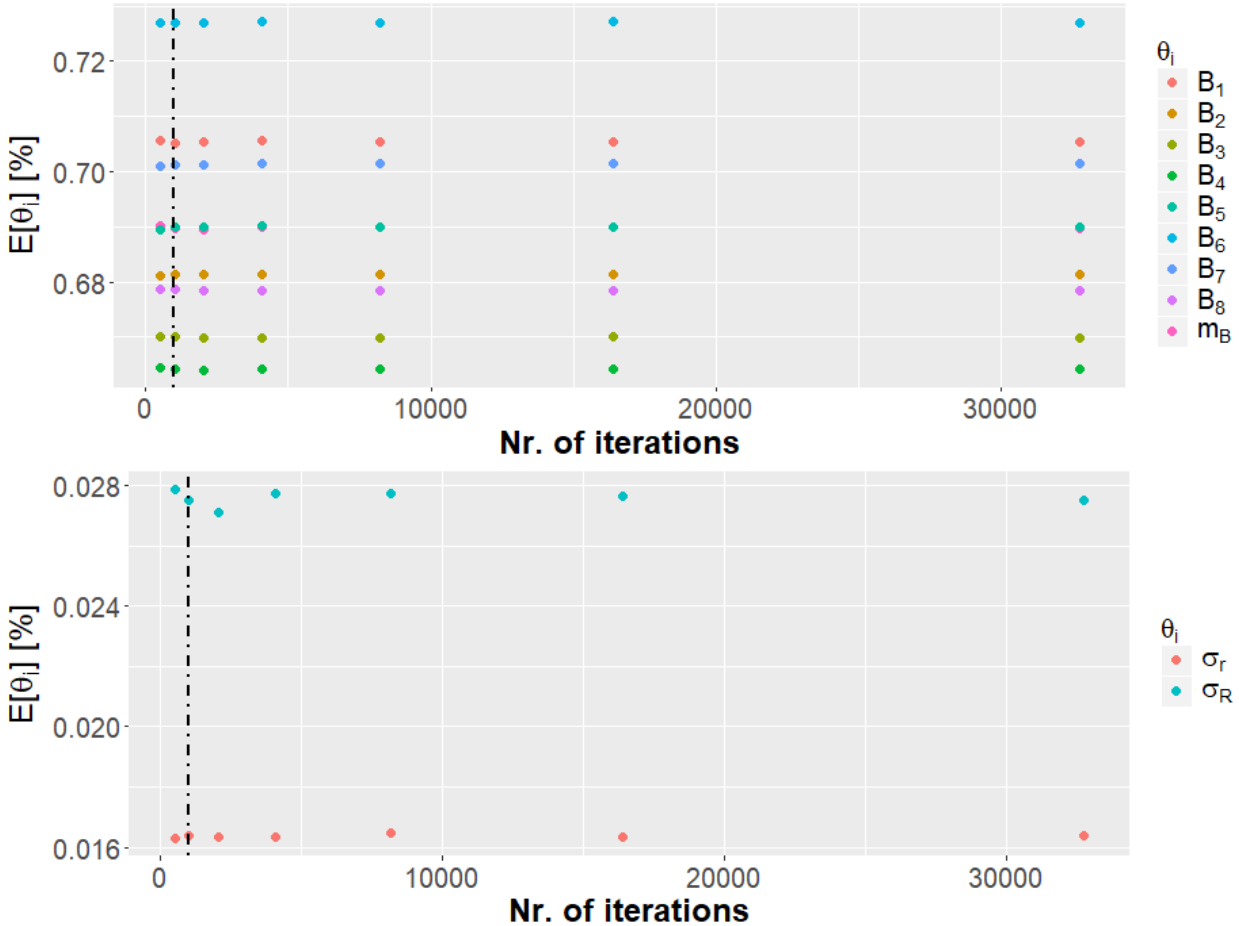

Figure 3.5: Sensitivity analysis on the number of iterations

The results of the sensitivity analysis on the number of iterations in each Markov chain are shown in Figure 3.5. In this case study, 1000 iterations (see the dot-dashed vertical line) are sufficient to provide stable results for the expected values of all parameter distributions except for the reproducibility standard deviation, where it may be better to run at least 5000 iterations. This is not a big issue, considering that even with 2^{15} iterations the MCMC simulation takes just seconds.

Given that there is no such a thing as the true prior, sensitivity analysis to alternative prior assumptions is vital and should be an integral part of Bayesian analysis. As a matter of fact, neglecting uncertainty about the choice of model has been called a “quiet scandal” in statistical practice [38], and, moreover, drawing conclusion on the basis of a single selected model can conceal the possibility that other plausible models would give different results. In any case it is good practice to explore the influence of different choices for the priors of the unknown scale parameters. Considering Gelman suggestions [39], in model 3.6 a uniform prior was used on both scales of the standard deviations, over a “large” range. Looking at the results in Figure 3.4, it seems reasonable to test the improper gamma prior on both the inverse-variance of repeatability and reproducibility. Under this assumption for priors the accuracy probability structure becomes:

$$\begin{aligned} Y_{ik}^{(j)}|\beta_i^{(j)},\sigma_r^{(j)} &\sim i.i.d.\quad Normal\left(\beta_i^{(j)}, 1/\sqrt{\tau_r^{(j)}}\right) \\ \beta_i^{(j)}|\mu_B^{(j)},\sigma_R^{(j)} &\sim i.i.d.\quad Normal\left(\mu_B^{(j)}, 1/\sqrt{\tau_R^{(j)}}\right). \\ \mu_B^{(j)} &\sim Normal(0,\infty) \\ \tau_r^{(j)} &\sim Gamma(0,0) \\ \tau_R^{(j)} &\sim Gamma(0,0) \end{aligned} \qquad 3.11$$

The corresponding script for the hierarchical accuracy model in its MCMC formulation is

```
model {
for (i in 1:p)
{
for(k in 1:a) {
xik[i,k] ~ dnorm (B[i], tau.r)
}
B[i] ~ dnorm (mu.B, tau.R)
}
mu.B ~ dnorm (0.0, 1.0E-6)
tau.R ~ dgamma (0.001, 0.001)
tau.r ~ dgamma (0.001, 0.001)
}
```

3.12

This model must be stored in a separate file, e.g. “sulfur3.txt” as it is named in the ANNEX B §7, in an appropriate directory, say C:/PhD/Thesis, in order to be got into the MCMC routine. Finally, running the MCMC solver through the R function `bugs()` of R2OpenBUGS package we can draw the expected values of the general mean, the repeatability and reproducibility parameters. The outcome is summarized in Table 3-4.

$Level_j$	$m_B^{(j)}$	$s_r^{(j)}$	$s_R^{(j)}$
1	0.690	0.018	0.028

Table 3-4: Expected values [%] of the marginal posterior probability distributions for the accuracy parameters according to probability structure formulated in the equation 3.11

The difference from the point estimators is negligible for repeatability and reproducibility standard deviations, even if there is a slight increase for the precision parameters. The same conclusion can be drawn looking at the 95 % credibility intervals for trueness, as expressed in equation 3.10, and at the expanded precisions in both repeatability and reproducibility conditions, summarized in Table 3-5. In both precision estimates the 5 % unilateral credibility interval is used instead of the 95 %, as in equations 3.8 and 3.9, because the chosen precision parameters in the last probability structure 3.11 are the reciprocal of the variances.

$Level_j$	$\hat{\delta}^{(j)}_{lower}$	$\hat{\delta}^{(j)}_{upper}$	$u_r^{(j)}$	$u_R^{(j)}$
1	-0.025	0.025	0.024	0.054

Table 3-5: Trueness limits and expanded precision under both repeatability and reproducibility conditions as mass percentage for Sulfur content of coal experiment according to probability structure formulated in the equation 3.11

Given the set of candidate models, we would like to say something about which is "better", or even "best". For this reason the effective number of parameters p_D and Deviance Information Criterion DIC are particularly useful [10]. The minimum for both p_D and DIC is intended to identify the "better" model. Actually, the `bugs()` function already report these parameters and their outcome is reported in Table 3-6.

Model	Precision prior	$p_D^{(1)}$	$DIC^{(1)}$
3.04	Uniform	7.908	-139.5
3.11	Gamma	8.658	-135.9

Table 3-6: Model comparison statistics

We note that DIC can legitimately be negative. So, the smaller for both p_D and DIC are in 3.4. Besides, it is important to highlight that only differences between models in DIC are important, and not the absolute values. Hence, we select the model 3.4.

Oncethe model is chosen we replicate the Bayesian calculation to evaluate the accuracy of the remaining three levels for the batch factor. Thanks to the integration with R through the package R2OpenBUGS, repeating the calculations is not time-consuming (see the whole script in annex B §7). The more relevant results are summarized in the following figures.

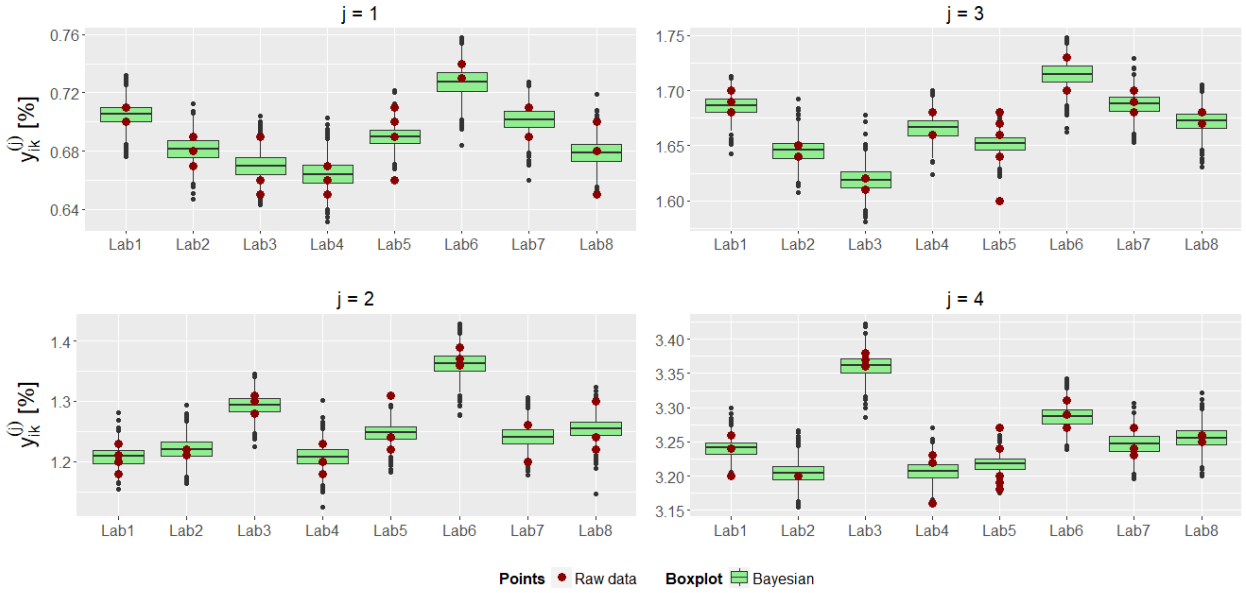

Figure 3.6: Comparison between red points representing specific-laboratory measurements (raw data) plotted with respect to green boxplots describing posterior data for each specific-laboratory parameter as result of MCMC solution

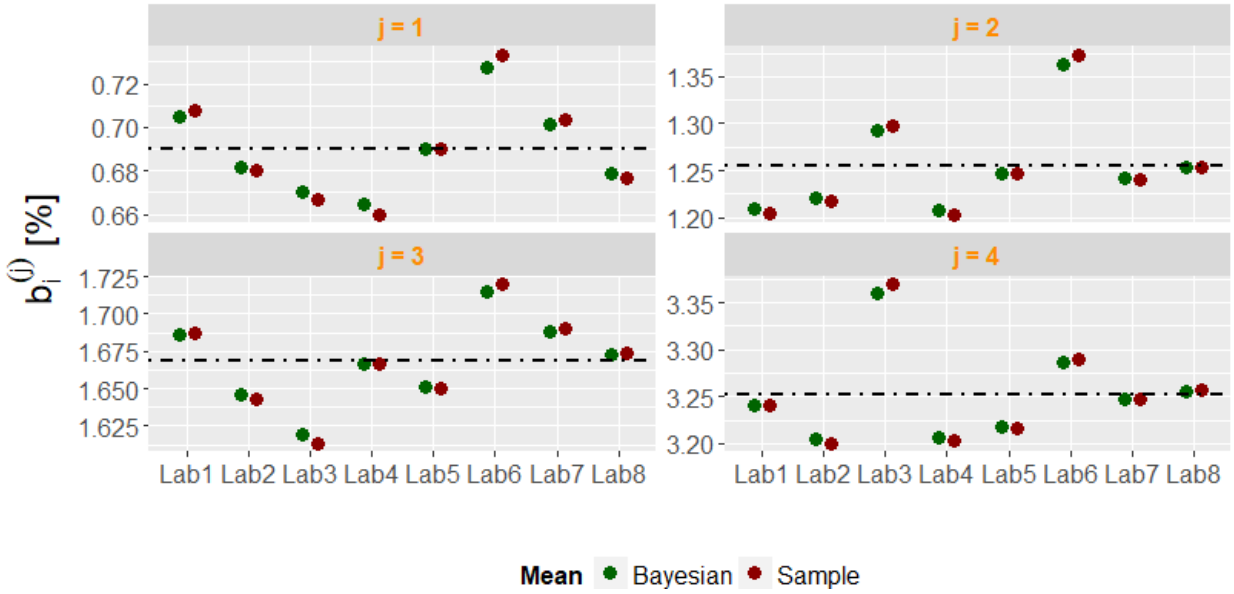

Figure 3.7: Comparison between the expected values of the Bayesian posterior probability distributions for each laboratory-specific parameter and the relative sample means

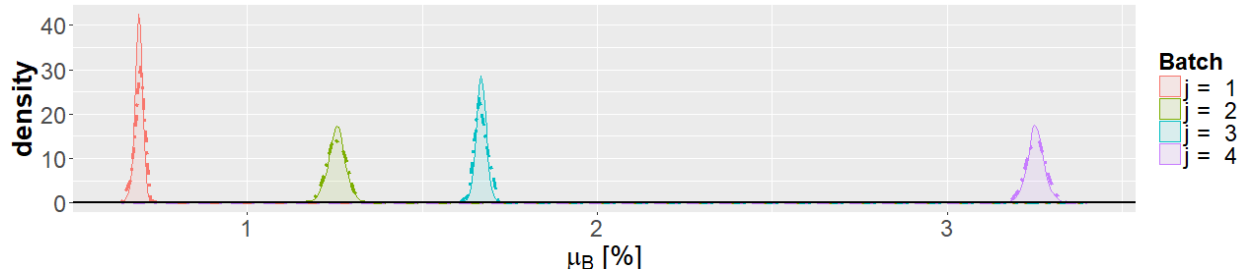

Figure 3.8: Posterior densities of the overall mean parameter for each level of the materials' batches

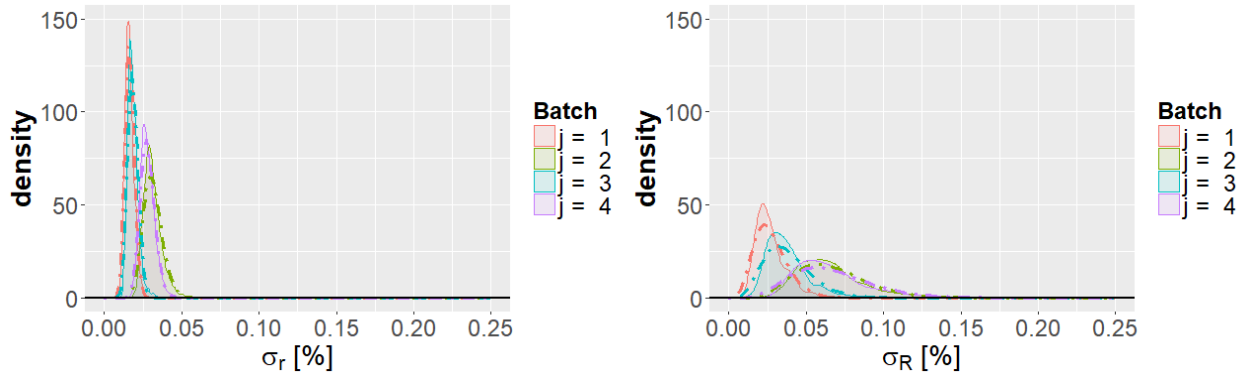

Figure 3.9: Posterior densities of the precision parameters under both repeatability and reproducibility conditions for each level of the materials' batches

In addition, the final results expressed in terms of standard deviations to be compared with the ANOVA results (see Table 2-7) are summarized in the following table.

$Level_j$	$m_B^{(j)}$	$s_r^{(j)}$	$s_R^{(j)}$
1	0.690	0.016	0.027
2	1.255	0.031	0.066
3	1.669	0.018	0.038
4	3.253	0.028	0.067

Table 3-7: Computed values of $m_B^{(j)}$, $s_r^{(j)}$ and $s_R^{(j)}$ for sulfur content of coal [%] through the accuracy Bayesian model 3.4

In conclusion, the accuracy Bayesian model 3.4 overall yields the same results given in section 2.1. The differences between the point estimates in Table 3-7 and those in Table 2-7 are at the level of the least significant digit of the measurements, so we can assume that are negligible. The main differences are:

i. the shrinkage to the population mean for all the laboratory-specific parameters, that is the tendency of the unit-specific parameters to be closer to the overall mean value;
ii. the slight increase in the point estimators for standard deviations under repeatability and reproducibility conditions, due to the use of the improper uniform prior.

What we appreciate most is the possibility of describing the accuracy of each material's batch with three marginal probability distributions, each uniquely linked to one of the concepts of trueness, repeatability precision and reproducibility precision. In the next section we see how this increased descriptive capability can be leveraged for the accuracy evaluation of a product line.

3.4 INDUSTRIAL APPLICATION CASE: VST TEST EQUIPMENTS PERFORMANCES

The accuracy hierarchical Bayesian model in Figure 3.1 can be applied to the industrial case analyzed in section 2.3. As already emphasized in section 2.2, moving from the domain of the inter-laboratory accuracy experiment to the production accuracy experiment, the reproducibility precision turns into homogeneity precision due to the different environment conditions, as summarized in Table 2-8. The final outcome reported in Table 2-11 is the benchmark against which to compare the following Bayesian results.

The first step is to arrange the data in Table 2-10 according to BUGS requirements and split theminto $q = 4$ subsets grouped by each of the specific level for the reference accepted value factor, that was named "jump". The experimental plan for the LVDT accuracy evaluation is balanced, so we do not need to deal with missing values as we can realize looking at the BUGS input data in Table 3-8, that is the data subset for the level "3»2" of the reference accepted value factor.

```
print(data)
$p
[1] 18

$n
[1] 4

$yik
       [,1]  [,2]  [,3]  [,4]
 [1,] 1.000 1.003 1.002 1.002
 [2,] 1.003 0.996 0.999 0.997
 [3,] 0.998 0.999 0.999 1.001
 [4,] 1.005 1.005 1.009 1.008
 [5,] 1.004 1.003 1.004 1.004
 [6,] 0.998 0.999 1.003 0.999
 [7,] 1.003 1.004 1.004 1.000
 [8,] 0.999 1.003 0.999 0.998
 [9,] 1.014 1.013 1.013 1.013
[10,] 0.998 0.999 0.999 1.000
[11,] 1.004 1.006 1.007 1.012
[12,] 1.001 1.001 1.003 1.003
[13,] 0.991 0.990 0.999 0.995
[14,] 0.999 0.998 0.999 0.998
[15,] 1.009 1.009 1.009 1.009
[16,] 1.001 1.001 1.003 1.003
[17,] 0.995 0.995 1.002 0.999
[18,] 0.995 0.995 0.996 0.998
```

Table 3-8: inputs list of the Hierarchical Bayesian model for the first "jump" of the LVDT accuracy experiment

The hierarchical model has the same probability structure as expressed in equation 3.4. So, we still have the two-levels Hierarchical Bayesian model with improper priors. The few differences are about objects naming, so it is written as:

```
model {
for (i in 1:p)
{
for(k in 1:n) {
yik[i,k] ~ dnorm (B[i], tau.r)
}
B[i] ~ dnorm (mu.B, tau.H)
}
mu.B ~ dnorm (0.0, 1.0E-6)
tau.H <- pow(sigma.H, -2)
sigma.H ~ dunif (0, 1000)
tau.r <- pow(sigma.r, -2)
sigma.r ~ dunif (0, 1000)
}
```

3.13

Again, this model must be stored in a separate file, e.g. "LVDT_model.txt" as it is named in ANNEX C §8, in an appropriate directory, say C:/PhD/Thesis, in order to be got into the MCMC routine. Then, in order to have a clever initialization of the Metropolis algorithm, the starting points are randomly sampled from tight distributions of the same families of those assumed for the

hyperparameters. The Hierarchical Bayesian problem is then solved through the R function `bugs()` of R2OpenBUGS package and MCMC posterior samples, summary statistics and other related information are available for further analysis.

It is possible to have a quick summary of the main statistics correlated to the MCMC simulation, as in Table 3-9 below.

```
Inference for Bugs model at "C:/PhD/Thesis/LVDT_model.txt",
Current: 3 chains, each with 1500 iterations (first 750 discarded)
Cumulative: n.sims = 2250 iterations saved
```

	mean	sd	2.5%	25%	50%	75%	97.5%	Rhat	n.eff
B[1]	1.002	0.001	1.000	1.001	1.002	1.002	1.004	1.001	2000
B[2]	0.999	0.001	0.997	0.998	0.999	1.000	1.001	1.001	2200
B[3]	0.999	0.001	0.997	0.999	0.999	1.000	1.001	1.001	2200
B[4]	1.007	0.001	1.004	1.006	1.007	1.007	1.008	1.002	940
B[5]	1.004	0.001	1.002	1.003	1.004	1.004	1.006	1.002	2200
B[6]	1.000	0.001	0.998	0.999	1.000	1.001	1.002	1.002	1700
B[7]	1.003	0.001	1.001	1.002	1.003	1.003	1.005	1.001	2200
B[8]	1.000	0.001	0.998	0.999	1.000	1.001	1.002	1.000	2200
B[9]	1.013	0.001	1.011	1.012	1.013	1.013	1.015	1.002	1700
B[10]	0.999	0.001	0.997	0.998	0.999	1.000	1.001	1.002	1100
B[11]	1.007	0.001	1.005	1.006	1.007	1.008	1.009	1.001	2200
B[12]	1.002	0.001	1.000	1.001	1.002	1.003	1.004	1.002	1500
B[13]	0.994	0.001	0.992	0.993	0.994	0.995	0.996	1.005	430
B[14]	0.999	0.001	0.997	0.998	0.999	0.999	1.001	1.001	2200
B[15]	1.009	0.001	1.007	1.008	1.009	1.009	1.011	1.000	2200
B[16]	1.002	0.001	1.000	1.001	1.002	1.003	1.004	1.000	2200
B[17]	0.998	0.001	0.996	0.997	0.998	0.999	1.000	1.001	2200
B[18]	0.996	0.001	0.994	0.996	0.996	0.997	0.998	1.000	2200
mu.B	1.002	0.001	0.999	1.001	1.002	1.003	1.004	1.002	980
sigma.R	0.005	0.001	0.004	0.004	0.005	0.006	0.008	1.002	1600
sigma.r	0.002	0.000	0.002	0.002	0.002	0.002	0.003	1.001	2200
deviance	-684.536	7.336	-696.600	-689.900	-685.300	-680.500	-667.967	1.000	2200

```
For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = Dbar-Dhat)
pD = 18.090 and DIC = -666.400
DIC is an estimate of expected predictive error (lower deviance is better).
```

Table 3-9: Summary of the main statistics for an MCMC simulation run with the package R2OpenBUGS

In the column “mean” of Table 3-9 we find the values to be compared with the point estimates evaluated with the ISO 5725-2 models. As we know, the point estimators of ANOVA models yield single-valued results for parameters, while the great advantage of operating in the Bayesian analysis framework is the possibility of dealing straightly with the marginal probability distributions for each hyperparameter. For example in Figure 3.10, we can compare the boxplots depicting the posterior probability distributions with the corresponding empirical measurements (raw data) grouped by product unit-specific parameter.

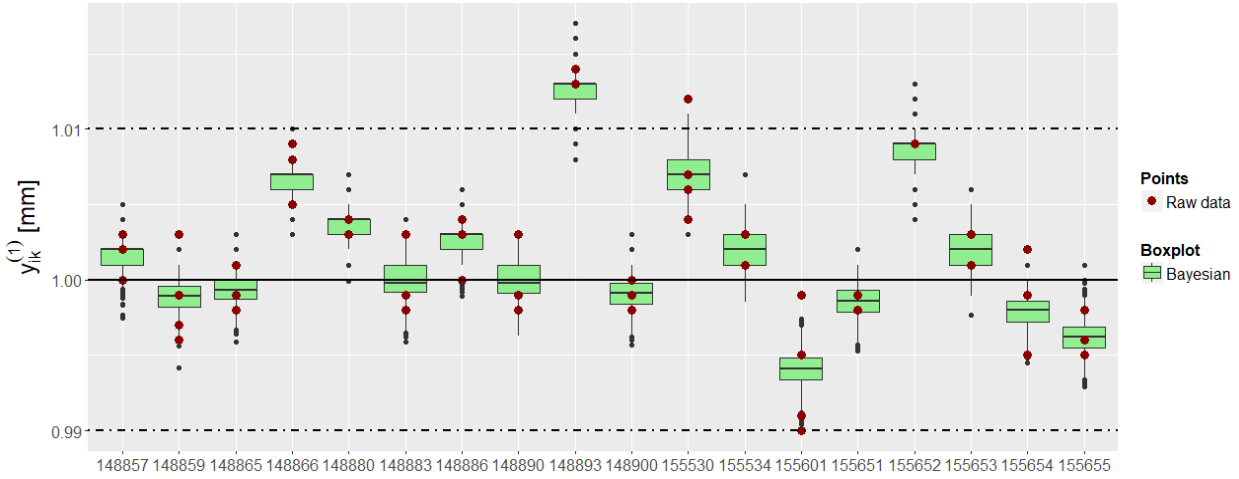

Figure 3.10: Comparison between red points representing specific-laboratory measurements (raw data) plotted respect to green boxplots describing posterior data for each specific-laboratory parameter as result of MCMC solution

Even if the sequence of boxplots for every equipment-specific parameter displays variations not homogeneously overlapping , only the LVDT having serial number 148893 is out of the acceptance range according to ISO 306 (dot-dashed lines). The same analysis can be repeated for the other

levels of the reference accepted value factor, as shown in Figure 3.11. Overall the result conveys a good fulfillment of the ISO 306 acceptance criteria for the penetration measurement. Also at the second level of the reference acceptance factor the measurements average of the LVDT 148893 it is slightly out of the thresholds. It is suspected to be a production fault and as such, it will be useful in the next chapter when data consistency techniques are discussed.

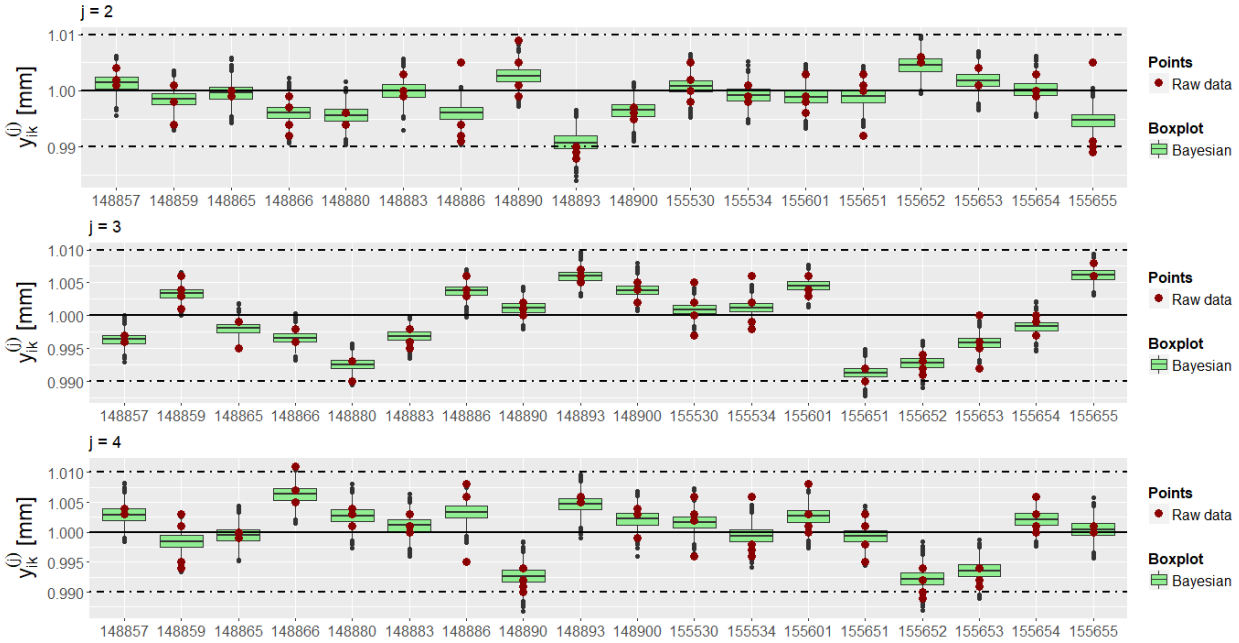

Figure 3.11: Comparison between red points representing specific-laboratory measurements (raw data) plotted respect to green boxplots describing posterior data for each specific-laboratory parameter as result of MCMC solution

Having some production faults does not affect the general mean parameter distributions, as can be seen in Figure 3.12. After Bayesian updating they are still normal, overall with the same variability span (similar shapes) and having slight differences in terms of specific mean values.

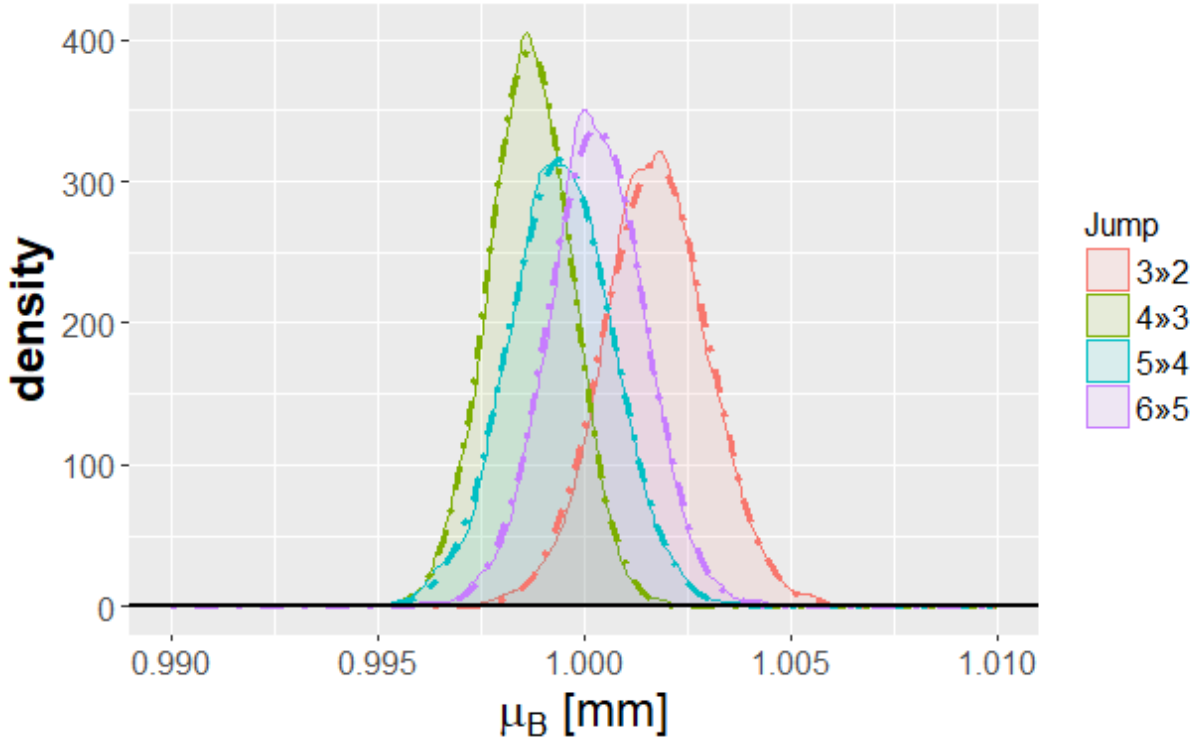

Figure 3.12: Marginal posterior probability densities for the general mean variable μ_B

Smooth distributions are also achieved for the repeatability and homogeneity standard deviation parameters, as shown in Figure 3.13. The figure shows a good fit between the MCMC

numerical realization for the posterior densities (solid lines) and the corresponding theoretical gamma densities (dot-dash lines). The latter PDFs were evaluated using shape and scale parameters calculated from the mean and variance of a random variable X having the gamma distribution [14].

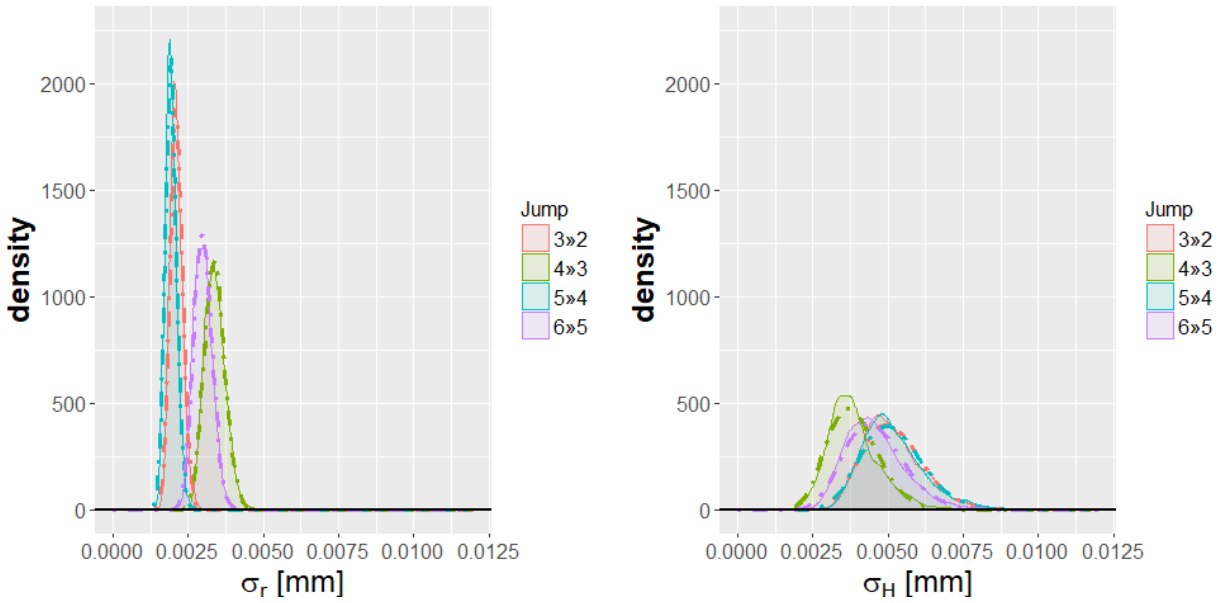

Figure 3.13: Marginal posterior probability densities for repeatability σ_r (left) and homogeneity σ_H (right) parameters

The validation of this model passes through the comparison between the point estimators given by ANOVA (Table 2-11) and the expected values of the marginal posteriors given by MCMC (Figure 3.12 and Figure 3.13). These are included in the output of the "bugs" function of the package R2OpenBUGS. The final result is summarized in Table 3-10.

j	Level$^{(j)}$	pD$^{(j)}$	DIC$^{(j)}$	$m_B^{(j)}$/mm	$s_r^{(j)}$/mm	$s_H^{(j)}$/mm
1	3»2	18.09	-666.4	1.0017	0.0021	0.0052
2	4»3	16.12	-600.2	0.9987	0.0034	0.0039
3	5»4	18.30	-679.3	0.9994	0.0019	0.0052
4	6»5	17.36	-616.6	1.0003	0.0030	0.0046

Table 3-10: Expected values of the posterior distributions for the accuracy parameters $m_B^{(j)}$, $s_r^{(j)}$ and $s_H^{(j)}$ based on the hierarchical Bayesian model 3.13

j	Level$^{(j)}$	$\Delta m_B^{(j)}$/mm	$\Delta s_r^{(j)}$/mm	$\Delta s_H^{(j)}$/mm
1	3»2	0	0e+00	0e+00
2	4»3	0	1e-04	-1e-03
3	5»4	0	0e+00	1e-04
4	6»5	0	1e-04	-5e-04

Table 3-11: Differences between the results obtained with the Hierarchical Bayesian model and those with the ANOVA

The results from the two models are very close. Looking at the differences (Table 3-11 and Figure 3.14), the largest of them (regarding homogeneity standard deviation) is of the same order of magnitude as the transducer resolution (1 μm), i.e., negligible.

In general, about repeatability, the Hierarchical Bayesian model gives the same outcome of ANOVA. Actually, at most for the jumps 4 » 3 mm and 6 » 5 mm the Bayesian result is 0.1 μm higher than the ANOVA that is to be considered negligible, being one-tenth the resolution.

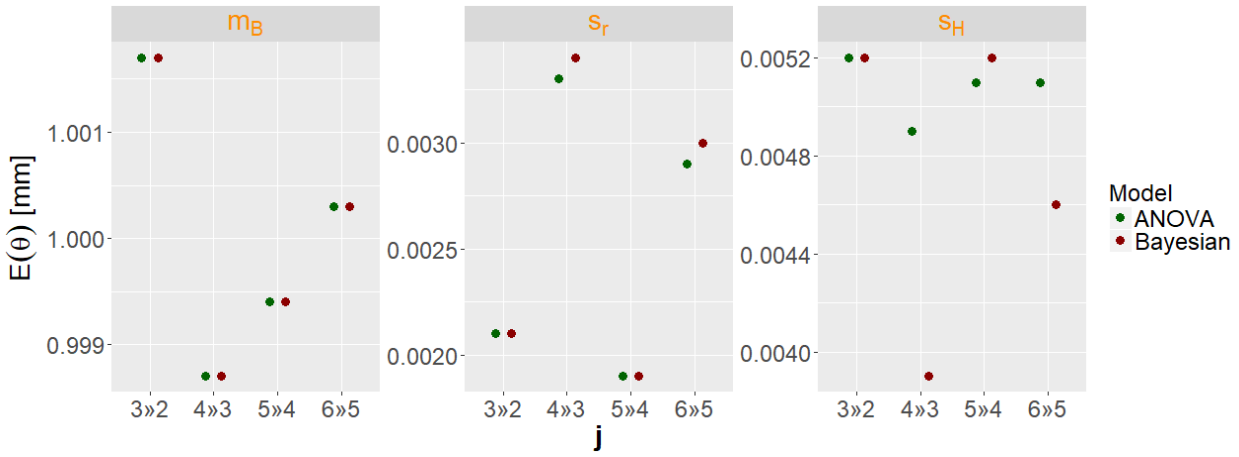

Figure 3.14: left to right, paired expected values, yield by ANOVA (green) and Bayesian (red), for general mean, repeatability and homogeneity parameters

The consistency of repeatability is as expected, whereas the behavior of homogeneity deserves deeper analysis. Looking at equation 2.7, the suspicion is that the what we identified as the homogeneity variable in the Hierarchical Bayesian model for accuracy evaluation (see Figure 3.1) is not rather the between-laboratory standard deviation (the ISO 5725 equivalent is reported in the equation 2.9). To remove this doubt we consider the mean values of the homogeneity standard deviation variables in Table 3-10 as the inter-laboratory standard deviation, that in this context is the inter-equipment standard deviation parameter $s_L^{(j)}$. Table 3-12 is as Table 3-10 with the last column heading changed.

j	Level$^{(j)}$	pD$^{(j)}$	DIC$^{(j)}$	$m_B^{(j)}$/mm	$s_r^{(j)}$/mm	$s_L^{(j)}$/mm
1	3»2	18.09	-666.4	1.0017	0.0021	0.0052
2	4»3	16.12	-600.2	0.9987	0.0034	0.0039
3	5»4	18.30	-679.3	0.9994	0.0019	0.0052
4	6»5	17.36	-616.6	1.0003	0.0030	0.0046

Table 3-12: Expected values of the posterior distributions for the accuracy parameter $m_B^{(j)}$, $s_r^{(j)}$ and $s_L^{(j)}$, i.e. interpreting the data in the last column as referring to the inter-equipment standard deviation rather than to homogeneity as it is in the hierarchical Bayesian model 3.13

Then, for each j we combine the squares of the point estimates of repeatability $s_r^{(j)}$ and inter-equipment standard deviations $s_L^{(j)}$, according to equation 2.7, to evaluate the homogeneity $s_H^{(j)}$ as their combination, according to ISO 5725-2. This result is reported in Table 3-13.

j	Level$^{(j)}$	$p_D^{(j)}$	DIC$^{(j)}$	$m_B^{(j)}$/mm	$s_r^{(j)}$/mm	$s_L^{(j)}$/mm	$s_H^{(j)}$/mm
1	3»2	18.09	-666.4	1.0017	0.0021	0.0052	0.0056
2	4»3	16.12	-600.2	0.9987	0.0034	0.0039	0.0052
3	5»4	18.30	-679.3	0.9994	0.0019	0.0052	0.0055
4	6»5	17.36	-616.6	1.0003	0.0030	0.0046	0.0055

Table 3-13: it adds to Table 3-12 the column $s_H^{(j)}$, achieved from the point estimates of $s_r^{(j)}$ and $s_L^{(j)}$, as the standard provides (see equation 2.7)

The new paired values are also given graphically in Figure 3.15. The overall mean and repeatability comparisons being unaffected, only the inter-equipment and homogeneity comparisons are shown. The paired values are still close, with a definitely better pattern. The

differences are almost constant and the Bayesian outcome is (consistently) slightly bigger than the ANOVA.

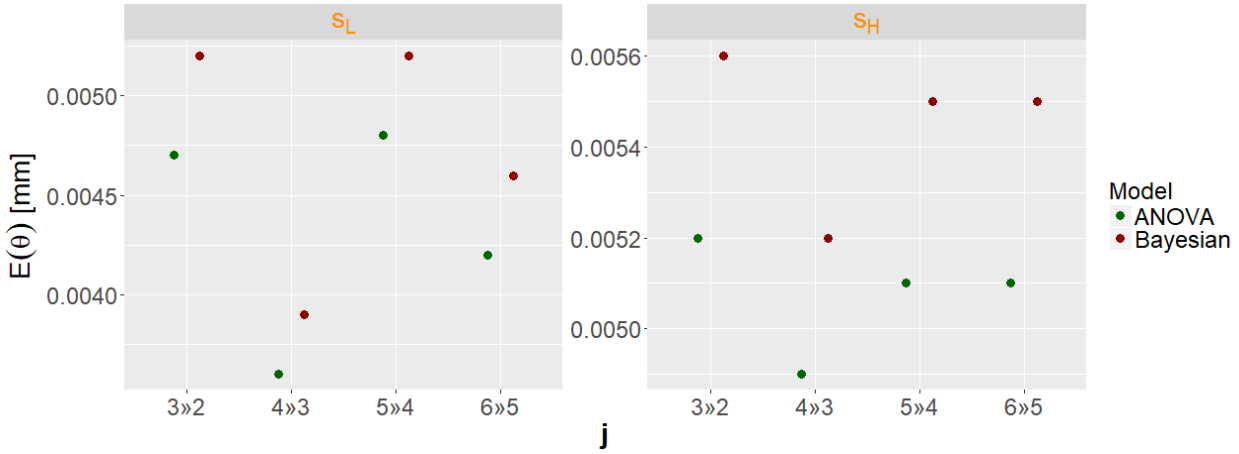

Figure 3.15: left to right, paired expected values, yield by ANOVA (green) and Bayesian (red), for inter-equipment and homogeneity parameters

Overall, the closeness of the results (well within resolution) confirms the validity of the hierarchical Bayesian model outcome, provided that we redefine the hyper-parameter for homogeneity as that for inter-laboratory variability in the proposed Hierarchical Bayesian model for Accuracy evaluation. Therefore, the Hierarchical Bayesian model for the accuracy evaluation becomes as in the schematic doodle graph of Figure 3.16.

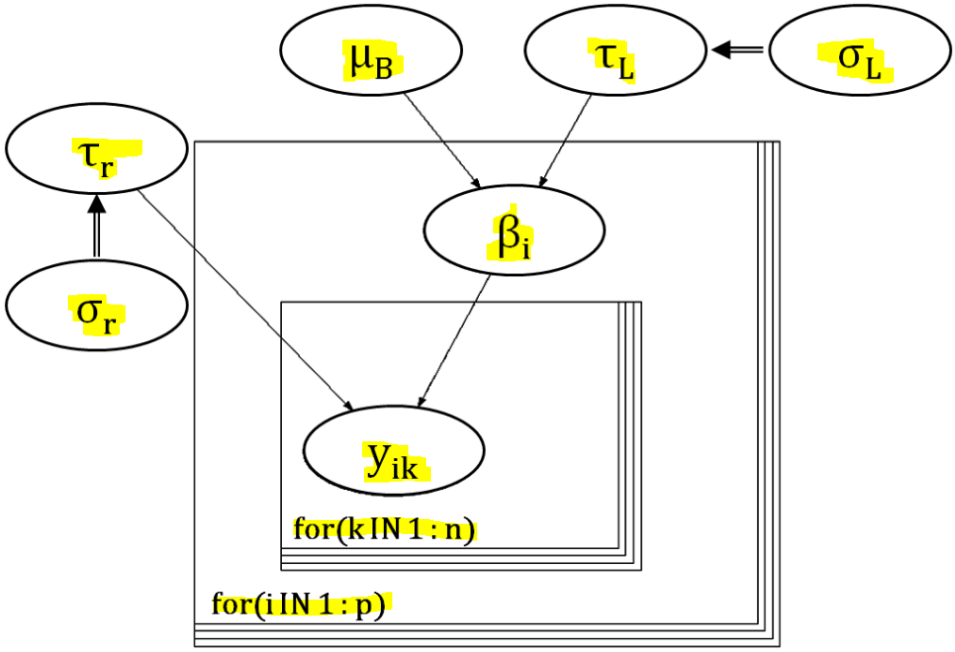

Figure 3.16: Doodle graph of the Hierarchical Bayesian model for the accuracy evaluation with the inter-laboratory node

The bound between the hyperparameters of repeatability and inter-equipment variability is not included in Figure 3.16, even though the flexibility of the MCMC algorithm would allow to introduce equation 2.7 directly in the model definition. As a matter of fact from the numerical point of view it would not be an issue to combine repeatability and inter-laboratory variability in order to achieve the random variable distribution for the homogeneity. By contrast, it is impossible to find a closed-form expression under the link of equation 2.7. At the moment, we leave aside the analytical implications and we focus on the design of the model according to the ISO 5725-2 constrains. It will be the purpose of other studies to rewrite the model under different assumptions in order to make the problem tractable and to solve it in terms of a closed-form expression. Here, the scope is to look for the model that better overlaps the ISO 5725-2 results, having the joint probability structure as close as possible to the "soul" of the standard itself. In order to create the logical link

between the nodes of repeatability and inter-laboratory variability, according to equation 2.7 and inherently to the joint probability structure, model 3.13 is modified as follows:

```
model {
for (i in 1:p)
{
for(k in 1:n) {
yik[i,k] ~ dnorm (B[i], tau.r)
}
B[i] ~ dnorm (mu.B, tau.L)
}
mu.B ~ dnorm (0.0, 1.0E-6)
tau.r <- pow(sigma.r, -2)
sigma.r ~ dunif (0, 1000)
tau.L <- pow(sigma.L, -2)
sigma.L ~ dunif (0, 1000)
sigma.H <- sqrt(pow(sigma.r, 2) + pow(sigma.L, 2))
}
```

3.14

This model must be stored in a separate file, e.g. “LVDT_repr.txt”, in an appropriate directory, say C:/PhD/Thesis.

The doodle graph used to represent model 3.14 is shown in Figure 3.17, where the double arrows linking the repeatability and inter-equipment standard deviation nodes to their inverse variances represent the transformation of equation 3.5, and the double arrows linking the repeatability and inter-equipment standard deviation nodes to the homogeneity standard deviation node represent equation 2.7.

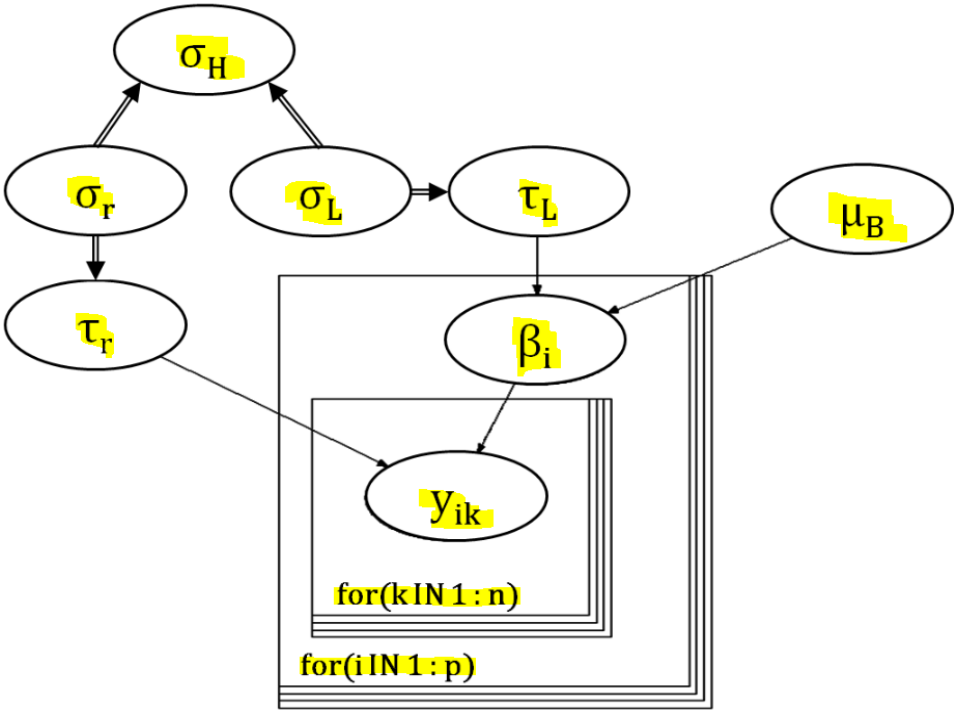

Figure 3.17: Hierarchical Bayesian model for the accuracy evaluation with both the inter-laboratory and the homogeneity nodes

Also in this case the OpenBUGS engine is able to numerically solve the Bayesian problem, using the prior probability structure

$$\begin{aligned} Y_{ik}^{(j)}|\beta_i^{(j)},\sigma_r^{(j)} &\sim i.i.d.\quad Normal\left(\beta_i^{(j)},\sigma_r^{(j)}\right) \\ \beta_i^{(j)}|\mu_B^{(j)},\sigma_L^{(j)} &\sim i.i.d.\quad Normal\left(\mu_B^{(j)},\sigma_L^{(j)}\right) \\ \mu_B^{(j)} &\sim Normal(0,\infty) \\ \sigma_r^{(j)} &\sim Uniform(0,\infty) \\ \sigma_L^{(j)} &\sim Uniform(0,\infty) \end{aligned} \quad . \qquad 3.15$$

The starting points are randomly sampled according to the narrow prior distributions of the parameters, and so the BUGS engine can run as many times as are the number of levels for the reference accepted value factor, i.e. q times. The final results for the expected values of the marginal probability distributions of interest are summarized in Table 3-14.

j	Level$^{(j)}$	$p_D^{(j)}$	DIC$^{(j)}$	$m_B^{(j)}$/mm	$s_r^{(j)}$/mm	$s_L^{(j)}$/mm	$s_H^{(j)}$/mm
1	3»2	18.09	-666.4	1.0017	0.0021	0.0052	0.0056
2	4»3	16.12	-600.2	0.9987	0.0034	0.0039	0.0052
3	5»4	18.31	-679.3	0.9994	0.0019	0.0052	0.0055
4	6»5	17.36	-616.6	1.0003	0.0030	0.0046	0.0055

Table 3-14: expected values of the marginal posterior distributions for the parameters m_j, s_{rj}, s_{Lj} and s_{Rj} based on the Hierarchical Bayesian model of Figure 3.17 for the LVDT accuracy evaluation

The final outcome of the q expected values for the parameters of interest is the same as that of Table 3-13. Therefore, we achieve the same results whether we get the homogeneity straight from the expected value of the marginal posterior distributions of the repeatability and inter-equipment variability or if we get the homogeneity through the joint probability distribution in the Bayesian formulation. The second option is preferred because it maximizes the amount of information. As a matter of fact, the Bayesian data updating provides the marginal posterior distribution for each of interest parameter, including the inter-equipment variability and the homogeneity, as shown in Figure 3.18.

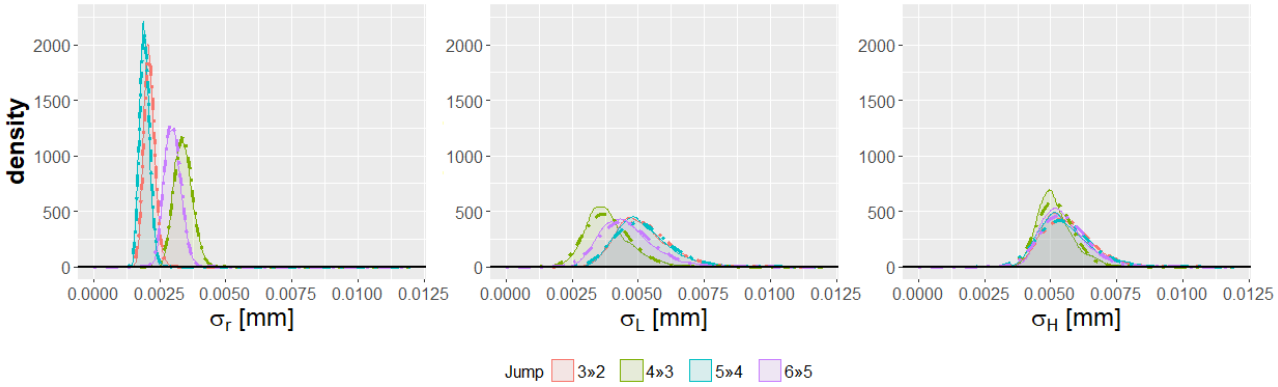

Figure 3.18: Marginal posterior probability density functions for repeatability σ_r, inter-equipment variability σ_L and homogeneity σ_H parameters (left to right)

It can be seen from Figure 3.18 that the marginal posterior distribution for the repeatability parameter is the same achieved with model 3.13 and the same holds for the inter-equipment distribution if compared to the homogeneity one (see Figure 3.13 to compare results). The dot-dash lines show the gamma distributions evaluated using shape and scale parameters calculated from the updated data for each level of the factor "Jump" using the classical equations providing their relationship with sample mean and variance[17]. Overall, the fit is not bad, especially for the repeatability parameter.

Finally, under the same assumptions used to obtain Table 2-14, we can arrange the final outcome of the hierarchical model 3.14 in terms of bias interval $\left[\hat{\delta}_{\text{lower}}^{(j)}, \hat{\delta}_{\text{upper}}^{(j)}\right]$ and expanded precisions for both repeatability $U_r^{(j)}$ and homogeneity $U_H^{(j)}$ conditions. Unlike the ISO 5725-2 framework, the proposed Hierarchical Bayesian model for accuracy evaluation provides a full description as well for bias in terms of posterior probability distribution. So, if using ISO 5725-2 the 95 % confidence interval for bias (Table 2-14) is the maximum amount of information about the trueness, the Bayesian framework provides a full description of the bias, and accordingly of the trueness. This represents the great advantage. In this way, the manufacturer can declare an interval with a stipulated level of credibility to describe also the trueness performance. As concerns expanded precisions, we can use the upper endpoint of the 95 % unilateral credible intervals

evaluated according to their marginal posterior probability distributions. In formulae, for a credibility level of 95 % we have

$$\begin{aligned} Bias \quad &\Rightarrow \quad \left[\hat{\delta}^{(j)}_{\text{lower}}, \hat{\delta}^{(j)}_{\text{upper}}\right] \coloneqq P\left(\hat{\delta}^{(j)}_{\text{lower}} \le \mu^{(j)}_B - \mu^{(j)} \le \hat{\delta}^{(j)}_{\text{upper}}\right) = 0.95 \\ Repeatability\ expanded\ precision \quad &\Rightarrow \quad u^j_r : \Phi_{\sigma^{(j)}_r}\left(u^{(j)}_r\right) = P\left(\sigma^{(j)}_r \le u^{(j)}_r\right) = 0.95 \\ Homogeneity\ expanded\ precision \quad &\Rightarrow \quad u^j_H : \Phi_{\sigma^{(j)}_H}\left(u^{(j)}_H\right) = P\left(\sigma^{(j)}_H \le u^{(j)}_H\right) = 0.95 \end{aligned} \quad , \qquad 3.16$$

where:
$\mu^{(j)}_B$ is the overall mean variable at each j-level of the reference accepted value $\mu^{(j)}$, described as a normal distribution (Figure 3.12) . The bias interval is thus defined by those values encompassing the 95 % of probability for the overall mean density shifted by the reference accepted value $\mu^{(j)}$; $\Phi_{\sigma^{(j)}_r}$ and $\Phi_{\sigma^{(j)}_H}$ are the cumulative distribution function for the repeatability $\sigma^{(j)}_r$ and reproducibility $\sigma^{(j)}_H$ parameters respectively. As we saw, both distributions can be described by gamma functions with their characteristic shape and scale parameters (Figure 3.18). Therefore, the expanded precisions are the value encompassing 95 % of probability.

According to equations 3.16, the accuracy outcome is drawn and the result is summarized in Table 3-15.

j	Level$^{(j)}$	$p_D^{(j)}$	DIC$^{(j)}$	$\hat{\delta}^{(j)}_{\text{lower}}$/mm	$\hat{\delta}^{(j)}_{\text{upper}}$/mm	$U_r^{(j)}$/mm	$U_H^{(j)}$/mm
1	3»2	18.09	-666.4	-0.001	0.004	0.0025	0.0073
2	4»3	16.12	-600.2	-0.003	0.001	0.0040	0.0065
3	5»4	18.31	-679.3	-0.003	0.002	0.0023	0.0074
4	6»5	17.36	-616.6	-0.002	0.003	0.0035	0.0071

Table 3-15: Bias and uncertainties under repeatability and reproducibility condition for the LVDT accuracy evaluation based on the Hierarchical Bayesian model of Figure 3.17

It should be stressed that, although the expected values for the standard deviation parameters in Table 3-14 are bigger than the point estimators provided by ANOVA (Table 2-11), the evaluated values for expanded precisions reverse the trend for both repeatability and homogeneity (see Table 3-15 against Table 2-14). The reason is the possibility of exploiting the whole amount of information encompassed by the marginal posterior distributions for bias, repeatability and homogeneity precisions instead of mixing the concepts of bias and uncertainty in the confidence intervals as in the ISO 5725-2 framework. Especially, it was the trueness to awaken some misunderstandings, to the point that many manufacturers preferred to describe their products only in terms of precision. On the other hand, the Hierarchical Bayesian model removes for the manufacturer the ambiguity in the description of the accuracy performances of a testing equipment family. The trueness can be described in terms of credibility interval for bias parameter. As a matter of fact, with respect to the confidence intervals for the constant parameter of the ANOVA model, the credibility intervals encompass the zero, so there is no reason for rejecting the hypothesis of null bias as it seemed looking at the ANOVA outcome in Table 2-14. So, as a consequence of the result in Table 3-15, there are no systematic errors to be corrected by reviewing the product design. The precision can be described in terms of the right endpoint of the unilateral credibility interval for both the parameters. These are expression of the random errors under repeatability or production homogeneity conditions, respectively, and so, they can introduce both positive and negative contribution.

This opportunity of error decomposition is really appreciated by the engineering mindset. The reductionist attempt to provide explanation in terms of smaller entities was already followed at the design stage of the new product project. The displacement measurement of LVDTs was split into three components. The first component was thought as the error introduced by the mechanics due to the tolerance chain that can cause an imperfect perpendicularity between the rod and the test specimen support (see Figure 2.2, item 4 and 9). This can be considered as a systematic error, different among equipments, because each specific-equipment, at the time at which it was

produced, had its own specific angle for the rod axis of the LVDT. The second component was the sum of the available clearance for the LVDT rod in order to guarantee a frictionless positioning and of the electronic inner variability. It expresses the random variability of the single equipment and, so, the random errors. Lastly, it was taken into account the inter-equipment variability. Then, in the discovery stage of the new product development project, a specific quantity requirement was stated for each of these entities in order to better control the design. So, every design choice was taken in order to address that specific requirement. Twice the sum square of the design requirements is roughly 0.01 mm, as it is stated in the standard ISO 306 for the accuracy accepted value [6, Par. 5.2.4]. The decomposition provides ±4 µm for the mechanical systematic error, ±4 µm for the intra-equipment random error and ±6 µm for the inter-equipment random error. This attitude in the Bayesian framework can be translated in the following proper priors:

$$\begin{aligned} \mu_B^{(j)} &\sim Unif(0.996, 1.004) \\ \sigma_r^{(j)} &\sim Unif(0, 0.008) \\ \sigma_L^{(j)} &\sim Unif(0, 0.012) \end{aligned} \qquad 3.17$$

Using these priors we can solve the MCMC problem after suitably modifying model 3.14 as stated below:

```
model {
for (i in 1:p)
{
for(k in 1:n) {
yik[i,k] ~ dnorm (B[i], tau.r)
}
B[i] ~ dnorm (mu.B, tau.L)
}
mu.B ~ dunif (0.995, 1.005)
tau.r <- pow(sigma.r, -2)
sigma.r ~ dunif (0, 0.008)
tau.L <- pow(sigma.L, -2)
sigma.L ~ dunif (0, 0.012)
sigma.R <- sqrt(pow(sigma.r, 2) + pow(sigma.L, 2))
}
```

3.18

The final result in terms of posterior distribution for accuracy parameters is depicted in Figure 3.19 below. Considering that, as seen in the uncertainty budget in Table 2-13, the reference accepted value can be modeled as a random variable described by a Dirac delta function which is zero everywhere except at 1 mm, then, according to 2.25, the trueness can be described by the same bell shapes of the general mean parameter μ_B in Figure 3.19 but shifted by -1 mm. The bias result thus obtained is depicted in Figure 3.20.

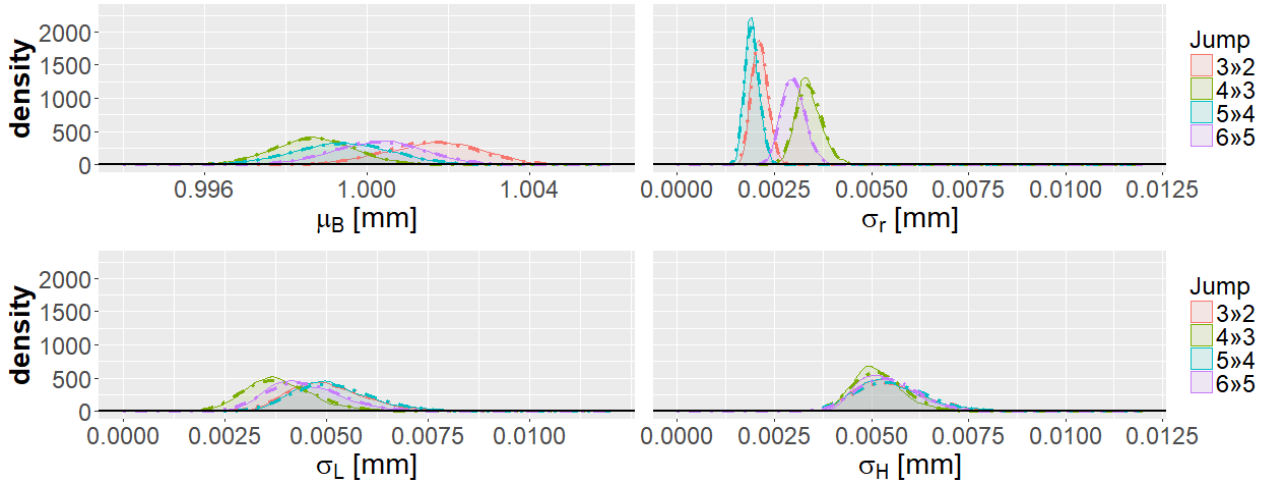

Figure 3.19: marginal posterior probability density functions for general mean μ_B, repeatability σ_r, inter-equipment variability σ_L and homogeneity σ_H parameters (from top left to right and then bottom left to right)

Hence, Figure 3.20 condenses all the information needed to describe the accuracy of a whole product line into an unique graph. Besides, bias is clearly distinguishable from precision and, so, the systematic error from the random error.

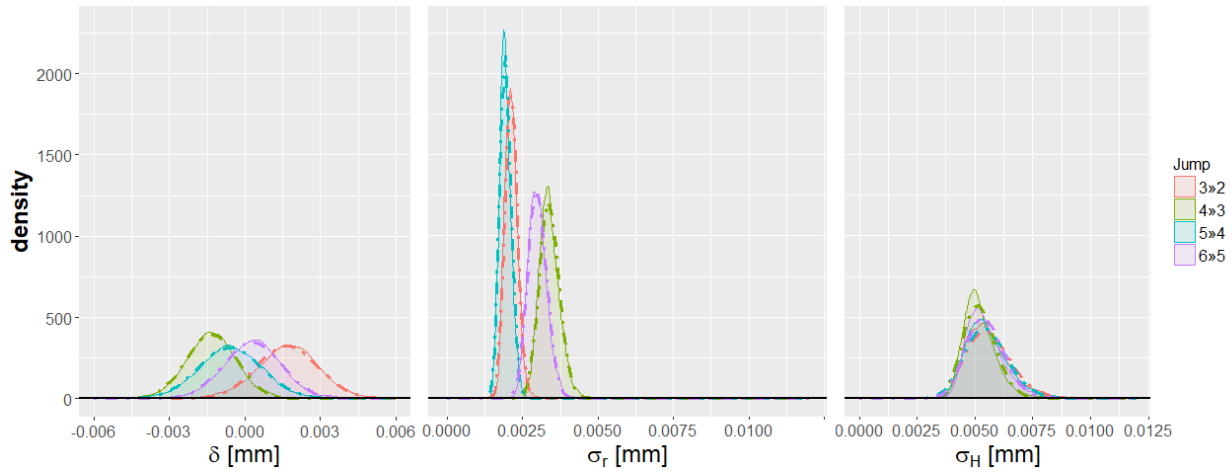

Figure 3.20: accuracy desciption in terms of the posterior density function of bias, repetability precision and homogeneity precision (from left to right)

Finally, we can apply the statistics the equation 3.16 to the result in Figure 3.20 aimed to complete the accuracy description under the prior conditions in 3.17. The outcome is summarized below in Table 3-16.

j	$Level^{(j)}$	$p_D^{(j)}$	$DIC^{(j)}$	$\hat{\delta}_{lower}^{(j)}/mm$	$\hat{\delta}_{upper}^{(j)}/mm$	$U_r^{(j)}/mm$	$U_H^{(j)}/mm$
1	3»2	18.09	-666.4	-0.001	0.004	0.0025	0.0073
2	4»3	16.12	-600.2	-0.003	0.001	0.0040	0.0065
3	5»4	18.31	-679.3	-0.003	0.002	0.0023	0.0073
4	6»5	17.36	-616.6	-0.002	0.003	0.0035	0.0069

Table 3-16: Bias and expanded precisions under repeatability and homogeneity conditions for the LVDT accuracy evaluation based on the Hierarchical Bayesian model of Figure 3.17 using the proper prior distributions conditions of 3.17

As expected, we have a “slight” reduction in terms of variability with respect to the outcome under improper prior conditions (see Table 3-15). This reduction is negligible, being well within the equipment resolution. Overall, we can emphasise that, having a negligible effect in the analysis due to the influence of the vague prior, this strongly suggests we have sufficient data to draw a robust conclusion.

We remark that informative prior distributions can be based on pure judgment, data from literature (in a broad sense, e.g. calibration report, scientific papers, etc.), or a mixture of them. Of course, selecting an informative prior distribution is never an automatic procedure but needs a deep understanding of the full process. The blind usage of the informative prior risks to heighten the model ambiguity and to twist the final result.

Looking at the outcome in Table 3-16, the manufacturer may conclude that its testing equipment family has the following accuracy description about the LVDT displacement measurements:

- trueness, evaluated in terms of bias, is in the range $[-3, 4]\ \mu m$;
- repeatability, evaluated in terms of expanded precision is $\pm 4\ \mu m$;
- homogeneity, evaluated in terms of expanded precision is $\pm 7\ \mu m$.

This synthesis of the statistics in Table 3-16 was achieved by considering the worst case for each parameter. But can we do better? A tentative answer to this question is given in the last section of this chapter.

3.5 DIFFERENT HYPOTESIS FOR THE JOINT PROBABILITY STRUCTURE

At the end of the previous section 3.4, we left pending a question about the possibility of doing something better in order to achieve a unique result summarizing the q-levels of the reference accepted value "jump". In ISO 5725-2 you cannot find any openings to the possibility of summarizing with unique parameters the statistics between the batch factor's levels. Nevertheless, for the LVDT accuracy experiment, it also seems reasonable to suppose that, although different, the parameters for the q-levels of the reference accepted value factor are similar in some sense, and so we may wish to assume they are exchangeable. So, extending the exchangeability assumption to the reference accepted value factors, we can rewrite the accuracy model by adding another hierarchical level, as shown in the doodle graph of Figure 3.21.

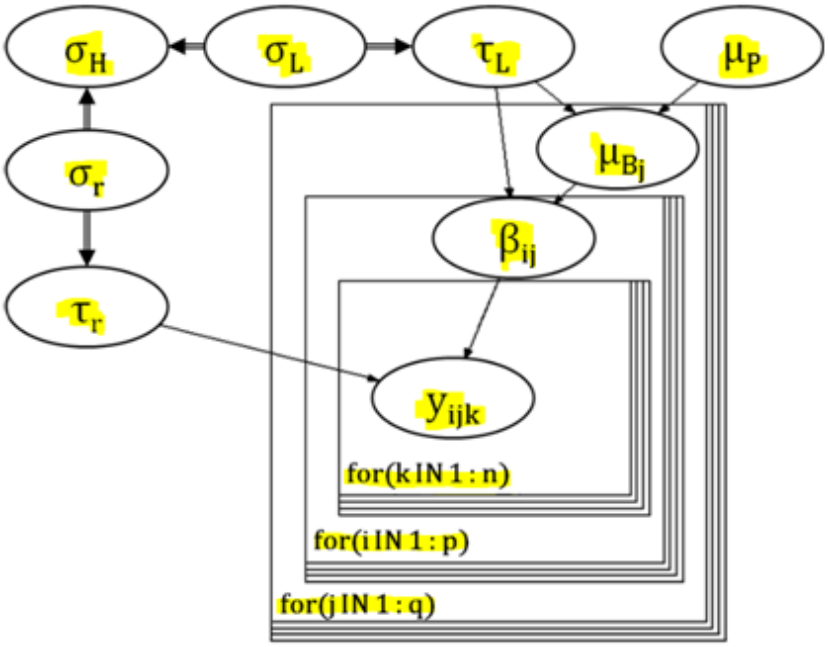

Figure 3.21: Doodle graph of the full Hierarchical Bayesian model for the accuracy evaluation of testing equipments

Hence, this full hierarchical Bayesian model has the inter-equipments means β_{ij} as first level, the reference accepted values means μ_{Bj} as second level, and the hyperparameters for the overall product line mean μ_P, the inter-equipments variability σ_L and the repeatability σ_r as third level.

The assumptions behind the full model in Figure 3.21 can be written as:

$$\begin{aligned} Y_{ijk}|\beta_{ij},\sigma_r &\sim i.i.d.\ \ Norm(\beta_{ij},\sigma_r) \\ \beta_{ij}|\mu_{B_j},\sigma_L &\sim i.i.d.\ \ Norm\left(\mu_{B_j},\sigma_L\right) \\ \mu_{B_j}\big|\ \mu_P,\sigma_L &\sim i.i.d.\ \ Norm(\mu_P,\sigma_L) \\ \mu_P &\sim Unif(0.996,1.004) \\ \sigma_r &\sim Unif(0,0.008) \\ \sigma_L &\sim Unif(0,0.012) \end{aligned} \quad . \qquad 3.19$$

The same scale parameter σ_L was assigned to both reference accepted value mean μ_{Bj} and specific-equipment mean β_{ij} parameters, because we expect that the inter-equipments variability conditions affect in the same way both parameters. Besides, under this assumption the homogeneity parameter is still the combination of repeatability and inter-equipment variability as stated in equation 2.7.

Given the measurement model 3.19 that relates the hyperparameters output quantities to the available measurement data, the distributions of the former are obtained by propagating the experimental evidence provided by observed data according to the axioms of probability calculus and by applying Bayes' theorem. Having a further hierarchical level, we need to add a further

subscript in the input data to identify the specific reference accepted value, so in this case the input data object for the “bugs” R function has the structure shown in the following Table 3-17.

```
> print(data)
$p
[1] 18

$q
[1] 4

$n
[1] 4

$xkij
, , 1
      [,1]  [,2]  [,3]  [,4]  [,5]  …  [,14] [,15] [,16] [,17] [,18]
[1,] 1.000 1.003 0.998 1.005 1.004  …  0.999 1.009 1.001 0.995 0.995
[2,] 1.003 0.996 0.999 1.005 1.003  …  0.998 1.009 1.001 0.995 0.995
[3,] 1.002 0.999 0.999 1.009 1.004  …  0.999 1.009 1.003 1.002 0.996
[4,] 1.002 0.997 1.001 1.008 1.004  …  0.998 1.009 1.003 0.999 0.998

, , 2
      [,1]  [,2]  [,3]  [,4]  [,5]  …  [,14] [,15] [,16] [,17] [,18]
[1,] 1.001 0.994 0.999 0.997 0.994  …  0.992 1.005 1.001 0.999 0.991
[2,] 1.001 1.001 1.000 0.992 0.994  …  1.003 1.006 1.004 1.000 1.005
[3,] 1.002 0.998 1.000 0.999 0.996  …  1.001 1.006 1.001 1.000 0.989
[4,] 1.004 1.001 1.000 0.994 0.996  …  1.000 1.006 1.004 1.003 0.990

, , 3
      [,1]  [,2]  [,3]  [,4]  [,5]  …  [,14] [,15] [,16] [,17] [,18]
[1,] 0.996 1.001 0.995 0.996 0.993  …  0.992 0.991 1.000 0.999 1.006
[2,] 0.997 1.004 0.999 0.998 0.993  …  0.992 0.992 0.992 1.000 1.006
[3,] 0.996 1.003 0.999 0.996 0.993  …  0.990 0.994 0.996 0.997 1.008
[4,] 0.996 1.006 0.999 0.996 0.990  …  0.990 0.993 0.995 0.997 1.006

, , 4
      [,1]  [,2]  [,3]  [,4]  [,5]  …  [,14] [,15] [,16] [,17] [,18]
[1,] 1.003 0.995 0.999 1.005 1.003  …  1.001 0.989 0.992 1.001 1.000
[2,] 1.004 1.003 0.999 1.007 1.001  …  1.003 0.990 0.994 1.000 1.000
[3,] 1.003 1.001 1.000 1.005 1.004  …  0.998 0.994 0.991 1.003 1.001
[4,] 1.003 0.994 1.000 1.011 1.004  …  0.995 0.992 0.994 1.006 1.001
```

Table 3-17: inputs list of the full Hierarchical Bayesian model for the LVDT accuracy experiment

The full model is then formulated according to the standards of the OpenBUGS code, as below.

```
model {
        m ~ dunif (0.997, 1.003)
        tau.L <- pow(sigma.L, -2)
        sigma.L ~ dunif (0, 0.008)
        tau.r <- pow(sigma.r, -2)
        sigma.r ~ dunif (0, 0.006)
        sigma.R <- sqrt(pow(sigma.r, 2) + pow(sigma.L, 2))
                for (j in 1:q) {
                mu.B[j] ~ dnorm (m, tau.L)
                for (i in 1:p) {
                        for (k in 1:n){
                                        xkij[k,i,j] ~ dnorm (B[i,j], tau.r)
                                        }
                                B[i,j] ~ dnorm (mu.B[j], tau.L)
                        }
        }
}
```

3.20

This model must be stored in a separate file, e.g. “LVDT_full.txt”, in an appropriate directory, say C:/PhD/Thesis.

We must be careful, if attempting to use informative priors, that our choice for the starting points does not influence the calculus solving time too much or even render the successive MCMC approximations not converging to the solution in the limits. Once again, this risk can be avoided by ensuring that the starting points are randomly selected from tighter proper prior distributions, where tighter means having a reduced scale parameter.

After this last expedient the MCMC solver can run to yield the result analysis. Nevertheless, before looking at the statistics for the accuracy description, we can provide a quick model check by comparing the expected value of the marginal posterior distributions for the reference accepted values means $E(\mu_{Bj})$ to those achieved by iterating q times the Hierarchical Bayesian model of Figure 3.17 and summarized in the column $m_B^{(j)}$ of Table 3-14. The four expected values for the full Hierarchical Bayesian model, given in Table 3-18, are almost exactly the same as those of Table 3-14, so that we can assume that the new joint probability structure does not introduce significant differences on the results.

m_{B_1}/mm	m_{B_2}/mm	m_{B_3}/mm	m_{B_4}/mm
1.0017	0.9988	0.9994	1.0003

Table 3-18: expected value of the marginal posterior distributions for the reference accepted values means $E(\mu_{Bj})$

Looking at Figure 3.22 we appreciate how the overall product line mean μ_P encompasses every q-th variable for reference accepted parameters μ_{B_j}, that is exactly what we are looking at to describe trueness: a unique statistic able to describe the population mean with respect to the reference accepted value. Again, the reference accepted value μ can be modelled as a random variable described by a Dirac delta probability function which is zero everywhere except at 1 mm. Then, according to 2.25, the trueness can be described by the same bell shapes of the overall product line mean μ_P in Figure 3.19 but shifted by -1 mm. The bias result thus obtained is depicted in Figure 3.22.

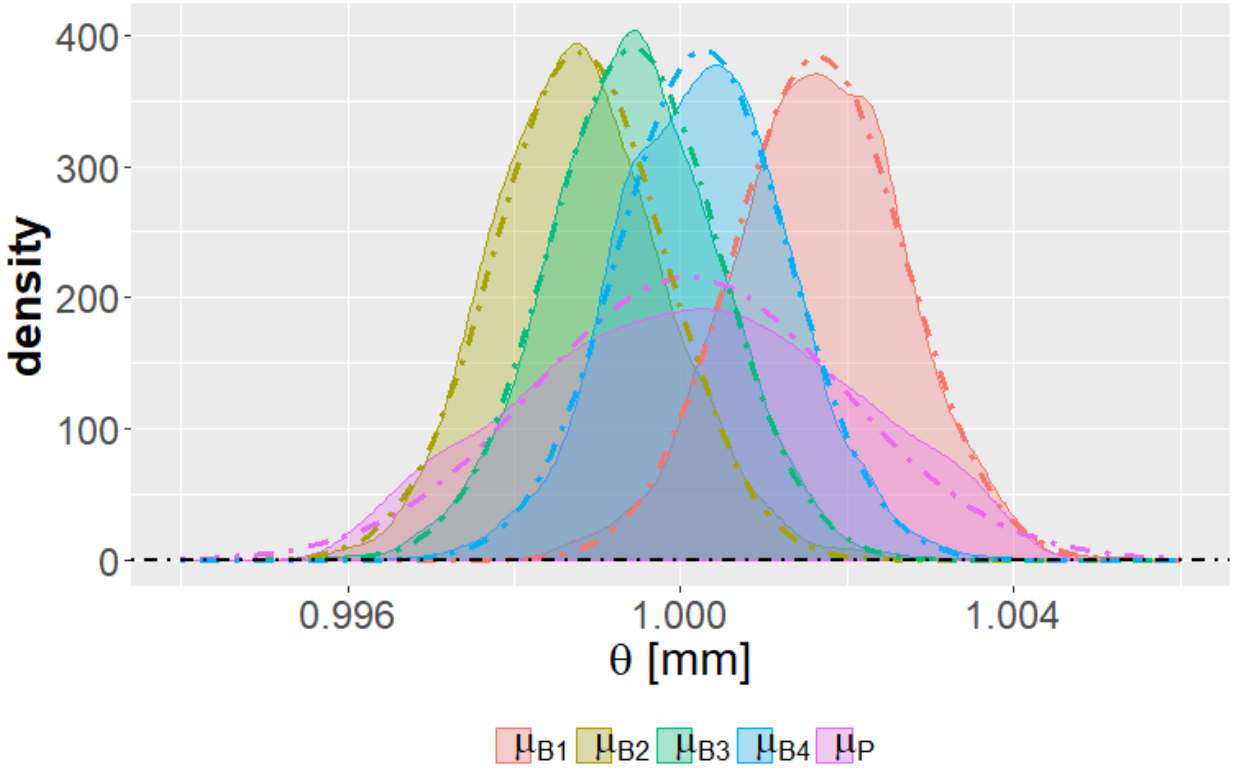

Figure 3.22: posterior probability density functions for the main mean parameters

The final result in terms of posterior density functions for the accuracy parameters of interest, i.e. trueness, repeatability precision and homogeneity precision, is shown below in Figure 3.23.

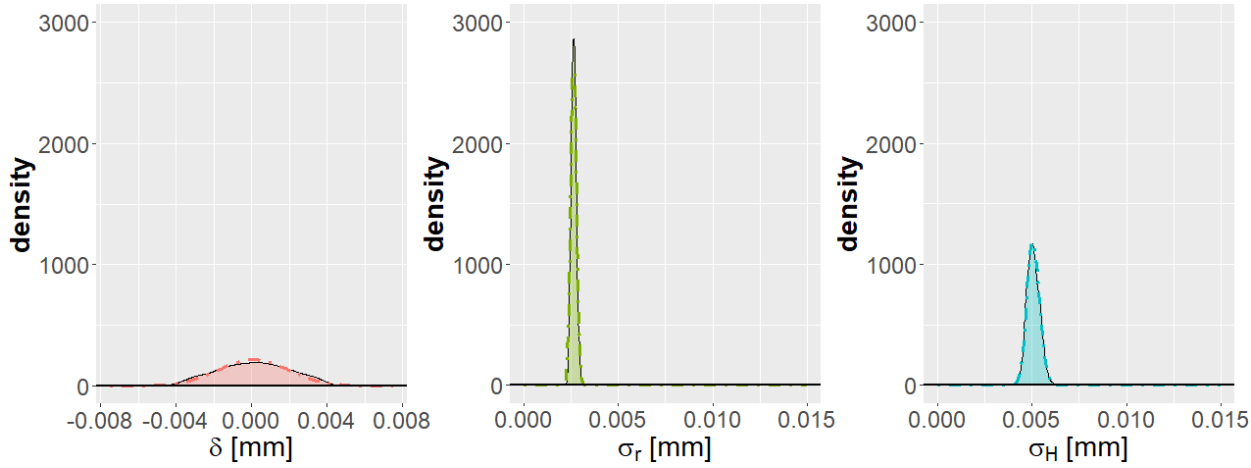

Figure 3.23: Marginal posterior density functions for bias δ, repeatability σ_r and homogeneity σ_H parameters

As expected, under a nested probability structure, repeatability and homogeneity parameters take advantage of the increased actual sample size. For them both we have tighter distributions (higher maximum value for the gamma probability density functions with respect to those in Figure 3.20) which means improved value for expanded precisions.

Finally, we can use the statistics 3.16 to the result in Figure 3.23 to complete the accuracy description with interval estimates. The outcome in Table 3-19 is no longer a table with q rows, like Table 3-16, because there is no more the distinction for each q-level of the reference accepted value factor. We have unique statistics, that are effective for the whole working measurement interval of the equipments population.

	p_D	DIC	$\hat{\delta}_{lower}$/mm	$\hat{\delta}_{upper}$/mm	U_r/mm	U_H/mm
95%	67.48	-2539	-0.0033	0.0035	0.0029	0.0056

Table 3-19: bias and expanded precisions under repeatability and homogeneity conditions for the full Hierarchical Bayesian model of the LVDT accuracy experiment

As already said, with respect to the synthesis that you can find in Table 3-16 at the end of section 3.4, the expanded precision evaluations benefit from the joint influence between the levels of the reference accepted value factor.

The manufacturer may now conclude that its testing equipments' family has the following accuracy description about the LVDT displacement measurements:

- trueness, evaluated in terms of bias, is in the range $[-3, 3]\ \mu m$;
- repeatability precision, evaluated in terms of expanded standard deviation is within $+3\ \mu m$;
- homogeneity precision, evaluated in terms of expanded standard deviation is within $+6\ \mu m$.

The clear distinction between trueness and precision descriptions is the greatest advantage of the Bayesian formulation. As a matter of fact in the Bayesian framework there is a greater distinction between the statistics associated to the concepts of trueness and precision with respect to the general linear model stated in the ISO 5725-2, where the trueness is described just by the estimate of the model general mean at which we may associate its confidence interval that is always based on the residuals mean square errors as also the repeatability and inter-laboratory standard deviations are. In conclusion, model 3.19 is able to describe the accuracy of the whole product line as what it had set out to achieve. Further researches are needed in order to give to the model more solid analytics foundations and to test it under a wider spectrum of sensitivity analysis that takes into account different sample sizes, others metrological domains, etc.

On the other hand, the manufacturers may now state up to which credible probability level are able to satisfy the ISO 306 requirement on the accuracy of displacement measurement. Without going into details of accuracy decomposition, in simple terms, $\hat{\theta}_1 \in \pm 4\ \mu m$ is assigned as requirement for trueness and $\hat{\theta}_2 \leq 6\ \mu m$ as requirement for homogeneity precision. Then, applying to the posterior samples that describe the accuracy parameters in Figure 3.23 the Empirical cumulative distribution function, that is a step function with jumps i/n at observation values, where

i is the number of tied observations at that value, it is possible evaluate the credible probability levels “C.P.” as reported in Table 3-20 below.

Accuracy	$\hat{\theta}$	C.P.
Trueness	±0.004 mm	99.5 %
Precision homogeneity	≤0.006 mm	99.5 %

Table 3-20: Credible Probability levels "C.P." of trueness and precision homogeneity requirements

In this way the accuracy evaluation frame has completely been turned upside down: from parameter estimation in term of fixed parameters with an assigned confidence level in the ISO 5725-2 framework, to credible probability estimation for specific target of variability encompassed by random variables describing the parameters in the proposed Bayesian framework.

The result of Table 3-20 is the cornerstone of the Instron requirement for the final validation of the penetration measurement accuracy on the new product line of VST test equipments. In this respect, it is recalled that ISO 306 requires to measure the penetration of the indenting tip into the test specimen to an accuracy of $\pm 0{,}01\ \mathrm{mm}$. It was decided to split the requirement as $\pm 0{,}004\ \mathrm{mm}$ for bias and $\leq 0{,}006\ \mathrm{mm}$ for precision. The final check is satisfied if the probability that the parameters (bias and precision taken individually) take a value in the required interval is higher than 0.95, consistently with their respective posterior distributions. Looking at the outcome in Table 3-20, the final check was fully satisfied. In this way the manufacturer knows that the risk that the shipped equipment does not meet the ISO 306 requirement is 0,5 %. Only in these few cases the adjustment of the displacement transducer is required, which implies an additional cost for the manufacturer. This industrial application is a first concrete acknowledgment of the goodness of the proposed model.

4 HIERARCHICAL BAYESIAN MODEL FOR PROFICIENCY TESTING

4.1 INTRODUCTION TO PROFICIENCY TESTING

A proficiency test is an inter-laboratory exercise aimed to determine the performance of participants in specific tests or measurements, and to monitor their continuing performance independently several times. The typical purposes of a proficiency test include the evaluation of laboratory performance, the identification of problems in laboratories, establishing effectiveness and comparability of test or measurements methods, the provision of additional confidence to laboratory customers, validation of uncertainty claims, and the education of participating laboratories [9].

In the last decades, proficiency testing has become an essential aspect of laboratory practice in all areas of testing, calibration and inspection. The testing schemes vary according to the needs of the sector in which they are used, the nature of the proficiency test items, the methods in use and the number of participants. However, in their simplest form, most proficiency testing possess the common feature of comparison of results obtained by one laboratory with those obtained by one or more different laboratories. So, the greatest common factor for the nature of this test method is the need of comparing performance, especially concerned with the assessment of participants and as such does not specifically address bias or precision.

The definition of proficiency testing is stated in ISO/IEC 17043 [40], whereas the ISO 13528 provides the requirements for the statistical design, validation criteria, review of results and reporting summary statistics [9]. The statistical techniques are consistent with other International Standards, particularly the ISO 5725 series, the ISO/IEC 98-3 (GUM) and the ISO/IEC Guide 99 (VIM). In its simpler scheme, performance statistics are typically derived by comparing the difference between a reported participant result and an assigned value with an allowable deviation or with an evaluation of the measurement uncertainty of the difference.

Although the proficiency scheme exhibits a strong Bayesian attitude, expressed in the statement of willing to compare participant results with respect to an assigned value, that remembers the prior value, the statistical design of ISO 13528 follows the frequentist formulation, based on the nature of errors assumption, the observed results and the nature of data. There are various types of data used in proficiency testing, including quantitative, categorical and ordinal. Here we consider continuous quantitative variables. In addition, we assume that the results from competent participants are approximately normally distributed.

In addition to what has already been mentioned, the basic approach for all purposes is to compare results on a proficiency test item x_i with the assigned value x_{pt}. Hence, usually for the performance evaluation, the difference is compared to an allowance for measurement error. The ISO 13528 standard indicates five ways to determine the assigned value x_{pt} [9, Par. 7]. The choice between these methods is the responsibility of the proficiency test provider. Once the assigned value has been determined, the approach for determining the evaluation criterion has to be chosen.

We adopt the ISO 13528:2015 notation. Accordingly, x denotes the measurement results, even if until now we have used y.

Coming back to the comparison criterions of ISO 13528:2015, a prescribed maximum permissible error may be used directly as δ_E (it is a threshold and not a parameter even if it is indicated by a Greek letter) for use with the estimates of deviation D_i, that is the simplest measure of participants' performance:

$$D_i = x_i - x_{pt} \tag{4.1}$$

The difference D_i may be also expressed as a percentage difference, calculated as:

$$D_i\% = 100\big(x_i - x_{pt}\big)/x_{pt}\,\% \tag{4.2}$$

The difference D_i or $D_i\%$ is usually compared with a criterion δ_E based on fitness purpose or with experience from previous rounds of a proficiency testing scheme. Use of D_i or $D_i\%$ generally

assumes symmetry of the distribution of participant results in the sense that the acceptable range is:

$$-\delta_E < D_i < \delta_E \qquad 4.3$$

The advantage of the deviance statistic is that participants have an intuitive understanding of the results, since they are tied directly to the measurement error. Instead, the disadvantages are that, being not standardized, it does not allow simple scanning of reports for action signals in proficiency testing scheme with multiple analyses or where fitness for purpose criteria can vary by level of the measurand. For all these purposes the z-score is preferred and it is calculated as:

$$z_i = \left(x_i + x_{pt}\right)/\sigma_{pt} \qquad 4.4$$

where σ_{pt} is the standard deviation for proficiency assessment, that can be interpreted as the population standard deviation of results from a hypothetical population of laboratories performing exactly in accordance with requirements [9]. If a regulatory requirement or a fitness-for-purpose goal is given as a standard deviation, then σ_{pt} may be used directly. Further details about the approaches for determining the standard deviation for proficiency testing are reported in the chapter 8 of ISO 13528.

The advantage of *z-score* with both these approaches for continuous schemes are:

a) performance z-scores have a consistent interpretation in terms of fitness for purpose from one round to the next;
b) performance z-scores are not subjected to the variation expected when estimating dispersion from reported results.

Otherwise, the value of the standard deviation σ_{pt} for proficiency assessment can be derived from a general model for the reproducibility of the measurement method. This method has the advantage of objectivity and consistency across measurands, as well as of being empirically based. It can be used for data from a previous collaborative study or from data obtained in the same round of a proficiency testing scheme. In the latter case, the value so obtained is called consensus value. This procedure invalidates the long-term monitoring, due to the fact that the comparability from year to year depends on the participants as well as on the used reference materials. Therefore, it is difficult for a participant to use values of z-score to look for trends over several proficiency rounds. Moreover, the standard deviation σ_{pt} can be unreliable when the number of participants is small. Lastly, when x_{pt} and σ_{pt} are calculated from participant results, the performance score is correlated with individual participant results, because individual results have an impact on both.

Besides, when there is concern about the uncertainty of the assigned value $u(x_{pt})$, for example as the clause 4.5 is satisfied,

$$u(x_{pt}) > 0{,}3\sigma_{pt}. \qquad 4.5$$

then the uncertainty can be taken into account by expanding the denominator of the performance score. This statistic is called z'-score and it is calculated as follows:

$$z'_i = \frac{x_i + x_{pt}}{\sqrt{\sigma_{pt}^2 + u^2(x_{pt})}} \qquad 4.6$$

The conventional interpretation of both z-scores and z'-scores is the following (see [40, Par. B.4.1.1]):

- $|z| < 2$ is considered to be acceptable;
- $2 < |z| < 3$ is considered to give a warning signal;
- $|z| > 3$ is considered to be unacceptable.

The justification of the usage of the limits of 2 and 3 for z-scores is easily understood. Measurements that are carried out correctly are assumed to generate results that can be described by a normal distribution with mean x_{pt} and standard deviation σ_{pt}. Hence, the z-scores will follow a

standard normal distribution. Under these assumptions, only about 5 % of scores would be expected to fall outside the range $|z| < 2$, and only about 0,3 % would be expected to fall outside the range $|z| < 3$. Because the probability of z-scores falling outside ±3 is so low, it is likely that there is an identifiable cause for the anomaly. Moreover, δ_E is closely related to σ_{pt} as used for z-scores. The relation is determined by the evaluation criterion for the proficiency assessment, so that $\delta_E = 3\sigma_{pt}$ sets the boundary within which the deviation in equation 4.1 is considered acceptable.

Finally, it is common, within a single round of a proficiency scheme, that more than one test item or measurand are included by mixing materials with different known levels of a property in specified proportions. In all these applications the designed levels are included such as to investigate the trends of random or systematic errors. For example, two similar proficiency test items may be used with the intention of treating them through a Youden plot [9]. Anyway, in this situation the results for each proficiency test item or for each measurand should be evaluated separately.

In the next sections we apply the proficiency testing scheme to the application case of the displacement measurements with the transducers of the VICAT test equipments. First, we follow ISO 13528. Then, we apply the Bayesian framework.

4.2 A SIMULATED PROFICIENCY TEST ON THE VST DEFLECTION MEASUREMENTS

In real proficiency tests, it is common that participants use testing equipement of different manufactures even if the test item or the measurand is the same. So, we cannot straightly use the previous dataset of the VST accuracy experiment. In order to better simulate the reality of proficiency testing schemes, we decided to enlarge the dataset in annex A (chapter 6) by adding data consistent with other levels of another factor for the manufacturer brand and for simplicity we focused on the first two levels of the reference accepted value factor, that are the jumps “3»2” and “4»3”. Therefore, from the VST accuracy experiement dataset we drawn out only the first level of the new factor, called "manufacturer" and, in order to add the other two levels, we resort to the random generation of numbers according to two different distributions. The former is supposed to describe the most competent testing equipments from an unknown manufacturer, and it is said “Best”. The latter is supposed to describe biased testing equipments from an unknown manufacturer, and it is said “Biased”. A normal distribution with mean equal to 1 mm and standard deviation equal to 5 µm describes the former. The latter is represented by a normal distribution having mean value of 1,006 mm and standard deviation of 4 µm. From each of these distributions $p = 18$ values $x_{.iz}$ are drawn in order to represent p different laboratories, described by normal distributions having the drawn value $x_{..z}$ as mean and the same standard deviation of the manufacturer level. In addition, for each laboratory $n = 4$ values x_{kiz} are drawn out according to the normal distribution having the specific laboratory mean value $x_{.iz}$ and standard deviation 3 µm. So, for each “jump” in formula we have:

$$\begin{aligned} x_{kiz} &\sim Norm(x_{.iz}, 0.003) \\ x_{.iz} &\sim Norm(x_{..z}, 0.005) \\ x_{..2} &\sim Norm(1.000, 0.005) \\ x_{..3} &\sim Norm(1.006, 0.004) \end{aligned} \qquad 4.7$$

where $z = 1,2,3$ is the subscript for the manufacturer brand, and, following the z numeric order, we have “Instron”, “Best” and “Biased”. This dataset in the rows below is called “simulated” in order to underline its origin from drawing of numerical random values, that are computer simulated. The simulated dataset was achieved through the random data generator of the software "R" and the reader can look the full script up in the annex D (chapter 9). The proficiency dataset so achieved is depicted in Figure 4.1 for the LVDT jump 3»2 and Figure 4.2 for jump 4»3.

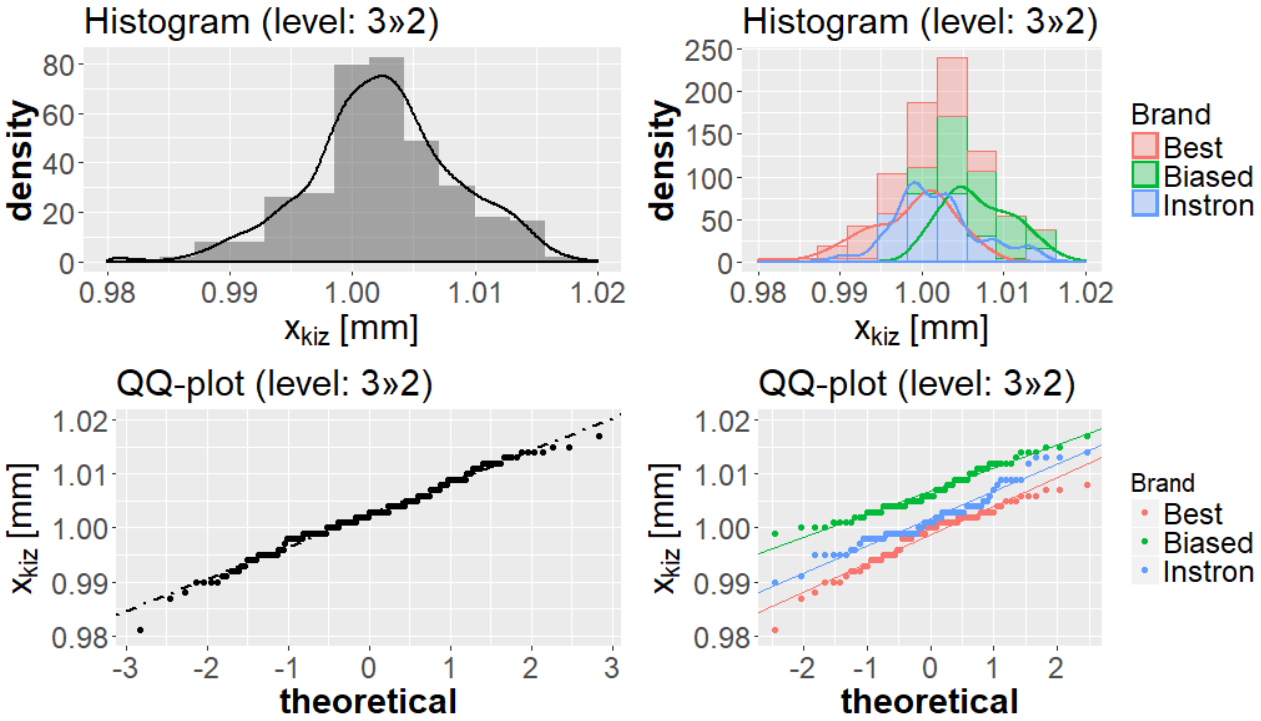

Figure 4.1: visual review for the level “3»2” of the simulated proficiency dataset

The visual check of data in Figure 4.1 and Figure 4.2 confirms the expected distribution of results. As a matter of fact, the histograms, useful and widely available data review tools, do not show any multimodal or asymmetric attitude in the density distribution. Likewise, the Q-Q (quantile-quantile) plots, a graphical method to compare the data distribution against the corresponding theoretical normal, show a good fit with the normality assumption.

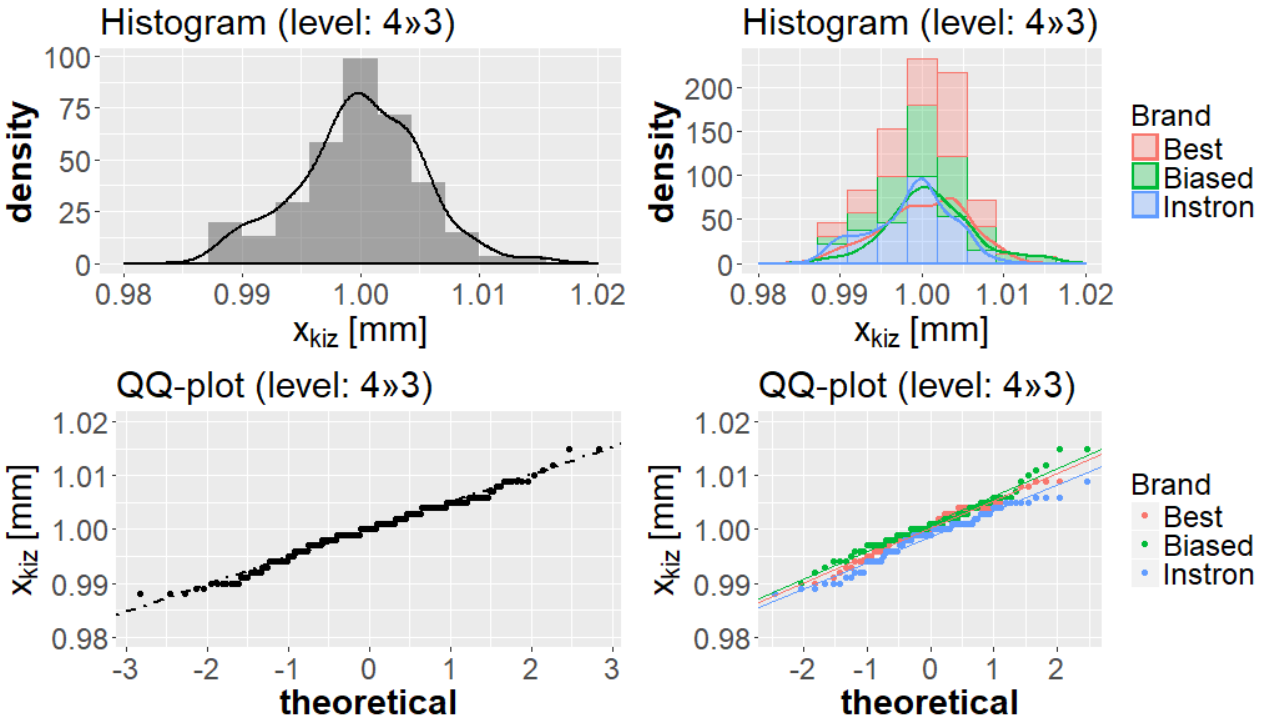

Figure 4.2: visual review for the level "4»3" of the simulated proficiency dataset

Besides, looking at the data of the level "3»2" (Figure 4.1) we see wider deviations in terms of both mean value and standard deviation (looking at the slopes of the QQ-lines) between the k manufacturers than those of level "4»3" (Figure 4.2).

Once the goodness of data fit with respect to the model assumptions has been checked, we can move to the proficiency performance evaluation with respect to the claimed expected value and standard deviation, which in this specific case are those used for the random numerical data drawing according to $x_{\_2}$ in equation 4.7. We are of course in a convenient position, as we know the best values, having simulated two-thirds of the dataset. Differently, we would act according to one of the approaches suggested by the ISO 13528 standard to determine the assigned mean value x_{pt} and the assigned standard deviation σ_{pt}. Summing up, for the simulated proficiency testing the assigned value attributed to both test items is:

$$\begin{aligned} x_{pt}^{3\text{»}2} &= x_{pt}^{(1)} = 1 \text{ mm} \\ x_{pt}^{4\text{»}3} &= x_{pt}^{(2)} = 1 \text{ mm} \end{aligned}, \qquad 4.8$$

and the measure of dispersion for proficiency assessment of both test items is:

$$\begin{aligned} \sigma_{pt}^{3\text{»}2} &= \sigma_{pt}^{(1)} = 0.005 \text{ mm} \\ \sigma_{pt}^{4\text{»}3} &= \sigma_{pt}^{(2)} = 0.005 \text{ mm} \end{aligned}. \qquad 4.9$$

We can now proceed to the statistical analysis for performance evaluation, being aware that data are consistent with the statistical assumptions for z-scores criteria. For dissertation simplicity, we focus only on the z-score in 4.4. The assessed dataset has n replicated measurements for each of the i, k participant laboratories. The subscript notation is a bit misleading: the k-measurements x_{k11} and x_{k12} are drawn up from two different laboratories even if the i-laboratory subscript is the same. As a matter of fact, every single laboratory is identified by the pair (i, z), where z adds the information about which kind of equipment the laboratory uses, or in other words

which manufacturer made the test equipment. That means that we are managing n replicated measurements by t different laboratories, where t is:

$$t = p \cdot r = 54 \qquad 4.10$$

Hence, the notation can be arranged by suppressing the k subscript and having $i = 1, \dots, t$.

We can calculate the expected value $x_{.i}$ for each participant laboratory and the associated standard deviation. The results are summarized in Figure 4.3, where the effect of the bias deliberately introduced in the random sampling can be appreciated, especially for the test item corresponding to the "jump 3»2". The two grey dashed and solid lines in Figure 4.3 delimit the $x_{pt} \pm 2\sigma_{pt}$ and $x_{pt} \pm 3\sigma_{pt}$ regions, respectively. Only the mean value for the LVDT with serial number 155991 in "jump 3»2" is on the threshold for unacceptable values.

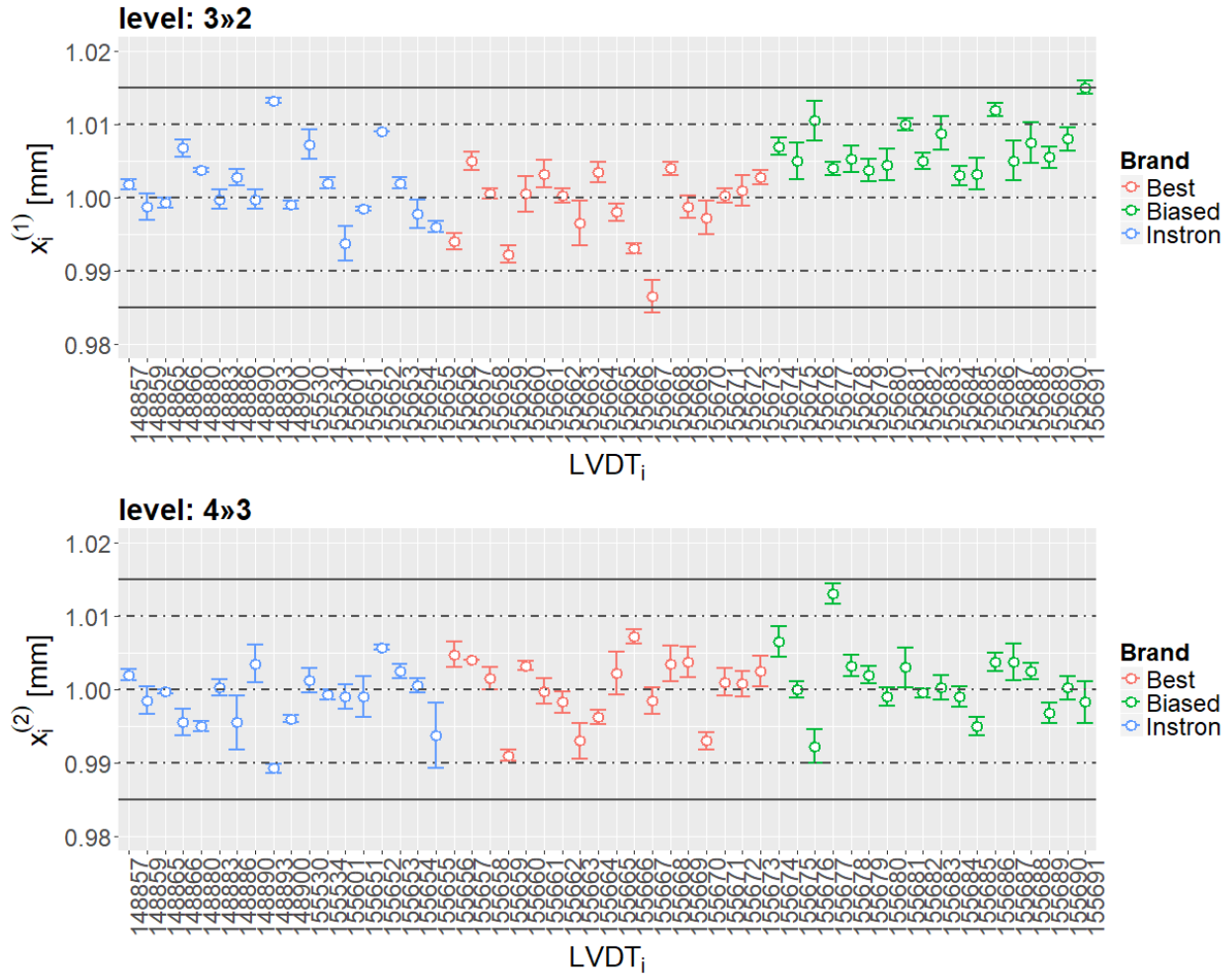

Figure 4.3: lab mean value ± lab standard deviation for each proficiency test item

In this simulated proficiency test we have two different test items, one for each level of the variable "jump", having the same expected value and best standard deviation as stated in equations 4.8 and 4.9 respectively. The only difference is the subset used over the whole LVDT measurement interval. In order to distinguish between the two different test items, a superscript was added to the laboratory mean in Figure 4.3 and, similarly for the calculated z-scores.

Combined performance scores for results from different proficiency test items can be useful for detecting persistent bias. Here, having two similar proficiency test items, the Youden Plot provides a very informative graphical method of studying the results [9]. The Youden graph is constructed by plotting the participant z-scores obtained on one of the proficiency test items against the other one. Using this graphical tool, the z-scores of the LVDT proficiency case are depicted in Figure 4.4.

Inspection of Figure 4.4 reveals two participants (LVDT numbers 148893 and 155691) in the bottom right-hand quadrant out from the solid line, and therefore they could have poor calibration having a high z-score on proficiency test item 3»2 and a negative score on proficiency test item

4»3. Overall, the real measurements, that are those from the Instron test equipments, are close to the origin, as it is for the “Best” population. Instead, the “Biased” population has an evident bias on the proficiency test item of the level 3»2, and the cloud of data is shifted to the right on the Youden graph of Figure 4.4. Concluded the visual review, we can move to the next section facing up the Bayesian formulation of the proficiency testing scheme.

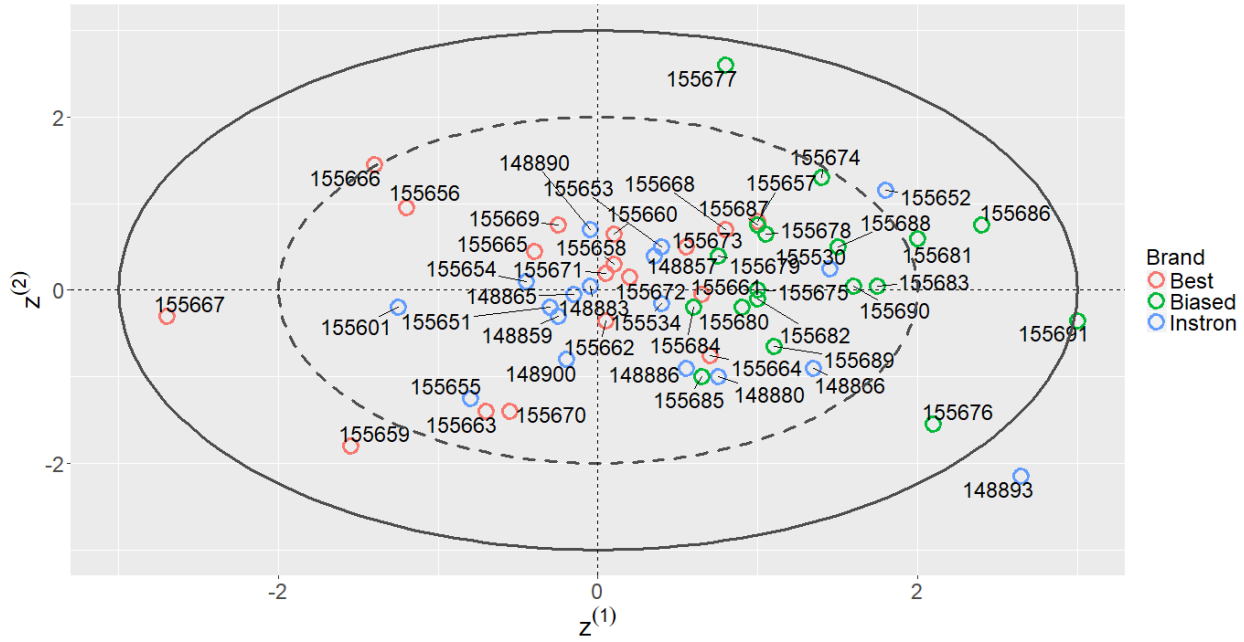

Figure 4.4: Youden Plot for the z-score of the jump 3»2, that is $z^{(1)}$, against the ones of the jump 4»3, that is $z^{(2)}$

4.3 HIERARCHICAL BAYESIAN MODEL FOR PROFICIENCY TESTING

In chapter 3 we have dealt with hierarchical models comprising three levels:

i. a likelihood conditional on unit-specific parameters;
ii. exchangeability assumptions for the unit-specific parameters;
iii. prior distributions for the population parameters, also referred to as hyperparameters.

There is, of course, no reason to stop us considering further application fields for the models previously introduced. Therefore, in this section we model the proficiency testing scheme for the simulated VST experiment according to a possible Bayesian formulation. The benchmark is the results achieved through the ISO 13528 scheme in section 4.2.

In a manner corresponding to the data structure of the simulated proficiency test on the VST deflection measurements (see left side of Figure 4.1 and Figure 4.2) and as implicitly stated in equation 4.7, we can write the following conditional probability structure based on the exchangeability assumptions:

$$\begin{aligned} x_{ki}^{(j)}|\chi_{\cdot i}^{(j)},\sigma_r^{(j)} &\sim i.i.d.\ \ Normal\left(\chi_{\cdot i}^{(j)},\sigma_r^{(j)}\right) \\ \chi_{\cdot i}^{(j)}|\chi_{\cdot\cdot}^{(j)},\sigma_L^{(j)} &\sim i.i.d.\ \ Normal\left(\chi_{\cdot\cdot}^{(j)},\sigma_L^{(j)}\right) \\ \chi_{\cdot\cdot}^{(j)} &\sim Normal(1,0.005) \quad \forall j \\ \sigma_r^{(j)} &\sim Uniform(0,0.006) \quad \forall j \\ \sigma_L^{(j)} &\sim Uniform(0,0.010) \quad \forall j \end{aligned} \tag{4.11}$$

where $k = 1,\dots,n$ is the subscript for the replicated measurement of each i-participant laboratory, so $i = 1,\dots,t$ is the subscript for the specific laboratory and $j = 1,2$ is the superscript for the specific proficiency test item. Looking at the prior assumption for the overall mean parameter $\chi_{\cdot\cdot}^{(j)}$ in 4.11, we see that is assumed to be coherent with the assigned value $x_{pt}^{(j)}$ of equation 4.8 and with the assigned standard deviation $\sigma_{pt}^{(j)}$ of equation 4.9. The same probability structure is depicted in Figure 4.5 through the corresponding Doodle graph, where the test item superscript was omitted for the sake of simplicity.

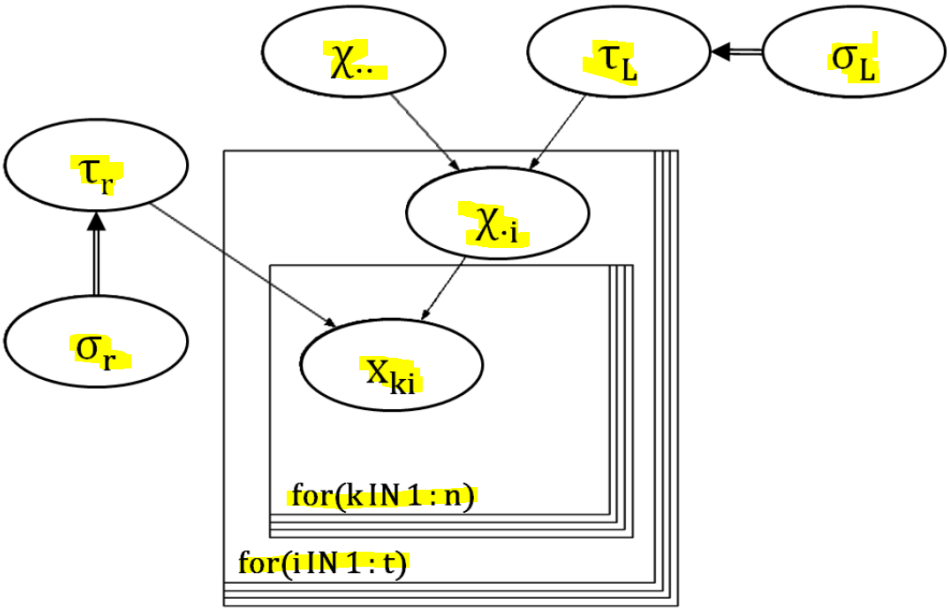

Figure 4.5: Hierarchical Bayesian probability structure for Proficiency testing scheme

As already mentioned, the proficiency scheme according to ISO 13528 has a strong Bayesian attitude, expressed in the statement of willing to compare participant results to an assigned value $x_{pt}^{(j)}$, for each j-test item. The Bayesian formulation in Figure 4.5 solves the conflict between proficiency attitude and ISO 13528 suggested statistics. Actually, we are able to include the information about the assigned value directly in the model through the choice of the proper prior

distribution for the hyper-parameter $x_{pt}^{(j)}$. The Bayesian rule propagates this information through the probability structure and the chosen prior affects the posterior distributions $x_i^{(j)}$ of each specific laboratory and, then, the final laboratory z-scores.

So, the hierarchical Bayesian model is written according to the standards of the OpenBUGS as

```
model {
        x ~ dnorm(1, 40000)
        tau.L <- pow(sigma.L, -2)
        sigma.L ~ dunif (0, 0.01)
        tau.r <- pow(sigma.r, -2)
        sigma.r ~ dunif (0, 0.006)

        for (i in 1:t) {                                          4.12
                        for (k in 1:n) {
                                        xki[k,i] ~ dnorm (xi[i], tau.r)
                                        }
                        xi[i] ~ dnorm (x, tau.L)
                        }
        }.
```

Then, to use the "R2OpenBUGS" package the next step is to arrange the simulated dataset as the OpenBUGS standard needs, that is the "list" R object in 4.13, having $n \times t$ dimension.

```
$t
[1] 54

$n
[1] 4

$xki                                                              4.13
      [,1]  [,2]  [,3]  [,4]  [,5] … [,50] [,51] [,52] [,53] [,54]
[1,] 1.000 1.003 0.998 1.005 1.004 … 1.000 1.002 1.006 1.007 1.015
[2,] 1.003 0.996 0.999 1.005 1.003 … 1.011 1.011 1.004 1.006 1.017
[3,] 1.002 0.999 0.999 1.009 1.004 … 1.003 1.005 1.009 1.007 1.013
[4,] 1.002 0.997 1.001 1.008 1.004 … 1.006 1.012 1.003 1.012 1.015
```

As in the previous cases, the starting points are randomly selected from tighter proper prior distributions, where tighter means having a reduced scale parameter. Eventually, the Bayesian numerical computation is performed using the "bugs" function of the "R2OpenBUGS" package.

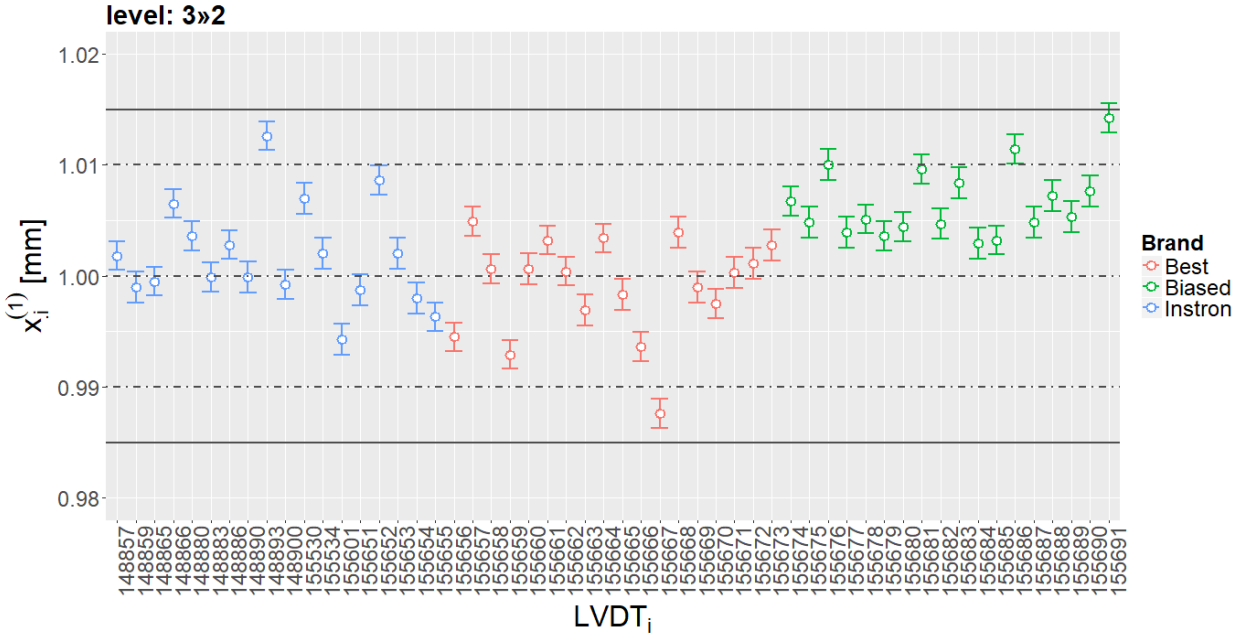

Figure 4.6: credible intervals (mean value ± standard deviation) for the laboratory specific parameters of the first proficiency test item, that is the "jump" 3»2

The results in terms of mean value (the point) and standard deviation (the error bars) of the posterior distributions for the specific laboratory parameters $\chi_{\cdot i}^{(j)}$ are shown in Figure 4.6 and Figure 4.7.

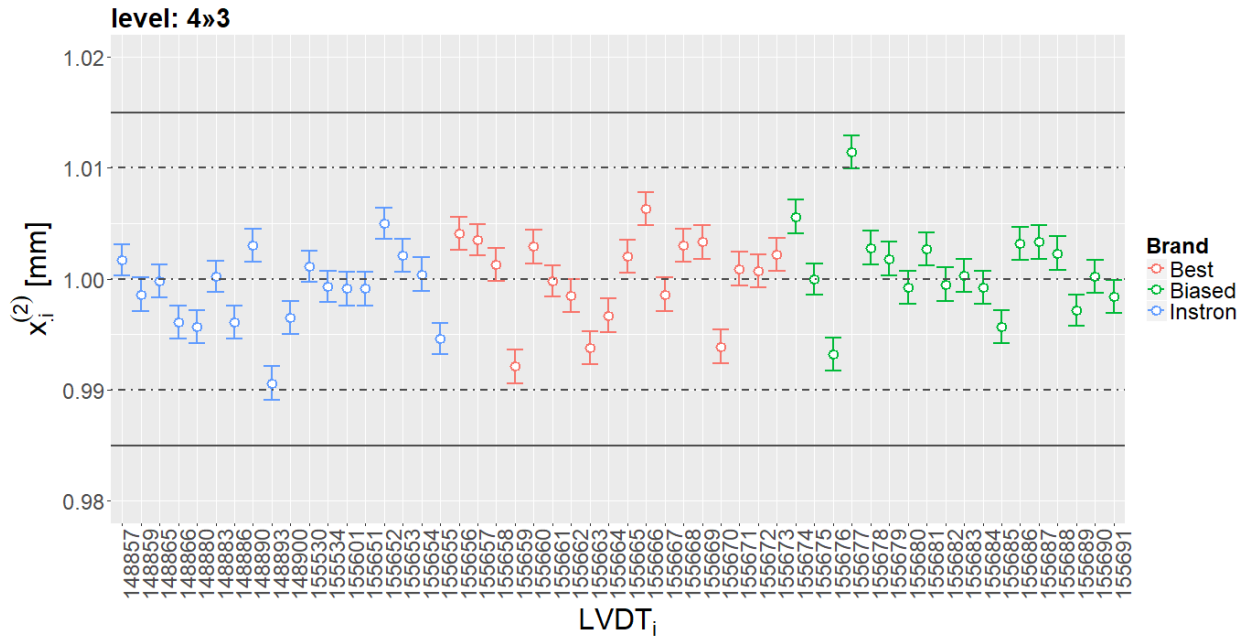

Figure 4.7: credible intervals (mean value ± standard deviation) for the laboratory specific parameters of the second proficiency test item, that is the "jump" 4»3

Going into the details of the Bayesian result, we the effects of fitting a hierarchical model based on exchangeability assumptions are enhanced strength, global smoothing of uncertainty and shrinkage to the mean. The first two effects can be deduced looking at the width of the error bars: under the exchangeability assumption each posterior borrows precision from the others via the joint influence on the estimation of the underlying population parameters. So, a side effect of this borrowing strength is that the variability about each specific parameter is spread more evenly across the laboratories, as can be observed from the more uniform width of credible intervals respect of the width of the error bars in Figure 4.3. The last effect, that is the shrinkage towards the mean, tends to pull extreme values towards the population mean, as can be seen especially for LVDT s/n 148893, 155667 and 155691 in Figure 4.6 compared to Figure 4.3.

Shrinkage towards the mean can initially appear somewhat undesirable, because we expect that it can affect the deviation scores and, consequently, the z-scores. We expect that this negative effect especially affects the result of the LVDT s/n 155691 for the first test item. Actually, the result lies on the solid line in Figure 4.3, whereas in Figure 4.6 it is farther from that threshold, within the zone between the solid line and dot-dash line.

Shrinkage is reflected in the Youden plot for the z-scores calculated according to equation 4.4 (Figure 4.8), where the assigned mean value x_{pt} and the assigned standard deviation σ_{pt} are those stated in equations 4.8 and 4.9, respectively. In fact, even if the Youden plot preserves the overall relative positions between z-scores, these are pulled in towards the origin of the plan and, as a consequence, the z-score of the LVDT s/n 155691 is pulled within the solid line.

The risk of misclassification can be reduced through a proper setting of the probability level for the credibility interval of the among-laboratories variability parameter $\sigma_L^{(j)}$, with which the two thresholds can be alternatively achieved instead of keeping on using ±2 and ±3. The thresholds optimization is beyond the scope of this research. Further studies need to be carried out in order to address the topic of how avoiding the misclassification if the proficiency statistics move into the Bayesian framework. At the moment, we merely seek to show that such proficiency statistics in the Bayesian framework are biased due to the shrinkage-to-the-mean effect. Further investigations would certainly help to provide a better sight of the concept.

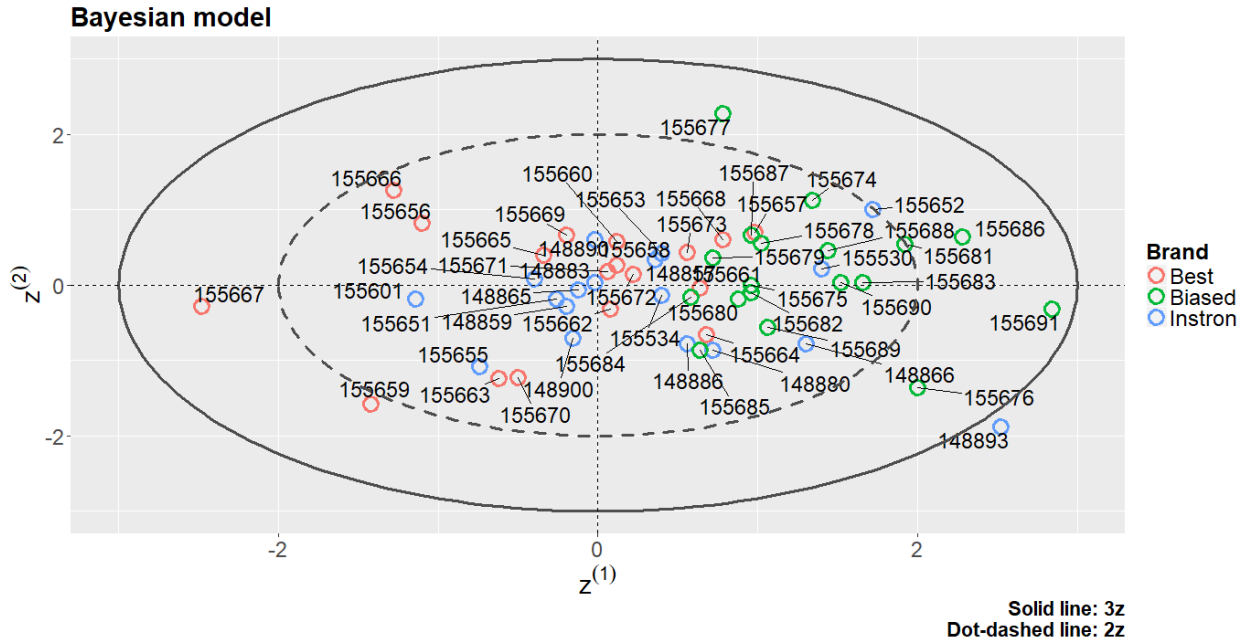

Figure 4.8: Youden Plot for the z-score of the jump 3»2, that is $z^{(1)}$, against the ones of the jump 4»3, that is $z^{(2)}$

4.4 BAYESIAN TECHNIQUES FOR DATA CONSISTENCY CHECKING

ISO/IEC 17043 and IUPAC Harmonized Protocol [41] recommend removing obvious blunders from a data set at the early stage of the analysis, prior to use of any test to identify statistical outliers. Actually, obvious blunders, such as reporting results in incorrect units or switching results from different proficiency test items, occur in most rounds of proficiency testing, and these results impair the performance of subsequent statistical methods. In these cases a visual review of the data, as the one performed in section 4.2, should be enough.

Then, in accordance with ISO 13528, outlier detection may be performed either to support visual review for anomalies or, coupled with outlier rejection, to provide a degree of resistance to extreme values when calculating summary statistics. In this respect, the standard suggests to apply the following techniques and tests, provided by ISO 5725-2:

- two measurements for graphical consistency techniques, called Mandel's h and k;
- Cochran's numerical outlier test;
- Grubb's numerical outlier test.

In addition, the standard provides recommendations for the level of confidence appropriate for outlier rejection in inter-laboratory studies for the determination of precision of test methods. In particular, it recommends rejection only at the 99 % level unless there is other strong reason to reject a particular result.

Before going ahead and introducing the Bayesian formulation for some consistency techniques and outlier detection tests, we emphasise that the techniques described in this chapter are more informal than those of the inferential methods. Classical statistical models generally depend on assumptions such as a linear model for error structure and normality distribution. Here we adapt these ideas to be generically applicable in Bayesian models, noting that this approach means that parameters have distributions and so, for example, residuals and deviations will be quantities with posterior distributions. In addition, within a Bayesian framework it is necessary to check for sensitivity to the prior and for conflict between prior and data.

Let us start and see Mandel's h statistic in order to understand if and under which assumptions it can be used in the Bayesian framework.

Mandel's h statistic is a simple indicator of relative deviation for grouped sets of observations. Let a set of observations $x_i^{(j)}$ be supposed as realizations of random variable $X_i^{(j)}$ and to be identically and independently distributed according to the normal distribution $\mathcal{N}(\mu, \sigma)$, where $i = 1, \dots, t$ denotes the group (usually the laboratory mean over n replicated measurements) and $j = 1, \dots, q$ is the test item index. The Mandel h statistic is given by:

$$H_i^{(j)} = \frac{X_i^{(j)} - \bar{X}^{(j)}}{\sqrt{\frac{1}{t-1}\sum_{i=1}^{t}\left(X_i^{(j)} - \bar{X}^{(j)}\right)^2}} \qquad 4.14$$

This is also called the between-laboratory consistency. Thompson [42] derived the exact distribution of the Mandel's h statistic. This distribution is symmetrically distributed around zero and the two-sided critical values in ISO 5725-2 can be also got as [43]:

$$h_{t;1-\alpha/2} = \frac{(t-1)t_{\nu;1-\alpha/2}}{\sqrt{t\left(t-2+t_{\nu;1-\alpha/2}^2\right)}} \qquad 4.15$$

where $t_{\nu;1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the t-distribution with $\nu = t-2$ degrees of freedom. This distribution has been implemented with the "qmandelh" function of the R-package "metRology".

Before moving to the Bayesian framework, we build the benchmark by applying the criterion as stated in ISO 5725-2 to the simulated proficiency dataset shown in section 4.2.

Mandel statistics are traditionally plotted for inter-laboratory study data, grouped by laboratory, to give a rapid graphical view of laboratory bias and precision. This plot produces a grouped, side-by-side bar plot. The final outcome for the Mandel h statistic applied to the simulated proficiency dataset according to ISO 5725-2 is depicted in Figure 4.9, where the solid lines encompass data

within a confidence level equal to the 99 %, whereas the dot-dashed lines within the 95 %. Therefore, we have:

- high consistency within the dot-dashed lines, that, given the symmetry of the $H_i^{(j)}$ distribution, are equal to $\pm h_{52;0.975} = \pm 1.934$;
- low consistency outside the solid lines, that, given the symmetry of the $H_i^{(j)}$ distribution, are equal to $\pm h_{52;0.995} = \pm 2.507$;
- medium consistency elsewhere.

Various patterns appear in the h plot. The majority of LVDTs tend to have both positive and negative h values at different test items, and in general the number of LVDTs giving negative values is approximately equal to those giving positive value, with the sole exception for the test item 3»2 of the "Biased" group of test equipments, that are all positive according to their random sampling strategy. Looking at the empirical data, i.e. those from Instron, only the serial number 148893 requires further investigation: it seems to be affected by a systematic error, arisen during the LVDT calibration procedure.

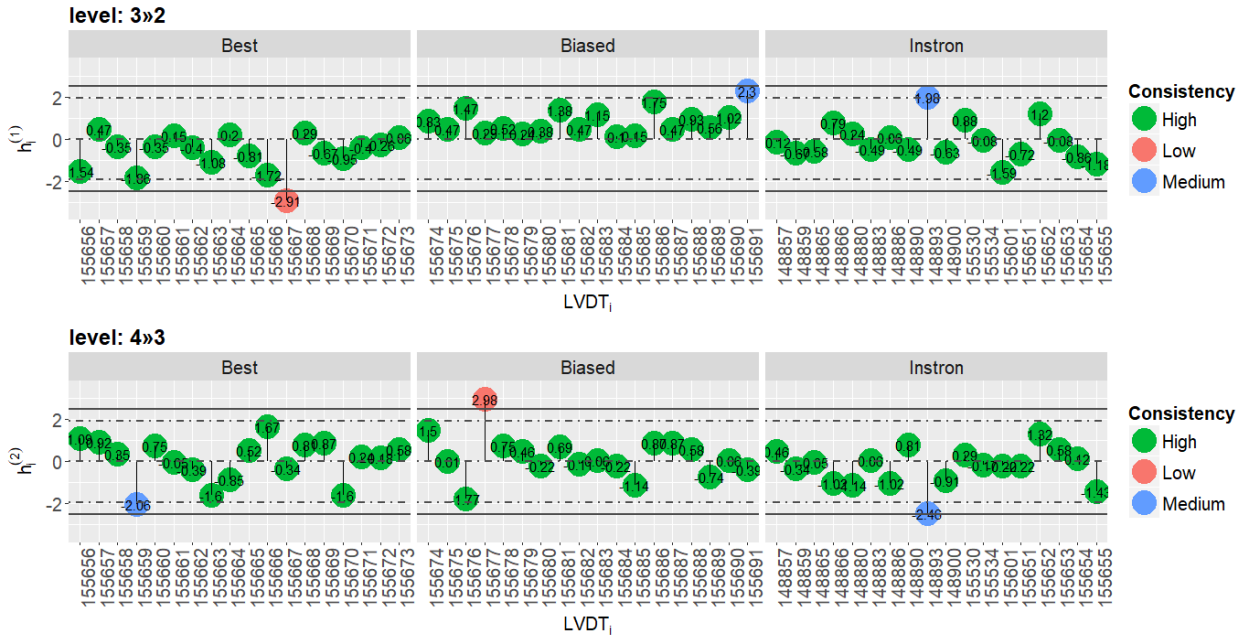

Figure 4.9: Mandel's between-laboratory consistency statistic according to ISO 5725-2, grouped by LVDT serial number, for the simulated proficiency testing of displacement measurements in VST apparatus

Now, we can move to the Bayesian formulation for the between-laboratory consistency statistic. Looking at the probability structure 4.11 for the Proficiency testing scheme, unless there are unavoidable changes due to the Bayesian updating, we have that $\chi_{\cdot i}^{(j)}$ are normal random variables, identically and independently distributed thanks to the exchangeability assumption. This is the same assumption behind the Mandel's h statistic for the between-laboratory consistency. Leveraging this similarity, we keep in the Bayesian framework equations 4.14 and 4.15, where the root of the observed data - the sample means of the replicated measurements within each of the LVDT groups in ISO 5725-2 - are in the Bayesian framework the mean of the posterior normal distribution for the specific LVDT parameters. Before using consistency and outlier detection techniques, the validity of the assumptions underlying the test, the assumption of normality in particular, should be demonstrated.

So, going back to the simulated proficiency test on the VST deflection measurements, first of all we have to check the consistency with the normality assumption for the posterior distributions of the specific LVDT parameters and, then, we can calculate the Mandel's h statistic. Figure 4.10 depicts the posterior distributions of all the specific LVDT parameters $\chi_{\cdot i}^{(j)}$ for each of the two test items and their population distribution that is a normal with parameters $x_{\cdot\cdot}^{(j)} = E\left(\chi_{\cdot\cdot}^{(j)}\right)$ and $s_L^{(j)} = E\left(\sigma_L^{(j)}\right)$. The $\chi_{\cdot i}^{(j)}$ bell shapes (coloured) look all very similar to normal distributions and their

population distribution (black dot-dashed) properly encompasses them all. Furthermore, the $\chi_{\cdot i}^{(j)}$ posterior distributions seem to have the same scale parameter and to be properly drawn from the marginal posterior distribution of the overall mean parameter. The normality assumption for the $\chi_{\cdot i}^{(j)}$ posterior distributions is double-checked with the Q–Q (quantile-quantile) plot (Figure 4.11), where the black dot-dashed line is the theoretical trend for a normal distribution having the mean value of the overall mean posterior distribution $\chi_{\cdot\cdot}^{(j)}$ as location parameter and the square root of the average variance of the specific LVDT distributions $\chi_{\cdot i}^{(j)}$ as scale parameter. Each coloured line is the simulated posterior set of realizations for a specific LVDT parameter.

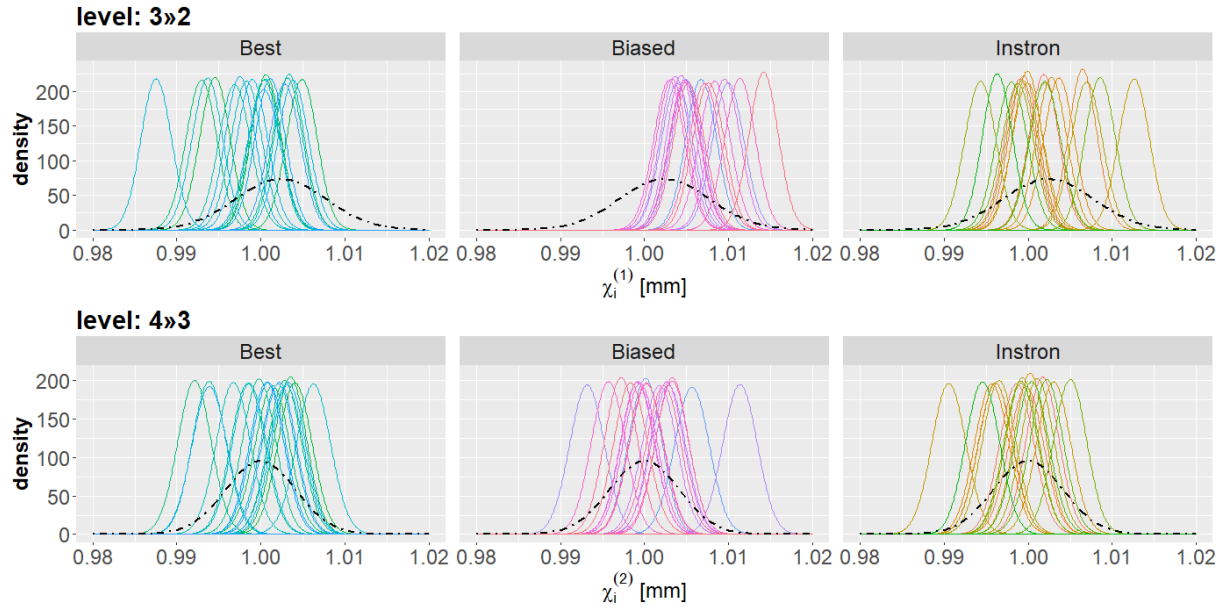

Figure 4.10: Posterior probability density functions for the specific LVDT parameters $\chi_{\cdot i}^{(j)}$, that are the colored curves, and for their population distribution $Normal\left(x_{\cdot\cdot}^{(j)}, s_L^{(j)}\right)$, that is the black dot-dashed curve

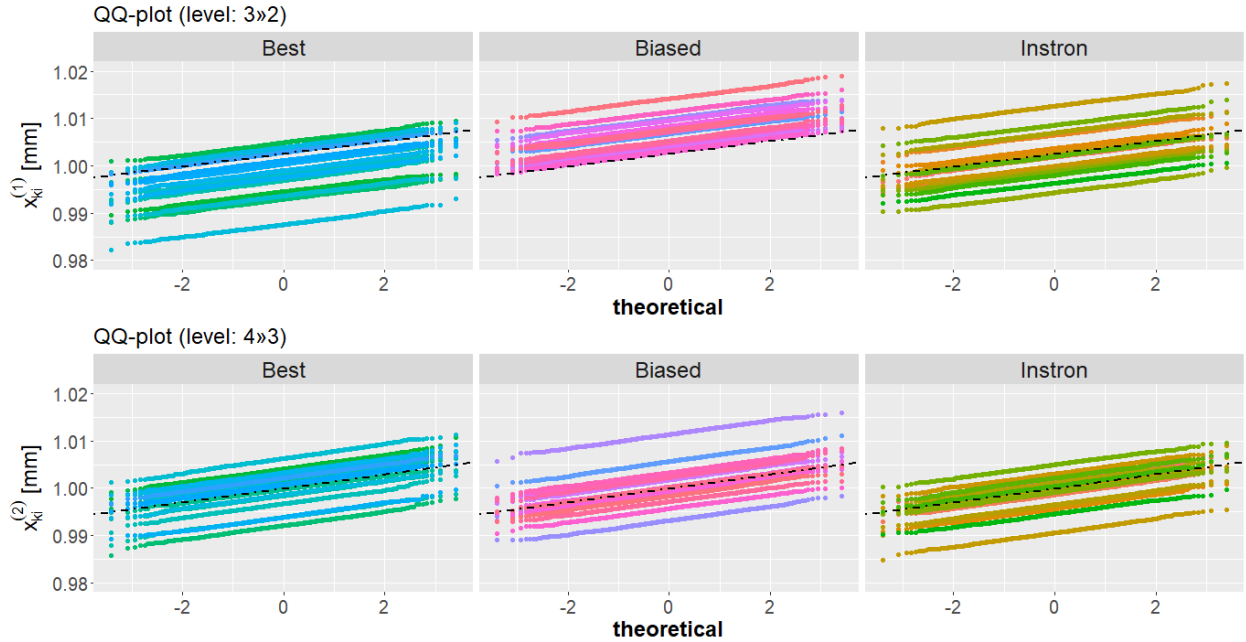

Figure 4.11: Q-Q plot of the posterior distributions for the specific LVDT parameters (one color for each LVDT) against normality assumption (black dot-dashed line)

Even more important for checking the assumptions underlying Mandel's h statistic, it is the identification of substantive departures from normality on the expected values of posterior probability distributions of the specific LVDT parameters with respect to their population

distribution. In other words, we ought to verify that $x_{\cdot i}^{(j)} = E\left(\chi_{\cdot i}^{(j)}\right)$, as stated in equation 4.11. To this end, the Q–Q probability plot was obtained for a normal distribution having as mean parameter the expected value of the posterior distribution for the assigned value parameter $x_{\cdot\cdot}^{(j)} = E\left(\chi_{\cdot\cdot}^{(j)}\right)$ and as standard deviation the expected value of the posterior distribution for the inter-laboratory variability $s_L^{(j)} = E\left(\sigma_L^{(j)}\right)$. The outcome is depicted in Figure 4.12.

Overall, we observe a good fit with respect to both normality assumptions. In addition, the examination of the plots in Figure 4.11 indicates that the specific LVDTs exhibit differences just on the shift parameter, as all the coloured lines share the same slope of the theoretical line, with different intercepts. The plots of Figure 4.12 support the applicability of the Mandel's h statistic in a Bayesian framework, thanks to the graphical confirmation of the normality assumption. Moreover, Figure 4.12 shows that the second test item has a better inter-laboratory variability, being the slope of the q-q line (red dot-dashed line) of the level "4»3" lower than that of the level "3»2". We also see that the posterior realization of the first test item is biased, being the intercept of the q-q line greater than 1 mm.

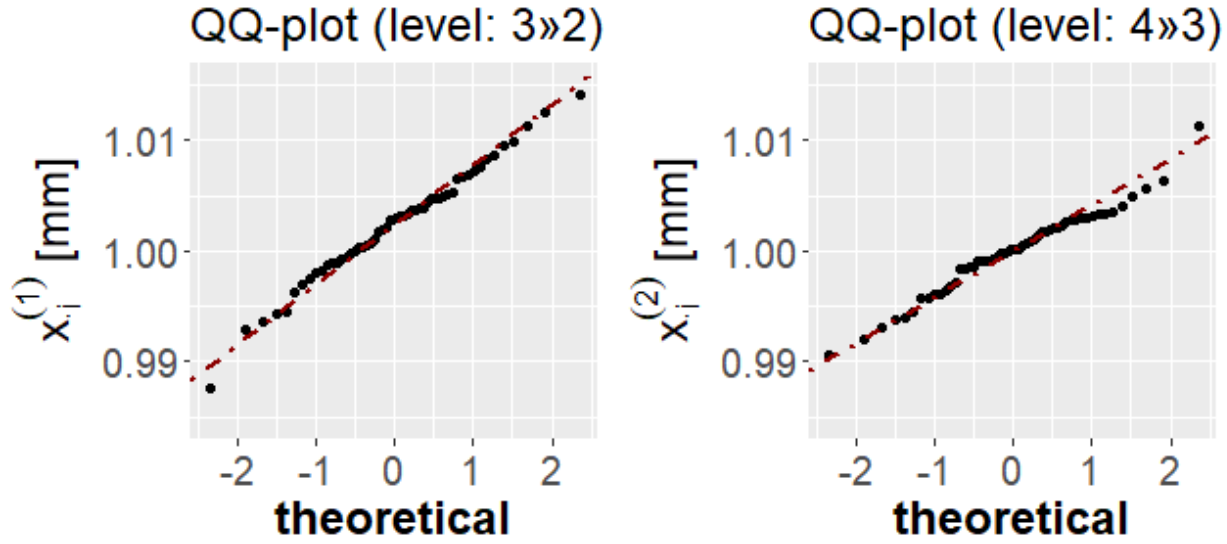

Figure 4.12: Q-Q plot of the posterior distributions for the expected values $x_i^{(j)}$ of the posterior distribution of specific LVDT parameters against normality assumption for their population (the red dot-dashed line is the theoretical trend)

Therefore, consistently with the posterior probability structure, we consider the mean value of the posterior specific LVDT distributions as the realizations of their population distribution. Accordingly we use the Mandel's h statistic in its form 4.14, with the expected value of the posterior distributions for the specific LVDTs $x_{\cdot i}^{(j)}$ replacing the sample mean $\bar{x}_i^{(j)}$. We can also keep the same thresholds for evaluating the level of consistency between-LVDT means.

The final outcome of the Mandel's h statistic for the Bayesian framework is depicted in Figure 4.13. The differences with respect to the ISO 5725-2 results are negligible, being of the same order of magnitude of the least significant digit (Figure 4.14) and the main differences are not for the lower consistency results. Therefore, in this case we have a negligible "shrinkage-to-the-mean effect with respect to the Mandel's h statistic in its natural frequentist framework. Further analysis and studies have to be performed before giving a universal validity to this conclusion.

Arguing along these lines, we can extend the same assumptions to the Grubbs test and, as for Mandel's h statistic we can try and replace the sample mean $\bar{x}_i^{(j)}$ of each LVDT's group of n-replicated measurement with the expected value of the posterior distributions for the specific LVDTs $x_{\cdot i}^{(j)}$. As many other statistical tests we use to distinguish between Single Grubbs test for the detection of an extreme laboratory mean as an outlier and the Double Grubbs test for the detection of two extreme laboratory means as outliers. The former uses the maximum or the minimum of Mandel's h statistic as test statistic, the latter uses the sum of squared deviations of the laboratory (LVDT in our case) means from their average without the two extreme means divided by that with the two extreme means as test statistic [43]. The null hypothesis of both the

Single and the Double Grubbs tests is that the $X_i^{(j)}$, with $i = 1, \dots, t$, are independently and identically distributed according to a normal distribution. This is also the null hypothesis of any goodness-of-fit-test for normality or any specific test for normality, e.g. the Shapiro–Wilk test. However, whereas the tests for normality are powerful against the alternative hypothesis that the $X_i^{(j)}$ are i.i.d. according to a non-normal distribution, the Grubbs tests as outlier test are powerful against the alternative hypothesis that all $X_i^{(j)}$ are i.i.d. normally distributed except the random variables modeling the extreme values of the sample, and that these follow a distribution with larger (smaller) mean [43].

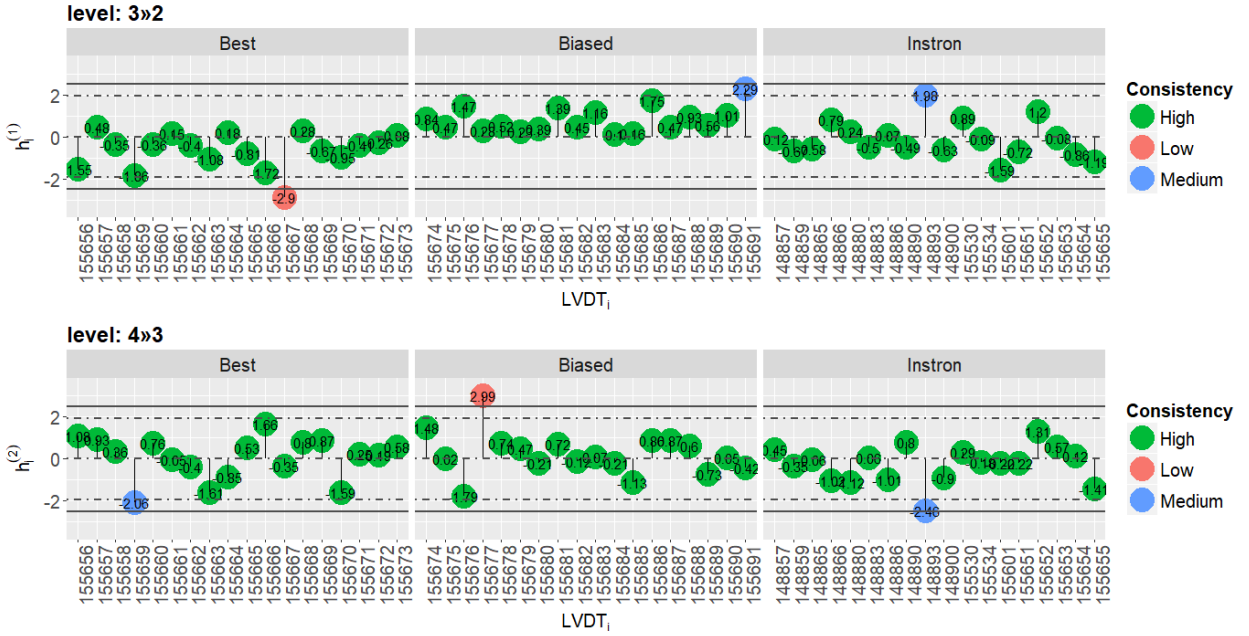

Figure 4.13: Mandel's between-laboratory consistency statistic for the Bayesian framework, grouped by LVDT serial number, for the simulated proficiency testing of displacement measurements in VST apparatus

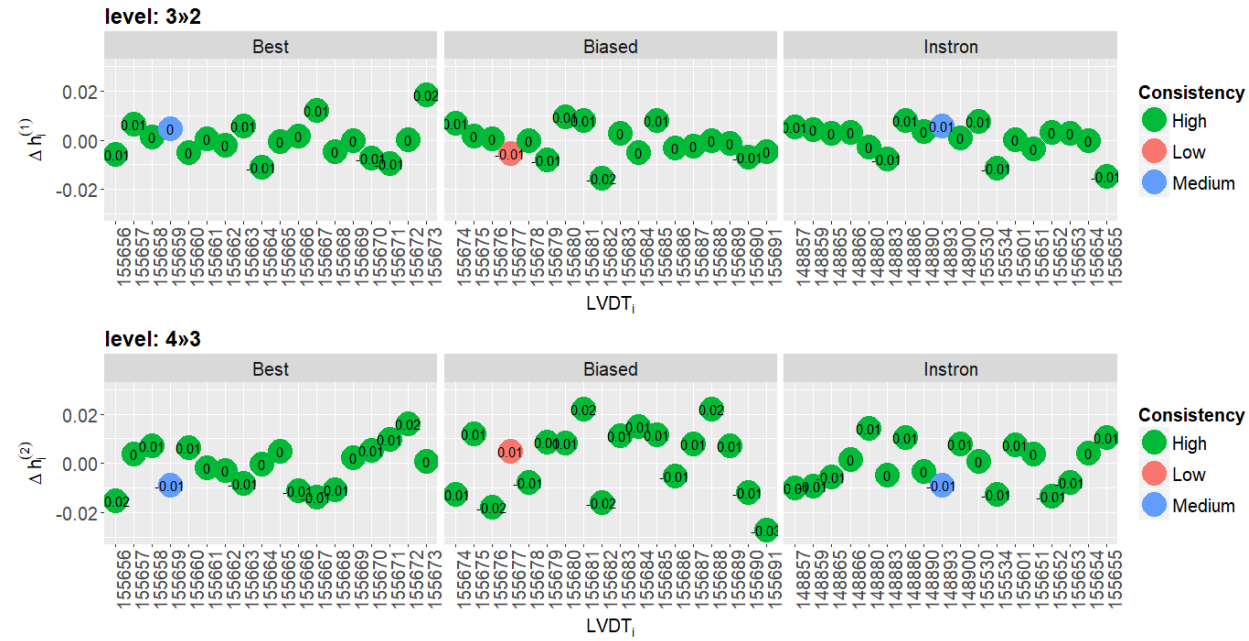

Figure 4.14: Differences $\Delta h_i^{(j)}$ between Mandel's statistic in the Bayesian framework against ones according to ISO 5725-2

It can be proved that the critical value, that is the $(1-\alpha)$-quantile $g_{t;1-\alpha}$ for the Single Grubbs test is approximated by the $(1-\alpha)$-quantile $h_{t;1-\alpha/t}$, by application of Bonferroni's inequality [43]. The ISO 5725-2 recommends:

- if the Grubbs statistic is less than or equal to its 5 % critical value, the specific laboratory is accepted as correct;
- if the Grubbs statistic is more than its 5 % critical value and less than or equal to its 1 % critical value, the specific laboratory is called a straggler;
- if the Grubbs statistic is more than its 1 % critical value, the specific laboratory is called a statistic outlier.

Therefore, in the present proficiency test, the critical values for stragglers are $\pm g_{t;1-0.05} = \pm 2.987$, and the critical values for outliers are $\pm g_{t;1-0.01} = \pm 3.369$. From results in Figure 4.9 and Figure 4.13, all LVDT data for the first proficiency test item ("jump" 3»2) can be accepted as correct, whereas for the second proficiency test item ("jump" 4»3) the LVDT having serial number 155677 is in the thin line between correct and straggler. More precisely, if we look at the results achieved according to ISO 5725-2 it can be accepted as correct ($2.98 < 2.987$), whereas in the Bayesian framework it ought to be a straggler ($2.99 > 2.987$). Looking at the same result from the p-value perspective, for the LVDT having serial number 155677 you get 5.05 % with ISO 5725-2 and 4.96 % with the Bayesian framework. This means that, even though the differences in Figure 4.14 seem to be negligible, the Single Grubbs test in the Bayesian framework tends to be less conservative, or in other words it is easier to reject the null hypothesis, at least in this present case.

Even if the present proficiency test needs not being checked against the Double Grubbs test, working with the expected values of the posterior probability distributions for the specific laboratory parameters also the Double Grubbs test can be performed in the Bayesian framework. Leaving to the curious reader the task of analyzing in detail the Double Grubbs statistic [43], [44], we can check if the lowest and highest value are two outliers on opposite tails in both ISO 5725-2 and Bayesian scenarios. In this respect, calculating the p-value according to the Double Grubbs statistic, we achieve for the first proficiency test item ("jump" 3»2) 13,67 % according to ISO 5725-2 against 14,54 % of the Bayesian framework, whereas for the second proficiency test item ("jump" 4»3) 5,74 % according to ISO 5725-2 against 5,44 % of the Bayesian framework. Again the results are not exactly the same but they are not so far between the two approaches, ISO 5725-2 and Bayesian.

Finally, about the usability of the Grubbs test on the expected value of the posterior probability distributions for the specific laboratory parameter we achieved good results, that means they are close to ones achieved on sample data. Nevertheless, further analysis and studies have to be performed before considering this conclusion an overall rule.

Lastly, we can consider the Mandel's k statistic and the Cochran test. Both are based on the assumption that $X_{ki}^{(j)}$, with $k = 1, \dots, n$ and $i = 1, \dots, t$, are independently and identically distributed according to the normal distribution $Normal(\mu_i, \sigma)$ and, to a certain degree, both look for consistency or detection of an extremely large within-laboratory standard deviation with respect to its mean value. So, the structure of both these statistics lays on the within-laboratory standard deviation or variance [43]. Nevertheless, we cannot apply them to the standard deviation of the posterior probability distributions for the specific laboratory parameters due to the borrowing of precision effect. As a matter of fact, the scale parameters of the bell-shaped curves for each LVDT in Figure 4.10 or the slopes of the Q-Q plots in Figure 4.11 are all very close to their mean value. Moreover, the critical value for both statistics is also based on the number n of replicated measurements within each laboratory. This number for the posterior distributions is a "big value" (size of thousands of simulated sample) due to the MCMC solving strategy and it pushes close to each other the critical values for stragglers and outliers or for medium and low consistency.

Anyway, the problem of testing outlying observations regards the data themselves, so it can leave out of consideration the Bayesian framework. In other words, before any metrology evaluation we can apply the graphical consistency techniques and the outlier detection tests in their original formulation [43], [44], [45]. Then, on the "checked" dataset we can apply the Bayesian statistics for proficiency testing, as in section 4.3, or for accuracy evaluation, as in section 3.2.

4.5 HIERARCHICAL BAYESIAN MODEL FOR MANUFACTURERS RATING

There is, of course, no reason to stop us considering models with more levels. The appropriate number of levels is governed by the structure of the data under consideration. So, if we want to leverage the proficiency testing scheme not only to determine the performance of participants but also to rate the test equipments made by different manufacturers, we can add a hierarchical level to the probability structure in order to consider this additional parameter.

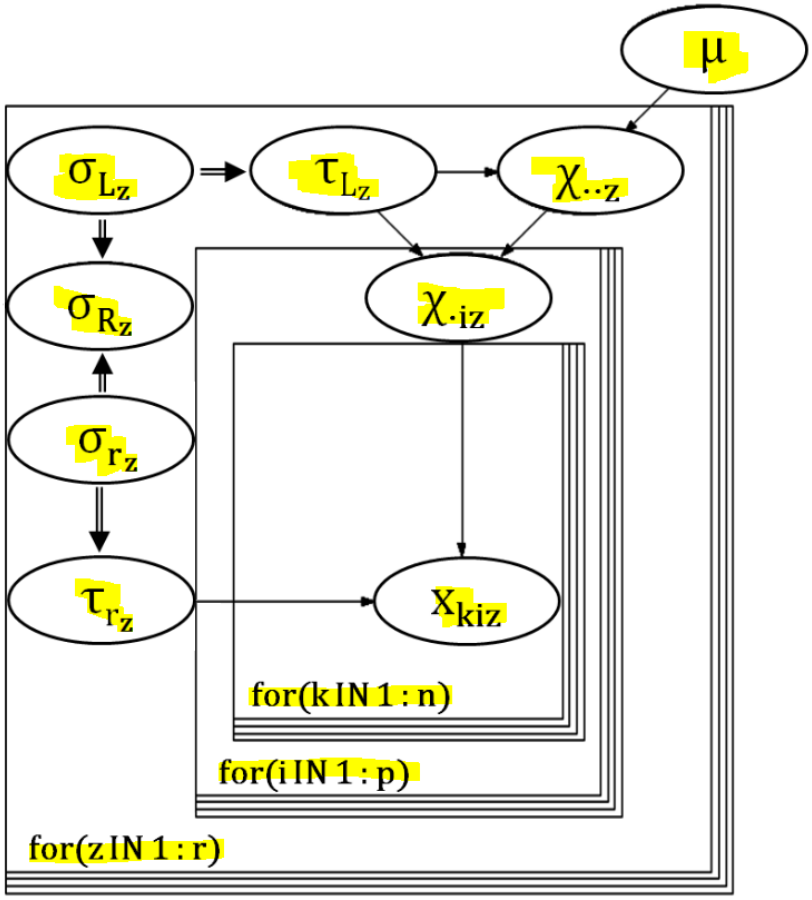

Figure 4.15: Hierarchical Bayesian probability structure for Proficiency testing scheme and manufacturers rating

The proposed Hierarchical Bayesian model for fulfilling this purpose is shown in Figure 4.15 and it can be summarized by the following conditional probability structure based on the exchangeability assumptions:

$$
\begin{aligned}
x_{kiz}^{(j)} | \chi_{\cdot iz}^{(j)}, \sigma_{r_z}^{(j)} &\sim i.i.d.\ \ Normal\left(\chi_{\cdot iz}^{(j)}, \sigma_{r_z}^{(j)}\right) \\
\chi_{\cdot iz}^{(j)} | \chi_{\cdot\cdot z}^{(j)}, \sigma_{L_z}^{(j)} &\sim i.i.d.\ \ Normal\left(\chi_{\cdot\cdot z}^{(j)}, \sigma_{L_z}^{(j)}\right) \\
\chi_{\cdot\cdot z}^{(j)} | \mu^{(j)}, \sigma_{L_z}^{(j)} &\sim i.i.d.\ \ Normal\left(\mu^{(j)}, \sigma_{L_z}^{(j)}\right) \\
\sigma_{L_z}^{(j)} &\sim Uniform(0{,}0.010) \quad \forall j, z \\
\sigma_{r_z}^{(j)} &\sim Uniform(0{,}0.006) \quad \forall j, z \\
\mu^{(j)} &\sim Norm(1{,}0.005) \quad \forall j
\end{aligned}
\qquad 4.16
$$

where $k = 1, \dots, n$ is the subscript for the replicated measurements of each i-participant laboratory, $i = 1, \dots, t$ is the subscript for the specific laboratory, $z = 1, \dots, r$ is the subscript for the manufacturer (one between Instron, Best and Biased) and $j = 1{,}2$ is the superscript for the specific proficiency test item.

The proposed probability structure, shown in Figure 4.15 and defined by equations 4.16, pursues the following founding idea. The accuracy description is split between the two top hierarchical levels. The first level expresses the characteristic properly related to the method

accuracy, that we supposed to be just the overall mean, whereas the second level describes mean, repeatability and inter-laboratory variability for a specific manufacturer. So, we lose the capability of capturing the overall method accuracy in order to highlight the differences between the equipments made by different manufacturers. Besides, being the dataset assumed to represent an inter-laboratory experiment, is it correct to talk about reproducibility instead of homogeneity. Of course, we are aware that this is misleading if we just consider the true experimental conditions for the Instron's data subset.

The hierarchical Bayesian model 4.16 is written following the OpenBUGS code as

```
model {
mu ~ dnorm (0, 40000)
    for (z in 1:r) {
        x.z[z] ~ dnorm (mu, tau.Lz[z])
        tau.Lz[z] <- pow(sigma.Lz[z], -2)
        sigma.Lz[z] ~ dunif (0, 0.010)
        tau.rz[z] <- pow(sigma.rz[z], -2)
        sigma.rz[z] ~ dunif (0, 0.006)
        sigma.Rz[z] <- sqrt(pow(sigma.rz[z], 2) + pow(sigma.Lz[z], 2))
                for (i in 1:p) {
                        for (k in 1:n){
                                xkiz[k,i,z] ~ dnorm (xiz[i,z], tau.rz[z])
                                        }
                                xiz[i,z] ~ dnorm (x.z[z], tau.Lz[z])
                                }
        }
    }.
```

4.17

To use the "R2OpenBUGS" package, the simulated dataset has to be arranged, according to the OpenBUGS rules, as a "list" R object in 4.13, having now $n \times p \times r$ dimensions:

```
$p
[1] 18

$r
[1] 3

$n
[1] 4

$xkiz
, , 1

      [,1]  [,2]  [,3]  [,4]  [,5]  …  [,14] [,15] [,16] [,17] [,18]
[1,] 1.000 1.003 0.998 1.005 1.004  …  0.999 1.009 1.001 0.995 0.995
[2,] 1.003 0.996 0.999 1.005 1.003  …  0.998 1.009 1.001 0.995 0.995
[3,] 1.002 0.999 0.999 1.009 1.004  …  0.999 1.009 1.003 1.002 0.996
[4,] 1.002 0.997 1.001 1.008 1.004  …  0.998 1.009 1.003 0.999 0.998

, , 2

      [,1]  [,2]  [,3]  [,4]  [,5]  …  [,14] [,15] [,16] [,17] [,18]
[1,] 0.996 1.007 1.000 0.994 1.003  …  0.995 1.002 0.998 0.998 1.002
[2,] 0.995 1.007 1.001 0.991 1.005  …  1.001 0.999 1.001 1.001 1.003
[3,] 0.993 1.003 0.999 0.994 0.998  …  1.000 0.993 1.000 0.999 1.001
[4,] 0.992 1.003 1.002 0.990 0.996  …  0.999 0.995 1.002 1.006 1.005

, , 3

      [,1]  [,2]  [,3]  [,4]  [,5]  …  [,14] [,15] [,16] [,17] [,18]
[1,] 1.004 1.003 1.014 1.002 1.007  …  1.000 1.002 1.006 1.007 1.015
[2,] 1.009 1.011 1.014 1.006 1.001  …  1.011 1.011 1.004 1.006 1.017
[3,] 1.007 1.005 1.010 1.004 1.008  …  1.003 1.005 1.009 1.007 1.013
[4,] 1.008 1.001 1.004 1.004 1.005  …  1.006 1.012 1.003 1.012 1.015
```

4.18

As in the previous cases, the starting points are randomly selected from tighter proper prior distributions, where tighter means having a reduced scale parameter. Finally, the Bayesian numerical computation is performed using the “bugs” function of the “R2OpenBUGS” package.

The results are shown in Figure 4.16 in terms of mean value (points) and standard deviation (error bars) of the posterior distributions for the specific laboratory parameters $\chi_{\cdot i}^{(j)}$. The outcome is compared with sample values: the pink points denote the expected values of the specific LVDTs posterior distributions and the associated error bars span two standard deviations of their posterior distributions, whereas the light blue triangles and associated error bars are for the sample values, the same shown in Figure 4.3. Again we see the typical behavior of the Bayesian statistics, that are the borrowing of precision between the specific laboratory parameters and the shrinkage to the mean.

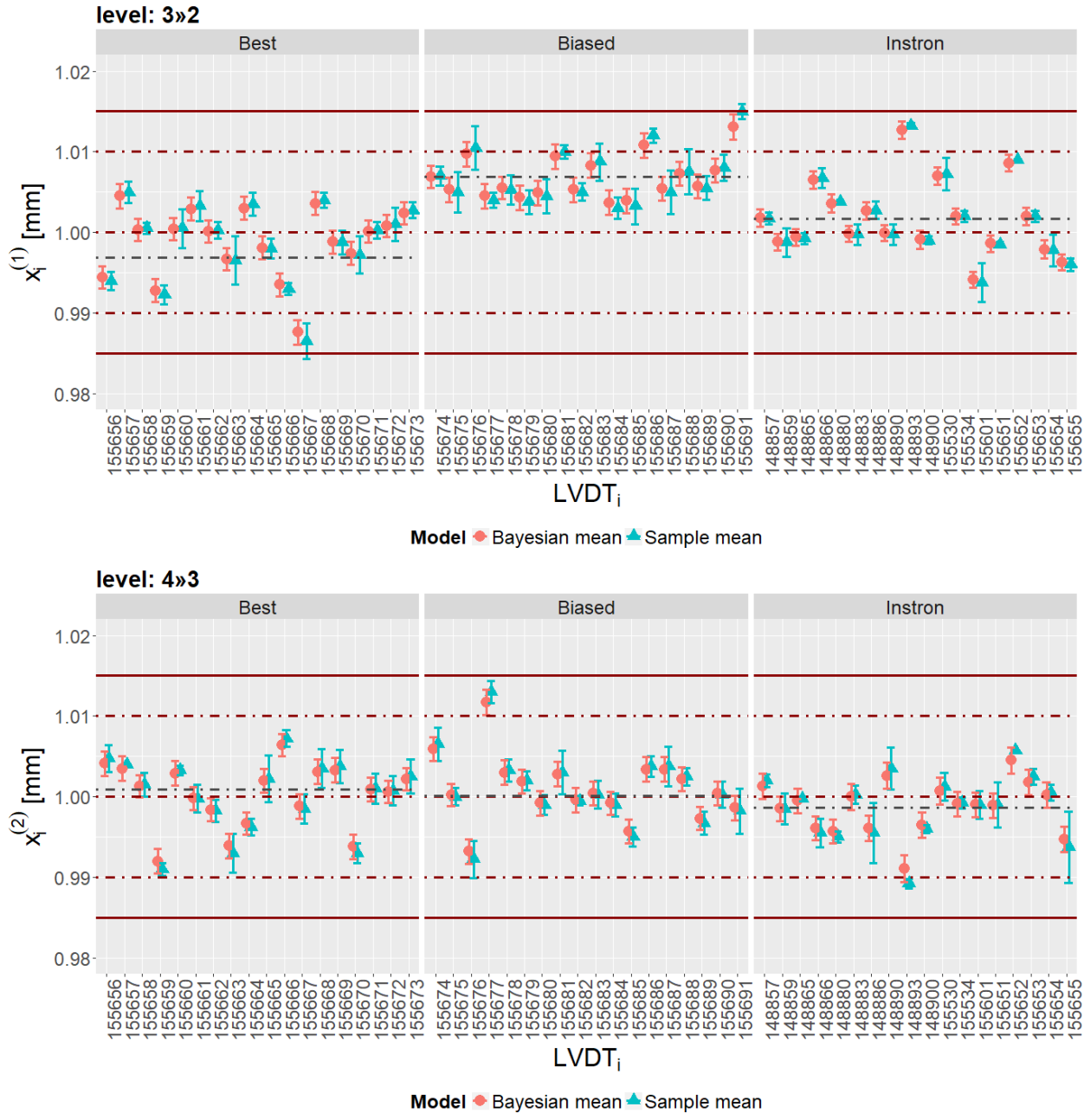

Figure 4.16: Bayesian posterior compared to sample mean value ± lab standard deviation for each proficiency test item

The shrinkage to the mean deserves attention. The center of gravity for each manufacturer group is not the general mean but the specific manufacturer mean value (grey dot-dashed lines). This effect is more evident when looking at the “Biased” manufacturer results for the first test item (“jump 3»2”). For this reason, we observe slight differences in the expected values of the posterior

distribution for the specific LVDT parameters with respect to those shown in Figure 4.6 and Figure 4.7, that were achieved with the Bayesian model 4.11.

We can calculate the z-scores on the specific LVDT results of the hierarchical Bayesian model 4.16 arranging them in the Youden plot of Figure 4.17. Even if overall there is a good level of agreement of the results in Figure 4.17 with respect to those in Figure 4.8, we can see that the left side of the cloud of the green points is a bit shifted to the right with respect to what is shown in Figure 4.8 for the results of the hierarchical Bayesian model 4.11, and this is a consequence of the shrinkage to Biased manufacturer mean for the first test item. The effect is more evident for the z-scores of the LVDT having serial number 155677 in the Figure 4.17 with respect to those in Figure 4.8.

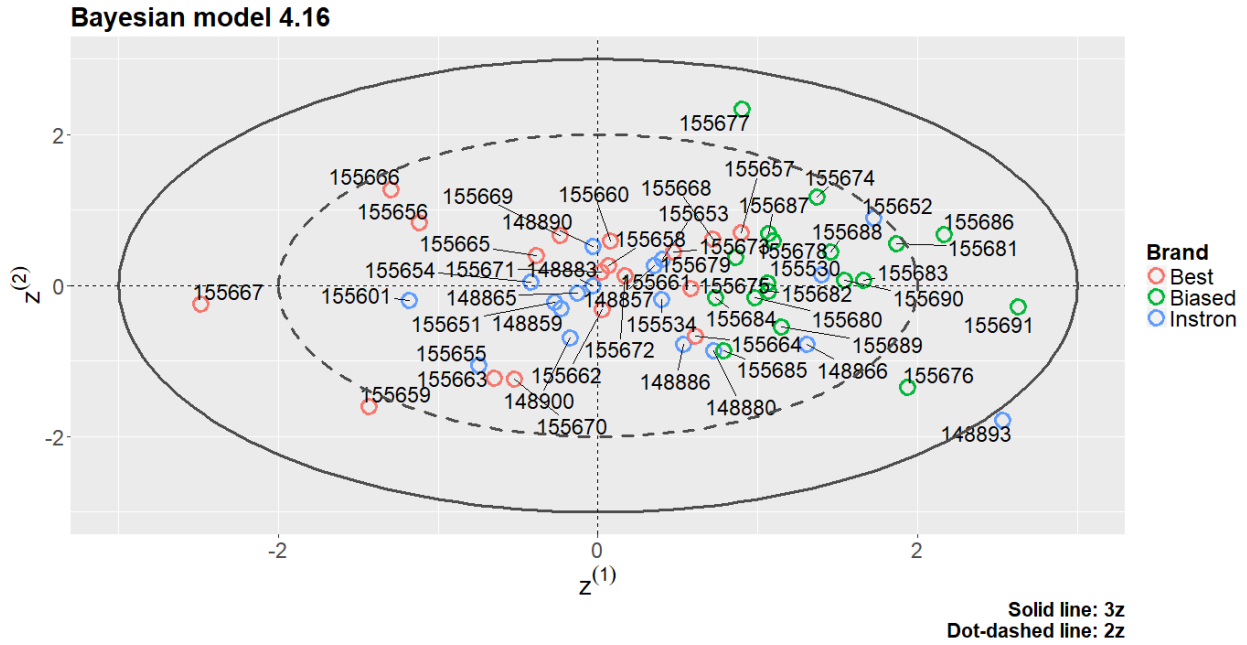

Figure 4.17: Youden Plot for the z-score of the jump 3»2, that is $z^{(1)}$, against the ones of the jump 4»3, that is $z^{(2)}$

Now, we are wondering if the model is able to catch the same accuracy description within each manufacturer group as the stand alone model. For this purpose, we compare the point estimators obtained using model 3.15 to those achieved using model 4.16. The comparison is summarized in Table 4-1 where the superscript identifies the first or the second test item. The subscript is always 1 because we are looking just inside the the manufacturer group, Instron.

	$x_1^{(1)}$/mm	$\sigma_{r_1}^{(1)}$/mm	$\sigma_{L_1}^{(1)}$/mm	$\sigma_{R_1}^{(1)}$/mm	$x_1^{(2)}$/mm	$\sigma_{r_1}^{(2)}$/mm	$\sigma_{L_1}^{(2)}$/mm	$\sigma_{R_1}^{(2)}$/mm
Model 4.16	1.002	0.0021	0.005	0.0054	0.999	0.0034	0.0037	0.0051
Model 3.15	1.002	0.0021	0.0052	0.0056	0.999	0.0034	0.0039	0.0052

Table 4-1: Point estimator comparison between the outcome of the model 3.15 against 4.16 for general manufacturer mean and repeatability, inter-LVDTs and reproducibility standard deviations

Table 4-1 shows a good agreement even if the results were achieved by two different paths and enlarging the dataset as input of the model 4.16 for the proficiency testing scheme. We have slight differences only on the point estimators of the inter-laboratory (LVDT) variability. The goodness of this results arises by having a balanced experimental plan even for the proficiency testing dataset and enough laboratories for each manufacturer. We refer to future studies to answer the question on the effect of unbalanced experimental plan or smaller-sized datasets,.

In connection with the accuracy description, we can analyze the posterior distributions of the parameters connected to its definition. We first examine in Figure 4.18 the overall mean for the LVDT displacement measurement in the equipments performing the VST method. The posterior distributions for the two test items give an idea of the trueness for the measurand. Considering that, in this case, we have a reference accepted value that can be modeled as a random variable described by a Dirac delta probability function which is zero everywhere except at 1 mm, then,

according to 2.25, the trueness can be described by the same bell shapes of Figure 4.18 shifted by -1 mm. We also remind that the assumption for the reference accepted value is drawn up as a consequence of the uncertainty budget analysis in Table 2-13.

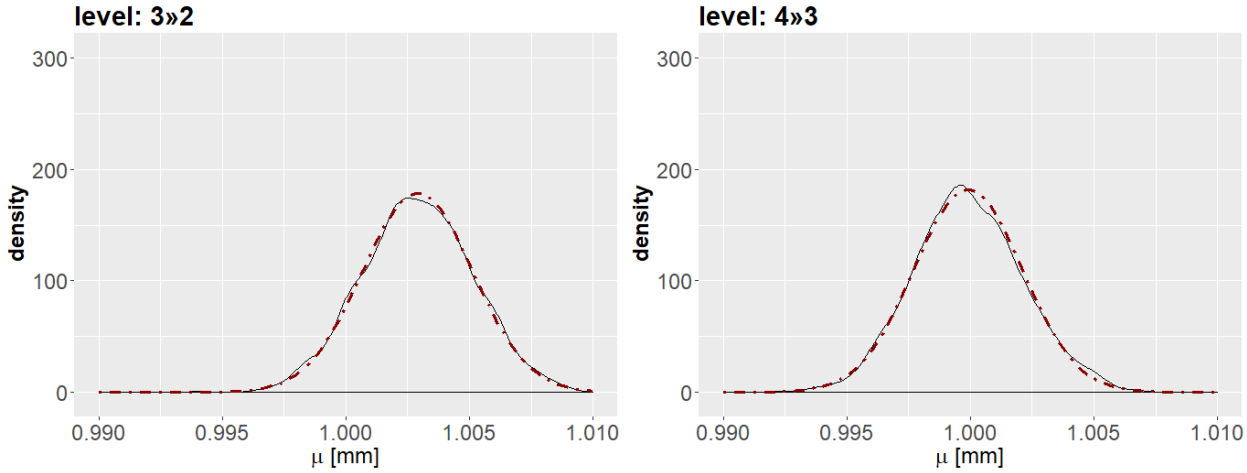

Figure 4.18: posterior probability distribution of the overall mean for the VST method as outcome of the simulated proficiency testing experiment

Likewise, we can consider the posterior distributions for the specific manufacturer mean parameters $\chi^{j}_{\cdot\cdot z}$. Figure 4.19 depicts their outcome after Bayesian updating according to model 4.16. Again we can think about them as the expression of the measurement trueness because of the assumption for the reference accepted value as the Dirac delta function which is zero everywhere except at 1 mm.

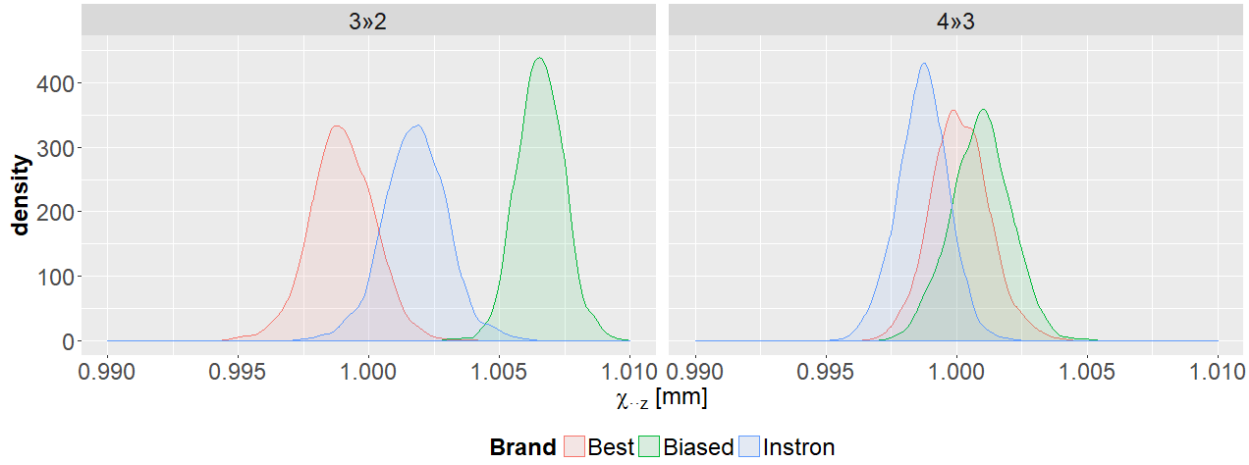

Figure 4.19: posterior probability distribution of the specific manufacturer mean for the VST method as outcome of the simulated proficiency testing experiment

In both Figure 4.18 and Figure 4.19 there is a good fit with respect to the normality assumption. The posterior normal distribution of the overall mean is wider than those for the specific manufacturer means and encompasses them. Lastly, there is an evident bias on the overall mean, due to the Biased subset of data.

We can then analyze the posterior probability density functions of the variability parameters, i.e. repeatability, inter-laboratory and reproducibility standard deviations shown in Figure 4.20. As we could expect, we have a good fit with the associated gamma theoretical distribution just for repeatability and inter-laboratory posterior distributions. As a matter of fact, in Figure 4.20 there is a good overlapping between the solid line and the dot-dashed line, the former depicts the numerical solution of the Bayesian hierarchical model whereas the latter is related to the theoretical distribution. Instead, the gamma assumption is not so good for the reproducibility parameter, being the combination of two gammas having different scale parameters as stated in the equation 2.7.

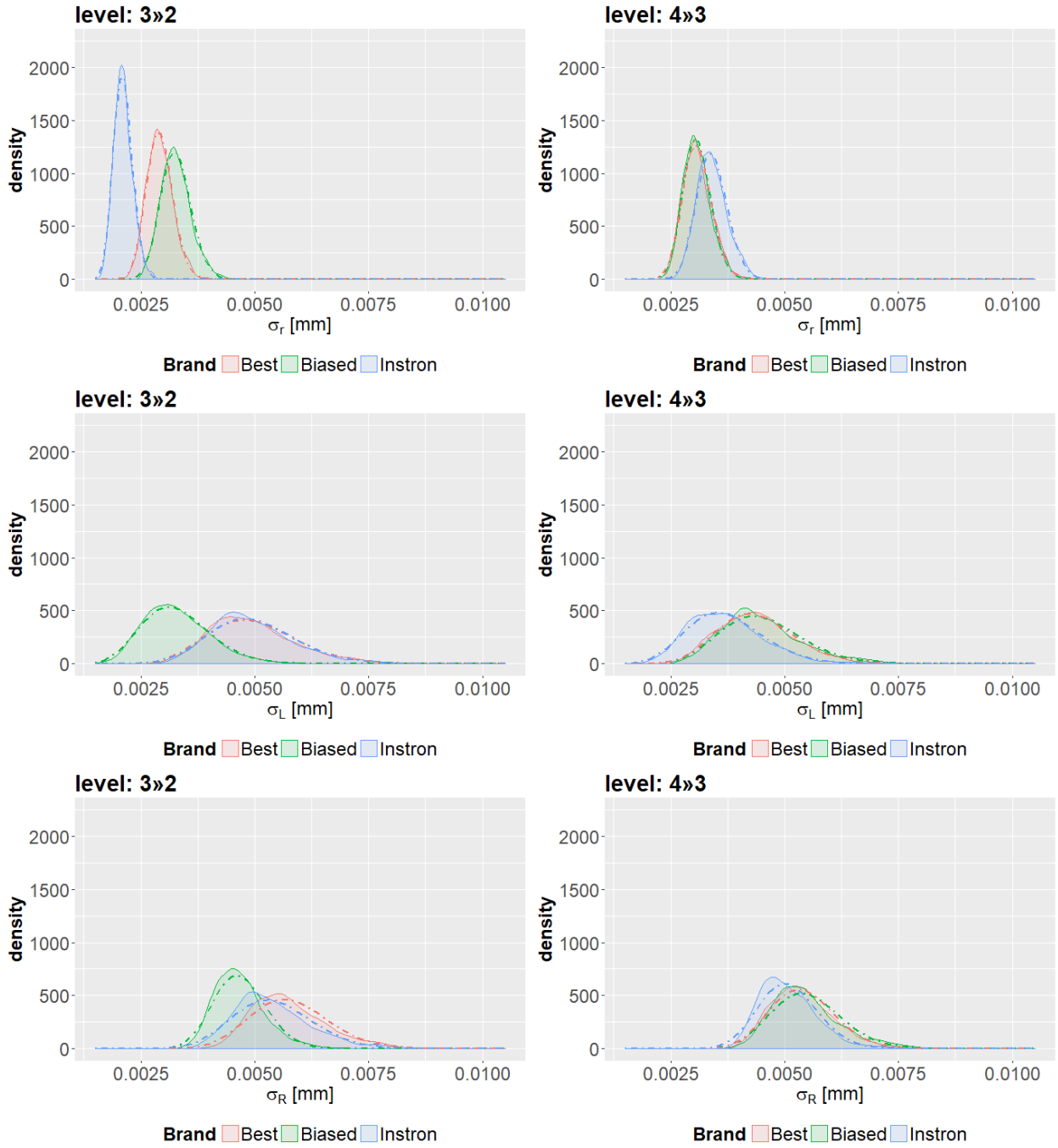

Figure 4.20: Posterior probability distributions of precision parameters under repeatability, inter-laboratory variability and reproducibility condition grouped by manufacturers (different colors) and by test items (different columns)

Figure 4.18, Figure 4.19 and Figure 4.20 highlight the great advantage of the Hierarchical Bayesian models: in spite of their greater calculus complexity, they yield a full description in terms of probability distributions for each of the concepts defining accuracy. We have not just confidence intervals for the point estimators of the accuracy parameters, achieved as linear combinations of the residuals' mean square error and propagating the normality assumption for residuals, but in the presented Hierarchical Bayesian models we also have the full probability distributions, being the parameters random variables, each with its own density. So, we see the bias as a normal distribution more or less centered in zero, the repeatability as a gamma function and the reproducibility having a density not so far from a gamma function.

Using equations 3.16, from the posterior probability distributions of the accuracy parameters we can draw up the credibility intervals for each of them. These results are summarized in Table 4-2. Once again, looking at the results for Instron product line, we have a good level of agreement with the results of the accuracy experiment in Table 3-16. On the basis of this information it can be

concluded that, performing the hierarchical Bayesian model 4.16 for a balanced experimental proficiency dataset with this cardinality and without outliers on the dataset, we do not introduce any distortion on the accuracy statistics for each manufacturers product line, that are, bias and uncertainties in Table 3-16.

	Level: 3»2				Level: 4»3			
	$\hat{\delta}^{(1)}_{z_{lower}}$/mm	$\hat{\delta}^{(1)}_{z_{upper}}$/mm	$U^{(1)}_{r_z}$/mm	$U^{(1)}_{R_z}$/mm	$\hat{\delta}^{(2)}_{z_{lower}}$/mm	$\hat{\delta}^{(2)}_{z_{upper}}$/mm	$U^{(2)}_{r_z}$/mm	$U^{(2)}_{R_z}$/mm
Instron	-0.001	0.004	0.0025	0.0070	-0.003	0.001	0.0040	0.0062
Best	-0.003	0.001	0.0034	0.0074	-0.002	0.002	0.0036	0.0067
Biased	0.005	0.008	0.0039	0.0057	-0.001	0.003	0.0036	0.0070

Table 4-2: Bias and uncertainties under repeatability and reproducibility conditions for the LVDT proficiency testing scheme based on the Hierarchical Bayesian having the conditional probability structure of the equation 4.16

After the model validation, confident that we did not introduced distortion, finally we can introduce the rating metrics in order to judge which manufacturers' product line is performing better regarding the ISO 306 standard requirement. The aim of the rating is to evaluate the ability of each manufacturers' measuring system to produce output within specification limits of the standard. The proposed idea for the rating statistics is borrowed from the Statistical Process Control (SPC) methods. So, it was thought to adopt something similar to process capability C_p and C_{pk} statistics. These indices measure how much "natural variation" a process experiences relative to its specification limits and allows different processes to be compared with respect to how well an organization controls them. Both these statistics assume a two-sided specification, if μ and σ are the mean and standard deviation, respectively, of the normal data for the process under control and USL and LSL are the upper and lower specification limits, respectively, then the population capability indices are defined as follows:

$$\hat{C}_p = \frac{USL - LSL}{6\hat{\sigma}}$$
$$\hat{C}_{pk} = min\left\{\frac{\hat{\mu} - LSL}{3\hat{\sigma}}, \frac{USL - \hat{\mu}}{3\hat{\sigma}}\right\} \qquad 4.19$$

Working on the VST example, the specification limits can be easily associated with the ISO 306 requirement for accuracy, so:

$$ULS = -LSL = 0.01\ mm \qquad 4.20$$

The process mean value can be thought as the specific manufacturer product line bias, in formula:

$$\hat{\mu}_z^{(j)} \coloneqq \left[x^{(j)}_{\cdot\cdot z} : \Phi_{\chi^{(j)}_{\cdot\cdot z}}\left(x^{(j)}_{\cdot\cdot z}\right) = .5\right] - 1 \qquad 4.21$$

where 1 is measured in millimeters and it the reference accepted value for the deflection measurements of the VST accuracy experiment. We did not use the generic letter, as stated in 2.25, in order to avoid conflict between symbols, being μ also the choice for indicating the overall mean parameter of the Bayesian model having the probability structure in 4.16.

Lastly, to estimate the half width of the "natural process variability, instead of three standard deviations, we use the quantile function of the posterior probability distribution of the reproducibility parameter for each z-specific manufacturer, that is, the inverse of its posterior probability cumulate distribution function, corresponding to a probability equal to 99,73 %, that represents the probability associated to a standard normal random variable taking on a value less than or equal to 3. In formula we have:

$$3\hat{\sigma}_z^{(j)} \coloneqq s^{(j)}_{R_z} : \Phi_{\sigma^{(j)}_{R_z}}\left(s^{(j)}_{R_z}\right) = .9973 \qquad 4.22$$

Therefore, the manufacturer rating equations are:

$$\widehat{R_{p_z}}^{(j)} = \frac{USL - LSL}{6\hat{\sigma}_z^{(j)}}$$
$$\widehat{R_{pk_z}}^{(j)} = min\left\{\frac{\hat{\mu}_z^{(j)} - LSL}{3\hat{\sigma}_z^{(j)}}, \frac{USL - \hat{\mu}_z^{(j)}}{3\hat{\sigma}_z^{(j)}}\right\} \qquad 4.23$$

where $\widehat{R_{p_z}}^{(j)}$ describes, under the assumption of zero bias, the capability of the product line of the z-specific manufacture for the j-level of the reference accepted value factor to provide measurements within the specific limits range, whereas $\widehat{R_{pk_z}}^{(j)}$ describes the capability of the product line of the z-specific manufacture for the j-level of the reference accepted value factor to provide measurements centred between the specification limits.

Therefore, applying the rating metrics in 4.23 to the outcome of the Hierarchical Bayesian model 4.16 we achieve the final result that is summarized in Table 4-3.

	$\hat{R}_{p_z}^{(1)}$	$\hat{R}_{pk_z}^{(1)}$	$\hat{R}_{p_z}^{(2)}$	$\hat{R}_{pk_z}^{(2)}$
Instron	1.149	0.945	1.255	1.093
Best	1.087	0.977	1.197	1.180
Biased	1.428	0.490	1.197	1.086

Table 4-3: Manufacturers' product lines ratings

Higher ratings correspond to better manufacturer performance. The optimum value for $\widehat{R_{p_z}}^{(j)}$ should be 2. Anyway, above 1 the performance is satisfactory. More complex is the judgment about $\widehat{R_{pk_z}}^{(j)}$ statistic. As a matter of fact, it depends also on the $\widehat{R_{p_z}}^{(j)}$ index so that, if $\widehat{R_{p_z}}^{(j)} \gg 1$, then $\widehat{R_{pk_z}}^{(j)} < 1$ could be acceptable. That is exactly what happened for the "Biased" manufacturer, for whom looking at the Youden plot outcome in Figure 4.17 we don't have any low consistency results, even if its $\widehat{R_{pk_3}}^{(1)}$ is less than 1.

Hence, we did not lose consistency in accuracy evaluation for the each specific product line, even if the general problem has been modelled as fully nested experimental design with four different hierarchical levels: the upper level is for test method, the second is for the manufacturer product lines, the third is for the specific testing equipments and the bottom level is for residuals.

5 CONCLUSIONS

At the practical level, hierarchical models proved to be flexible tools for combining information from both cognitive sources, that are measurements and prior knowledge. The flexibility guarantees a complete accuracy description in terms of probability density function for trueness, evaluated as bias from a reference accepted value, and for precision under repeatability and homogeneity or reproducibility conditions. So, instead of the point estimators described in ISO 5725-2, that yield single-valued results, the great advantage of the hierarchical Bayesian accuracy model, according to the proposed probability structure of equation 3.15 or, at a higher level, of equation 3.19, is the possibility to have a numerical approximation of the whole marginal probability distribution for each accuracy parameter. This leads to a better fit with the metrology attitude of the last decade. Indeed, the recent GUM-S1 appears to move the metrology guidelines towards the direction of working with probability distributions encoding the state of knowledge about quantities. This attitude is fully satisfied in both accuracy experiments considered in this research. As we apply the hierarchical Bayesian accuracy model, the final outcome is not just the expected value but the whole marginal posterior distribution for each of the parameter describing the accuracy (see Figure 3.4 for Sulfur content experiment or Figure 3.20 for LVDT displacement experiment).

Conceptually, this allows to overcome the criticism often moved at the GUM, that of being based on a mixture of frequentist and Bayesian thinking. In particular, the GUM Type A uncertainty evaluations are frequentist, whereas the Type B evaluations, using state-of-knowledge distributions, have a Bayesian attitude. In a fully Bayesian framework the inference is based on the state-of-knowledge, expressed in terms of prior distributions that are upgraded as functions of data evidence (observed measurements).

The most important result from the model validation viewpoint is the consistency of the Bayesian outcome, expressed as expected values for accuracy parameters, with the bias and expanded precision estimates according to ISO 5725-2. In the case of the LVDT displacement experiment, the difference between ISO 5725-2 outcome (Table 2-11) and the Bayesian one (Table 3-14) is smaller than the instrument resolution. Therefore, there is no evidence for asserting that the bias and precision point estimates are different, be they obtained with ISO 5725-2 or the proposed Bayesian statistics. This statement is no more completely true as we move from point estimation to interval estimation, that is what manufacturers need to characterise their whole product lines of test equipments. In fact, comparing the ISO 5725-2 interval estimates (Table 2-14) to the Bayesian (Table 3-14), you may gather that the ISO 5725-2 is not able to properly separate random and systematic errors. As explained in section 2.1, the statistics describing trueness and precision according ISO 5725-2 descend from the ANOVA general linear model, so that the variance decomposition follows the ANOVA rules. This means that we are able to distinguish between the within-groups and among-groups variances, the former expressing repeatability and the latter expressing inter-laboratory variability. However, we are unable to evaluate the fraction of variability due to the systematic effects. Undoubtedly, looking at the result reported in Table 2-14, the product line trueness of the displacement measurement as bias interval estimate according to ANOVA model does not make much sense. Three intervals over four do not encompass the zero. This result suggests that the product line displacement measurements are systematically biased. By contrast, the proposed Bayesian model does not come to the same conclusion. Looking at the outcome in Table 3-14, there is no enough evidence to infer that the displacement measurements of the product line population are systematically biased. We believe in the Bayesian result because it provides a better variance decomposition, being the proposed Bayesian model also able to describe the systematic variability in terms of probability distribution for the general mean. So, the Bayesian model seems definitely better to describe the accuracy of a whole product line, in a way consistent with the concepts of trueness and precision, as they are stated in the International Vocabulary of Metrology. Looking at the final result, the customer realizes that, buying one Instron VST equipment for testing samples 3 mm thick, the displacement measurement can be affect by a systematic error in the range $[-1 + 4]$ µm and by a random error up to 7.3 µm under homogeneity conditions, both having a credibility level of 95 %.

Moreover, we can easily leverage the exchangeability assumption for the reference accepted value, thought in order to analyze the metrological performance for a discrete subset of sample thicknesses allowed by the ISO 306, and adding another hierarchical level it is possible to have a

complete overview of the accuracy associated with the displacement measurements in the Instron VST test equipment. The overall accuracy outcome is summarized in Table 3-19 as interval estimates for specific credibility levels or in Table 3-20 as credibility levels given some specific interval estimates.

The proposed hierarchical models for accuracy evaluation have been already tested under two experimental plans with different sample sizes and in one case including missing values, under different assumptions for priors, under different choices for MCMC numerical conditions. Of course we are far from a full sensitivity analysis and further researches are welcome. The main pending questions are how much small the experimental sample size can be without affecting the accuracy outcome and if the sensitivity to prior assumptions depends on sample size.

The greater disadvantage is the greater calculus complexity with respect to ISO 5725-2, especially because a possible solution in closed form was not analyzed. Nowadays, even where no closed-form algebraic formulas are available, Bayesian statistics of interest can be calculated using a computer simulation technique, popularly known as Markov Chain Monte Carlo (MCMC) [16]. Even complicated hierarchical Bayesian models as those used in this study can be solved easily thanks to the IT progress. However, it should be pointed out that the Bayesian framework needs a deeper level of statistical knowledge, which might be a deterrent to their spread in metrology.

In order to have a complete overview of the Bayesian framework in metrology, the present research proofs how it is possible to exploit the hierarchical models flexibility for proficiency testing schemes. The benchmarks are the z-scores statistics according to ISO 13528. Overall, also in this case there is a good agreement between the Bayesian outcome and its benchmark. Nevertheless, we need to pay attention to the shrinkage-to-the-mean effect of the hierarchical Bayesian models that can lead to misclassifications with respect to the benchmark. Also this topic deserves to be further analyzed focusing on possible choices for comparison criterion.

As concerns proficiency testing, the graphical consistency technique based on Mandel h-statistic and Grubb's tests for outliers detection were successfully introduced straight on the Bayesian outcome after proving the consistency of the underlying assumptions. However, we were unableto do the same with the Mandel k-statistic and Cochran test due to the attitude, said borrowing of precision, of the hierarchical Bayesian models.

Lastly, adding another level in the hierarchical Bayesian model for proficiency comparison in was possible to maintain the overall specific laboratory scores and to have at the same time a full accuracy evaluation of each product family from different manufacturers involved in the inter-laboratory test. This has been possible thanks to the evidence that with the fully nested hierarchical model established in equation 4.16, we did not lose consistency in accuracy evaluation for the specific Instron product line with respect to the outcome of the stand alone accuracy evaluation, even if the experimental dataset was enlarged by adding data drawn from the product line of other manufacturers. So, leveraging this "accuracy-retaining property" within each product line involved in the proficiency testing experiment, two scores for evaluating the equipments performance of different manufacturers are presented. The first statistic rates the product line variability with respect to an assigned target, while the second rates the behaviour of the different product line as a combined function of their bias and variability. In this way, the proficiency test scheme can be exploited not only for laboratories comparisons but also in order to compare the performance of the product lines of different manufacturers. Further analysis and studies have to be performed before giving universal validity to this conclusion, especially because unbalanced experimental plan have not already been tested. Actually, there are concerns that in the case of unbalanced plan as concerns the number of equipments for each product line the "accuracy-retaining property" is no more realistic. In this way the risk is that the final rating is not a reliable projection of the product line accuracy performance but a function of the represented market share within the subset of the participants to the proficiency testing.

In conclusion, let we remark once again that, as result of this research, the proposed hierarchical Bayesian models can be used as an alternative to those given in ISO 5725-2 and ISO 13528, the main international standards for the accuracy evaluation in inter-laboratory experiments and proficiency testing schemes, respectively. The Bayesian framework yields results consistent with those of the established methods, while offering an enhanced descriptive capability for accuracy evaluation, as the result is not just a point estimator or a confidence interval but the full

probability density for each accuracy parameter, that are trueness, precision under repeatability, homogeneity and reproducibility conditions. This really ought to be appreciated by people who are not deep inside the metrology environment because they can clearly distinguish all the accuracy components, as it was for some colleagues in Instron.

6 ANNEX A: VST ACCURACY EXPERIMENT MEASUREMENTS

#	y_{ijk}	Jump	LVDT	#	y_{ijk}	Jump	LVDT	#	y_{ijk}	Jump	LVDT
1	0.996	6»5	155534	97	1.006	6»5	148886	193	1.003	6»5	148880
2	1.006	6»5	155534	98	1.008	6»5	148886	194	1.001	6»5	148880
3	0.997	6»5	155534	99	1.006	6»5	148886	195	1.004	6»5	148880
4	0.998	6»5	155534	100	0.995	6»5	148886	196	1.004	6»5	148880
5	0.999	5»4	155534	101	1.006	5»4	148886	197	0.993	5»4	148880
6	0.998	5»4	155534	102	1.003	5»4	148886	198	0.993	5»4	148880
7	1.002	5»4	155534	103	1.004	5»4	148886	199	0.993	5»4	148880
8	1.006	5»4	155534	104	1.003	5»4	148886	200	0.990	5»4	148880
9	0.999	4»3	155534	105	0.991	4»3	148886	201	0.994	4»3	148880
10	0.999	4»3	155534	106	0.992	4»3	148886	202	0.994	4»3	148880
11	0.998	4»3	155534	107	0.994	4»3	148886	203	0.996	4»3	148880
12	1.011	4»3	155534	108	1.005	4»3	148886	204	0.996	4»3	148880
13	1.001	3»2	155534	109	1.003	3»2	148886	205	1.004	3»2	148880
14	1.001	3»2	155534	110	1.004	3»2	148886	206	1.003	3»2	148880
15	1.003	3»2	155534	111	1.004	3»2	148886	207	1.004	3»2	148880
16	1.003	3»2	155534	112	1.000	3»2	148886	208	1.004	3»2	148880
17	0.989	6»5	155652	113	0.995	6»5	148859	209	0.991	6»5	148890
18	0.990	6»5	155652	114	1.003	6»5	148859	210	0.990	6»5	148890
19	0.994	6»5	155652	115	1.001	6»5	148859	211	0.992	6»5	148890
20	0.992	6»5	155652	116	0.994	6»5	148859	212	0.994	6»5	148890
21	0.991	5»4	155652	117	1.001	5»4	148859	213	1.001	5»4	148890
22	0.992	5»4	155652	118	1.004	5»4	148859	214	1.000	5»4	148890
23	0.994	5»4	155652	119	1.003	5»4	148859	215	1.002	5»4	148890
24	0.993	5»4	155652	120	1.006	5»4	148859	216	1.002	5»4	148890
25	1.005	4»3	155652	121	0.994	4»3	148859	217	1.001	4»3	148890
26	1.006	4»3	155652	122	1.001	4»3	148859	218	1.009	4»3	148890
27	1.006	4»3	155652	123	0.998	4»3	148859	219	1.005	4»3	148890
28	1.006	4»3	155652	124	1.001	4»3	148859	220	0.999	4»3	148890
29	1.009	3»2	155652	125	1.003	3»2	148859	221	0.999	3»2	148890
30	1.009	3»2	155652	126	0.996	3»2	148859	222	1.003	3»2	148890
31	1.009	3»2	155652	127	0.999	3»2	148859	223	0.999	3»2	148890
32	1.009	3»2	155652	128	0.997	3»2	148859	224	0.998	3»2	148890
33	0.996	6»5	155530	129	1.001	6»5	148883	225	1.005	6»5	148866
34	1.006	6»5	155530	130	1.000	6»5	148883	226	1.007	6»5	148866
35	1.003	6»5	155530	131	1.003	6»5	148883	227	1.005	6»5	148866
36	1.002	6»5	155530	132	1.001	6»5	148883	228	1.011	6»5	148866
37	0.997	5»4	155530	133	0.995	5»4	148883	229	0.996	5»4	148866
38	1.005	5»4	155530	134	0.998	5»4	148883	230	0.998	5»4	148866
39	1.000	5»4	155530	135	0.998	5»4	148883	231	0.996	5»4	148866
40	1.002	5»4	155530	136	0.996	5»4	148883	232	0.996	5»4	148866
41	1.002	4»3	155530	137	0.999	4»3	148883	233	0.997	4»3	148866
42	1.000	4»3	155530	138	0.999	4»3	148883	234	0.992	4»3	148866
43	0.998	4»3	155530	139	1.003	4»3	148883	235	0.999	4»3	148866
44	1.005	4»3	155530	140	1.000	4»3	148883	236	0.994	4»3	148866
45	1.004	3»2	155530	141	0.998	3»2	148883	237	1.005	3»2	148866
46	1.006	3»2	155530	142	0.999	3»2	148883	238	1.005	3»2	148866
47	1.007	3»2	155530	143	1.003	3»2	148883	239	1.009	3»2	148866
48	1.012	3»2	155530	144	0.999	3»2	148883	240	1.008	3»2	148866

#	y_{ijk}	Jump	LVDT	#	y_{ijk}	Jump	LVDT	#	y_{ijk}	Jump	LVDT
49	0.999	6»5	148865	145	1.001	6»5	155651	241	1.005	6»5	148893
50	0.999	6»5	148865	146	1.003	6»5	155651	242	1.005	6»5	148893
51	1.000	6»5	148865	147	0.998	6»5	155651	243	1.005	6»5	148893
52	1.000	6»5	148865	148	0.995	6»5	155651	244	1.006	6»5	148893
53	0.995	5»4	148865	149	0.992	5»4	155651	245	1.005	5»4	148893
54	0.999	5»4	148865	150	0.992	5»4	155651	246	1.006	5»4	148893
55	0.999	5»4	148865	151	0.990	5»4	155651	247	1.007	5»4	148893
56	0.999	5»4	148865	152	0.990	5»4	155651	248	1.007	5»4	148893
57	0.999	4»3	148865	153	0.992	4»3	155651	249	0.990	4»3	148893
58	1.000	4»3	148865	154	1.003	4»3	155651	250	0.988	4»3	148893
59	1.000	4»3	148865	155	1.001	4»3	155651	251	0.990	4»3	148893
60	1.000	4»3	148865	156	1.000	4»3	155651	252	0.989	4»3	148893
61	0.998	3»2	148865	157	0.999	3»2	155651	253	1.014	3»2	148893
62	0.999	3»2	148865	158	0.998	3»2	155651	254	1.013	3»2	148893
63	0.999	3»2	148865	159	0.999	3»2	155651	255	1.013	3»2	148893
64	1.001	3»2	148865	160	0.998	3»2	155651	256	1.013	3»2	148893
65	1.000	6»5	155655	161	1.003	6»5	148857	257	0.992	6»5	155653
66	1.000	6»5	155655	162	1.004	6»5	148857	258	0.994	6»5	155653
67	1.001	6»5	155655	163	1.003	6»5	148857	259	0.991	6»5	155653
68	1.001	6»5	155655	164	1.003	6»5	148857	260	0.994	6»5	155653
69	1.006	5»4	155655	165	0.996	5»4	148857	261	1.000	5»4	155653
70	1.006	5»4	155655	166	0.997	5»4	148857	262	0.992	5»4	155653
71	1.008	5»4	155655	167	0.996	5»4	148857	263	0.996	5»4	155653
72	1.006	5»4	155655	168	0.996	5»4	148857	264	0.995	5»4	155653
73	0.991	4»3	155655	169	1.001	4»3	148857	265	1.001	4»3	155653
74	1.005	4»3	155655	170	1.001	4»3	148857	266	1.004	4»3	155653
75	0.989	4»3	155655	171	1.002	4»3	148857	267	1.001	4»3	155653
76	0.990	4»3	155655	172	1.004	4»3	148857	268	1.004	4»3	155653
77	0.995	3»2	155655	173	1.000	3»2	148857	269	1.001	3»2	155653
78	0.995	3»2	155655	174	1.003	3»2	148857	270	1.001	3»2	155653
79	0.996	3»2	155655	175	1.002	3»2	148857	271	1.003	3»2	155653
80	0.998	3»2	155655	176	1.002	3»2	148857	272	1.003	3»2	155653
81	1.001	6»5	155654	177	1.003	6»5	148900	273	1.000	6»5	155601
82	1.000	6»5	155654	178	0.999	6»5	148900	274	1.001	6»5	155601
83	1.003	6»5	155654	179	1.004	6»5	148900	275	1.008	6»5	155601
84	1.006	6»5	155654	180	1.004	6»5	148900	276	1.003	6»5	155601
85	0.999	5»4	155654	181	1.002	5»4	148900	277	1.006	5»4	155601
86	1.000	5»4	155654	182	1.004	5»4	148900	278	1.003	5»4	155601
87	0.997	5»4	155654	183	1.005	5»4	148900	279	1.006	5»4	155601
88	0.997	5»4	155654	184	1.005	5»4	148900	280	1.004	5»4	155601
89	0.999	4»3	155654	185	0.997	4»3	148900	281	0.996	4»3	155601
90	1.000	4»3	155654	186	0.995	4»3	148900	282	0.998	4»3	155601
91	1.000	4»3	155654	187	0.996	4»3	148900	283	0.999	4»3	155601
92	1.003	4»3	155654	188	0.996	4»3	148900	284	1.003	4»3	155601
93	0.995	3»2	155654	189	0.998	3»2	148900	285	0.991	3»2	155601
94	0.995	3»2	155654	190	0.999	3»2	148900	286	0.990	3»2	155601
95	1.002	3»2	155654	191	0.999	3»2	148900	287	0.999	3»2	155601
96	0.999	3»2	155654	192	1.000	3»2	148900	288	0.995	3»2	155601

7 ANNEX B: R SCRIPT FOR THE SULFUR CONTENT EXPERIMENT

```
setwd("C:/PhD/Thesis ")
d<-dim(data2)
i<-levels(data2$Laboratory)
p<-length(i)                    #nr. of Laboratories
#[1] 8
j<-levels(data2$Batch)
q<-length(j)                    #nr. of testing levels
#[1] 4
names(data2)[4] <- "xijk"       #sulfur measurements
attach(data2)

############ ISO 5725-2 Framework #############

### Arranging data like Figure 2
FormB <- tapply(xijk,list(Laboratory,Batch),mean)
round(FormB,3)                  #Cell means §7.2.9
#       LV1   LV2   LV3   LV4
#Lab1 0.708 1.205 1.688 3.240
#Lab2 0.680 1.217 1.643 3.200
#Lab3 0.667 1.297 1.613 3.370
#Lab4 0.660 1.203 1.667 3.203
#Lab5 0.690 1.248 1.650 3.216
#Lab6 0.733 1.373 1.720 3.290
#Lab7 0.703 1.240 1.690 3.247
#Lab8 0.677 1.253 1.673 3.257
FormC <- tapply(xijk,list(Laboratory,Batch),sd)
round(FormC,3)                  #Measure of cells spread §7.2.10
#       LV1   LV2   LV3   LV4
#Lab1 0.005 0.021 0.010 0.028
#Lab2 0.010 0.006 0.006 0.000
#Lab3 0.021 0.015 0.006 0.010
#Lab4 0.010 0.025 0.012 0.038
#Lab5 0.019 0.043 0.032 0.038
#Lab6 0.006 0.015 0.017 0.020
#Lab7 0.012 0.035 0.010 0.021
#Lab8 0.025 0.042 0.006 0.006
FormN <- tapply(xijk,list(Laboratory,Batch),length)
FormN                           #Number of replicates for each cell
#     LV1 LV2 LV3 LV4
#Lab1   4   4   4   4
#Lab2   3   3   3   3
#Lab3   3   3   3   3
#Lab4   3   3   3   3
#Lab5   5   4   5   5
#Lab6   3   3   3   3
#Lab7   3   3   3   3
#Lab8   3   3   3   3
FormV <- tapply(xijk,list(Laboratory,Batch),var)

AOV <- aov(xijk[Batch == j[1]]~Laboratory[Batch == j[1]])
summary(AOV)
#                          Df   Sum Sq   Mean Sq F value   Pr(>F)
#Laboratory[Batch == j[1]]  7 0.012555 0.0017935   7.849 0.000163 ***
#Residuals                 19 0.004342 0.0002285
#---
#Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

print(a<-model.tables(AOV, "means"),digits=4)
#Tables of means
#Grand mean
#
#0.6903704
#
# Laboratory[Batch == j[1]]
#      Lab1 Lab2   Lab3 Lab4 Lab5   Lab6   Lab7   Lab8
#    0.7075 0.68 0.6667 0.66 0.69 0.7333 0.7033 0.6767       #laboratory means
#rep 4.0000 3.00 3.0000 3.00 5.00 3.0000 3.0000 3.0000       #cell replicates
```

```
print(ni <- a$n$`Laboratory[Batch == j[1]]`)
#Laboratory[Batch == j[1]]
#Lab1 Lab2 Lab3 Lab4 Lab5 Lab6 Lab7 Lab8
#   4    3    3    3    5    3    3    3
print(ntot <- sum(ni))
#[1] 27        #that is the total number of measurements for the first batch
print(n1 <- 1/summary(AOV)[[1]]$'Df' [1]*(ntot-sum(ni^2)/ntot))
#[1] 3.354497  #that is the average number of measurements for the first batch

#initialization of the objected needed by the following for cycle
Table.B2 <- matrix(nrow = p, ncol = q*2)
mj <- c()              #the vector for the general means
ss2dj <- 1:q
ss2rj <- 1:q           #the vector of the repeatability variances
p2 <- 1:q              #the vector of the numbers of laboratory
n2 <- 1:q              #the vector of the replicates' average numbers

> for(w in 1:q){
+   AOV <- aov(xijk[Batch == j[w]]~Laboratory[Batch == j[w]])
+   ss2dj[w] <- summary(AOV)[[1]]$'Mean Sq'[1]
+   ss2rj[w] <- summary(AOV)[[1]]$'Mean Sq'[2]
+   p2[w] <- summary(AOV)[[1]]$'Df'[1]+1
+   ntot <-sum(summary(AOV)[[1]]$'Df')+1
+   a <- model.tables(AOV, "means")
+   ni <- a$n$`Laboratory[Batch == j[w]]`
+   Table.B2[,(2*w)]<-ni
+   Table.B2[,(2*w-1)]<-round(a$tables$`Laboratory[Batch == j[w]]`,3)
+   n2[w] <- 1/summary(AOV)[[1]]$'Df' [1]*(ntot-sum(ni^2)/ntot)
+   mj <- c(mj, a$tables$`Grand mean`)
+ }
rm(a, ni, ntot, n1)
print(ss2Lj <- (ss2dj-ss2rj)/n2)              #inter-laboratory variances
#[1] 0.0004665408 0.0028448195 0.0009171136 0.0027092114
print(ss2Rj <- ss2Lj + ss2rj)                 #reproducibility variances
[1] 0.0006950495 0.0036730603 0.0012087802 0.0033892114

### General mean for each level
print(xj <- round(apply(FormB*FormN,2,sum)/apply(FormN,2,sum),3))
#  LV1   LV2   LV3   LV4
#0.690 1.252 1.667 3.250
print(round(mj,3))
#[1] 0.690 1.252 1.667 3.250   #ANOVA leads to the same result of ISO 5725-2

TableB5 <- cbind(1:4, format(round(mj,3), nsmall = 3), round(sqrt(ss2rj),3),
                 round(sqrt(ss2Rj),3))
library(ggplot2)
library(grid)
library(gridExtra)
colnames(TableB5) <- c('Level[j]', paste0('hat(m)','^','(j)'),
                       paste0('s[r]','^','(j)'), paste0('s[R]','^','(j)'))
grid.newpage()
tt <- ttheme_default(colhead=list(fg_params = list(parse=TRUE)))
grid.table(TableB5, theme = tt)       #accuracy table according to ISO 5725-2



############ Bayesian Framework #############
library(R2OpenBUGS)
library(coda)
library(ggplot2)

data1 <- data2[Batch == "LV1",c(1,4)]
detach(data2)
attach(data1)

print(u <- tapply(xijk,list(Laboratory),length))
#Lab1 Lab2 Lab3 Lab4 Lab5 Lab6 Lab7 Lab8
#   4    3    3    3    5    3    3    3
```

```
print(u <- max(u)-tapply(xijk,list(Laboratory),length))
#Lab1 Lab2 Lab3 Lab4 Lab5 Lab6 Lab7 Lab8
#   1    2    2    2    0    2    2    2
a<-c()
for(w in 1:p){
   a<-c(a,xijk[Laboratory==i[w]],rep(NA,u[w]))
}
print(xik <- a)
# [1] 0.71 0.71 0.70 0.71   NA 0.69 0.67 0.68   NA   NA 0.66 0.65 0.69   NA   #
[15]   NA 0.67 0.65 0.66   NA   NA 0.70 0.69 0.66 0.71 0.69 0.73 0.74 0.73   #[
29]   NA   NA 0.71 0.71 0.69   NA   NA 0.70 0.65 0.68   NA   NA
a<-max(tapply(xijk,list(Laboratory),length))
print(data <- list (p=p, a=a, xik=t(structure(.Data=xik, .Dim=c(a,p)))))
#$p
#[1] 8
#
#$a
#[1] 5
#
#$xik
#     [,1] [,2] [,3] [,4] [,5]
#[1,] 0.71 0.71 0.70 0.71   NA
#[2,] 0.69 0.67 0.68   NA   NA
#[3,] 0.66 0.65 0.69   NA   NA
#[4,] 0.67 0.65 0.66   NA   NA
#[5,] 0.70 0.69 0.66 0.71 0.69
#[6,] 0.73 0.74 0.73   NA   NA
#[7,] 0.71 0.71 0.69   NA   NA
#[8,] 0.70 0.65 0.68   NA   NA

inits <- function(){
  list(B = rnorm(p, 0, 100), mu.B = rnorm(1, 0, 1), sigma.R = runif(1, 0, 100),

       sigma.r = runif(1, 0, 100))
}
> LV1.sim <- bugs(data, inits, model.file = "C:/PhD/Thesis/sulfur2a.txt",
       parameters = c("B", "mu.B", "sigma.R", "sigma.r"), n.iter = 1000)

print(LV1.sim)
#Inference for Bugs model at " C:/PhD/Thesis/sulfur2a.txt",
#Current: 3 chains, each with 1000 iterations (first 500 discarded)
#Cumulative: n.sims = 1500 iterations saved
#           mean  sd   2.5%    25%    50%    75%  97.5% Rhat n.eff
#B[1]        0.7 0.0    0.7    0.7    0.7    0.7    0.7    1  1500
#B[2]        0.7 0.0    0.7    0.7    0.7    0.7    0.7    1   800
#B[3]        0.7 0.0    0.7    0.7    0.7    0.7    0.7    1  1500
#B[4]        0.7 0.0    0.6    0.7    0.7    0.7    0.7    1   690
#B[5]        0.7 0.0    0.7    0.7    0.7    0.7    0.7    1  1100
#B[6]        0.7 0.0    0.7    0.7    0.7    0.7    0.7    1   610
#B[7]        0.7 0.0    0.7    0.7    0.7    0.7    0.7    1   830
#B[8]        0.7 0.0    0.7    0.7    0.7    0.7    0.7    1   220
#mu.B        0.7 0.0    0.7    0.7    0.7    0.7    0.7    1   520
#sigma.R     0.0 0.0    0.0    0.0    0.0    0.0    0.1    1   740
#sigma.r     0.0 0.0    0.0    0.0    0.0    0.0    0.0    1   790
#deviance -147.4 5.6 -155.7 -151.6 -148.4 -144.0 -134.8    1   970

#For each parameter, n.eff is a crude measure of effective sample size,
#and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

#DIC info (using the rule, pD = Dbar-Dhat)
#pD = 7.9 and DIC = -139.5
#DIC is an estimate of expected predictive error (lower deviance is better).

MCR <- LV1.sim$sims.list$B
n <- length(MCR[,1])
MCR <- as.vector(MCR)
MCR <- data.frame(y1ik = MCR, i = as.vector(t(matrix(rep(i, n), p, n))))
rm(n)
```

```
ggplot(data = MCR, aes(x=i, y=y1ik)) + geom_boxplot(aes(fill = 'lightgreen')) +

   labs(x = "", y = expression(paste(y[ik]^(1), '[%]'))) +
   theme(axis.text=element_text(size=12), axis.title=element_text(size=18, face
="bold"), legend.text=element_text(size=12), legend.title=element_text(size=12,
 face="bold")) +
   geom_point(data = data.frame(x=Laboratory,y =xijk), aes(x=x, y=y, color='dar
kred'), size=3) +
   scale_fill_identity(name = 'Boxplot', guide = 'legend',labels = c('Bayesian'
)) +
   scale_colour_manual(name = 'Points', values =c('darkred'='darkred'), labels
= c('Raw data'))
rm(MCR)

### Comparison
round(LV1.sim$mean$B,3)                 #Bayesian outcome
#[1] 0.705 0.681 0.670 0.664 0.690 0.727 0.701 0.679
round(FormB[,1],3)                      #ISO 5725-2 outcome
# Lab1  Lab2  Lab3  Lab4  Lab5  Lab6  Lab7  Lab8
#0.708 0.680 0.667 0.660 0.690 0.733 0.703 0.677
VS <- data.frame(c(round(FormB[,1],3), round(LV1.sim$mean$B,3)),c(i,i),
                 c(rep("Sample", p), rep("Bayesian", p)))
colnames(VS) <- c("Bi", "Laboratory_i", "Model")
ggplot(VS, aes(Laboratory_i, Bi)) +
geom_point(aes(colour = factor(Model)), position=position_dodge(0.5), size=3) +

   labs(x = "", y = expression(paste(b[i]^(1), ' [%]'))) +
   geom_hline(yintercept = LV1.sim$mean$mu.B, lty = 4, size = 1.05) +
   scale_color_manual(name = 'Mean', breaks = levels(VS$Model), values=c("darkg
reen", "darkred")) +
   theme(axis.text=element_text(size=12), axis.title=element_text(size=18,face=
"bold"), legend.text=element_text(size=12),
legend.title=element_text(size=12, face="bold"))
rm(VS)

TableHB5 <- data.frame(1, mB1 = format(round(LV1.sim$mean$mu.B, 3), nsmall=3),
                       sr1 = round(LV1.sim$mean$sigma.r, 3),
                       sR1 = round(LV1.sim$mean$sigma.R, 3))
colnames(TableHB5) <- c('Level[j]', paste0('m[B]','^','(j)'), paste0('s[r]','^'
,'(j)'), paste0('s[R]','^','(j)'))
grid.newpage()
grid.table(TableHB5, theme = tt)

### General mean posterior parameters
print(round(LV1.sim$mean$mu.B,3))
#[1] 0.69                        #expected value of the general mean parameter
print(round(LV1.sim$sd$mu.B,3))
#[1] 0.011                       #standard deviation of the general mean parameter
### Repeatability posterior parameters
print(round((LV1.sim$mean$sigma.r/LV1.sim$sd$sigma.r)^2,1))
#[1] 32.8                        #shape parameter of the repeatability parameter
print(round(LV1.sim$mean$sigma.r/LV1.sim$sd$sigma.r^2,0))
#[1] 2004                        #scale parameter of the repeatability parameter
### Reproducibility posterior parameters
print(round((LV1.sim$mean$sigma.R/LV1.sim$sd$sigma.R)^2,1))
#[1] 6.5                         #shape parameter of the reproducibity parameter
print(round(LV1.sim$mean$sigma.R/LV1.sim$sd$sigma.R^2,0))
#[1] 237                         #scale parameter of the reproducibity parameter

p1 <- ggplot(data.frame(muB = LV1.sim$sims.list$mu.B), aes(muB)) +
   geom_density(aes(color="Bayesian"), adjust=1, alpha=0.1) + ylim(0, 150) +
   labs(x = expression(paste(mu[B]^(1), ' [%]')), title = "General mean", subti
tle = paste0('Batch: ', j[1])) +
   theme(title=element_text(size=15,face="bold"), axis.text=element_text(size=1
8), axis.title=element_text(size=18,face="bold"), legend.position="bottom", leg
end.text=element_text(size=15)) +
   stat_function(aes(color = "Theoretical"), fun = dnorm, args = list(mean = LV
1.sim$mean$mu.B, sd = LV1.sim$sd$mu.B), lwd = 1.2, lty = 4) +
   geom_vline(xintercept = LV1.sim$mean$mu.B, col = 'darkgrey') +
```

```
   scale_colour_manual("Legend: ", values=c("Bayesian"="black","Theoretical"="d
arkred"))
a <- (LV1.sim$mean$sigma.r/LV1.sim$sd$sigma.r)^2
b <- LV1.sim$mean$sigma.r/LV1.sim$sd$sigma.r^2
p2 <- ggplot(data.frame(sr = LV1.sim$sims.list$sigma.r), aes(sr)) +
   geom_density(aes(color="Bayesian"), adjust=1, alpha=0.1) +
   xlim(0, .13) + ylim(0, 150) +
   labs(x = expression(paste(sigma[r]^(1), ' [%]')),
title="Repeatability standard deviation", subtitle = paste0('Batch: ', j[1])) +

   theme(title=element_text(size=15,face="bold"), axis.text=element_text(size=1
8), axis.title=element_text(size=18,face="bold"), legend.position = "bottom", l
egend.text=element_text(size=15)) +
   stat_function(aes(color = "Theoretical"), fun = dgamma, args = list(shape=a,
 rate = b), lwd = 1.2, lty = 4) +
   geom_vline(xintercept = LV1.sim$mean$sigma.r, col = 'darkgrey') +
   scale_colour_manual("Legend: ", values=c("Bayesian"="black", "Theoretical"="
darkred"))

a <- (LV1.sim$mean$sigma.R/LV1.sim$sd$sigma.R)^2
b <- LV1.sim$mean$sigma.R/LV1.sim$sd$sigma.R^2
p3 <- ggplot(data.frame(sR = LV1.sim$sims.list$sigma.R), aes(sR)) +
   geom_density(aes(color="Bayesian"), adjust=1, alpha=0.1) +
   xlim(0, .13) + ylim(0, 150) +
   labs(x = expression(paste(sigma[R]^(1), ' [%]')),
title="Reproducibility standard deviation", subtitle=paste0('Batch: ', j[1])) +

   theme(title=element_text(size=15,face="bold"), axis.text=element_text(size=1
8), axis.title=element_text(size=18,face="bold"), legend.position = "bottom", l
egend.text=element_text(size=15)) +
   stat_function(aes(color="Theoretical"), fun=dgamma, args=list(shape=a, rate=
b), lwd = 1.2, lty = 4) +
   geom_vline(xintercept = LV1.sim$mean$sigma.r, col = 'darkgrey') +
   scale_colour_manual("Legend: ", values = c("Bayesian" ="black", "Theoretical
" = "darkred"))
grid.arrange(p1,p2,p3, ncol = 3)

TableHB5 <- data.frame(1,
   dl = round(quantile(LV1.sim$sims.list$mu.B-LV1.sim$mean$mu.B, .025), 2),
   du = round(quantile(LV1.sim$sims.list$mu.B-LV1.sim$mean$mu.B, .975), 2),
   ur = round(quantile(LV1.sim$sims.list$sigma.r, .95),3),
   uR = round(quantile(LV1.sim$sims.list$sigma.R, .95),3))
colnames(TableHB5) <- c('Level[j]', paste0('hat(delta)[lower]','^','(j)'),
   paste0('hat(delta)[upper]','^','(j)'), paste0('u[r]','^','(j)'),
   paste0('u[R]','^','(j)'))
grid.newpage()
grid.table(TableHB5, theme = tt)


### Model checking
u <- 2^(9:15)
n <- length(u)
mt <- u
sr <- u
sR <- u
ti <- matrix(1,p,n)
for(w in 1:n){
   inits <- function(){
     list(B = rnorm(p, 0, 100), mu.B = rnorm(1, 0, 1),
          sigma.R = runif(1, 0, 100), sigma.r = runif(1, 0, 100))
   }
   LV1.sim2 <- bugs(data, inits, model.file = "C:/PhD/Thesis/sulfur2a.txt",
                    parameters = c("B", "mu.B", "sigma.R", "sigma.r"),
                    n.chains = 3, n.iter = u[w])
   ti[,w] <- LV1.sim2$mean$B
   mt[w] <- LV1.sim2$mean$mu.B
   sr[w] <- LV1.sim2$mean$sigma.r
   sR[w] <- LV1.sim2$mean$sigma.R
}
```

```
n <- length(u)
u <- data.frame(c(mt,as.vector(t(ti))),rep(u,p+1),
  c(rep('m[B]',n),as.vector(t(matrix(rep(paste0('B[',1:p,']'),n), nrow = p)))))
colnames(u) <- c('Fit', 'N.iter', 'Theta')
p1 <- ggplot(data = u, aes(x=N.iter, y=Fit)) +
   geom_point(aes(colour=factor(Theta)), size = 2.1) +
   labs(x = "Nr. of iterations", y=expression(paste('E[', theta[i], '] [%]')),
color=expression(theta[i])) +
   geom_vline(xintercept = 1000, lty = 4, size = 1.05) +
   theme(axis.text=element_text(size=15), axis.title=element_text(size=18,face=
"bold"), legend.text=element_text(size=15), legend.title=element_text(size=15))
 +
   scale_colour_discrete(breaks = levels(u$Theta), labels=c(expression(B[1]), e
xpression(B[2]), expression(B[3]), expression(B[4]), expression(B[5]), expressi
on(B[6]), expression(B[7]), expression(B[8]), expression(m[B])))

u <- 2^(9:15)
u <- data.frame(c(sr, sR), rep(u, 2), c(rep('sr', n), rep('sR', n)))
colnames(u) <- c('Fit', 'N.iter', 'Theta')
p2 <- ggplot(data = u, aes(x=N.iter, y=Fit)) +
   geom_point(aes(colour = factor(Theta)), size = 2.1) +
   labs(x = "Nr. of iterations",  =expression(paste('E[', theta[i], '] [%]')),
color=expression(theta[i])) +
   geom_vline(xintercept = 1000, lty = 4, size = 1.05) +
   theme(axis.text=element_text(size=15), axis.title=element_text(size=18,face=
"bold"), legend.text=element_text(size=15), legend.title=element_text(size=15))
 +
   scale_colour_discrete(breaks=levels(u$Theta),
labels=c(expression(paste(' ', sigma[r])), expression(paste('  ', sigma[R]))))
grid.arrange(p1,p2, nrow = 2)
rm(n,u)


### Another model having gamma priors for the inverse of variance parameters
inits <- function(){
   list(B = rnorm(p, 0, 100), mu.B = rnorm(1, 0, 1),
   tau.R = rgamma(1, 0.01, 0.01), tau.r = rgamma(1, 0.01, 0.01))
 }
LV1.sim3 <- bugs(data, inits, model.file = "C:/PhD/Thesis/sulfur3.txt",
                 parameters = c("B", "mu.B", "tau.R", "tau.r"),
                 n.chains = 3, n.iter = 1000)
TableHBB5 <- data.frame(1, mB1=format(round(LV1.sim3$mean$mu.B, 3), nsmall=3),
                        sr1 = round(sqrt(1/LV1.sim3$mean$tau.r), 3),
                        sR1 = round(sqrt(1/LV1.sim3$mean$tau.R), 3))
colnames(TableHBB5) <- c('Level[j]', paste0('m[B]','^','(j)'),
                         paste0('s[r]','^','(j)'), paste0('s[R]','^','(j)'))
grid.newpage()
grid.table(TableHBB5, theme = tt)
TableHBB5 <- data.frame(1, dl = round(qnorm(.025, 0, LV1.sim3$sd$mu.B), 3),
                 du=round(qnorm(.975, 0, LV1.sim3$sd$mu.B), 3),
                 ur=round(1/sqrt(quantile(LV1.sim3$sims.list$tau.r, .05)),3),
                 uR = round(1/sqrt(quantile(LV1.sim3$sims.list$tau.R, .05)),3))
colnames(TableHBB5) <- c('Level[j]', paste0('hat(delta)[lower]','^','(j)'),
                         paste0('hat(delta)[upper]','^','(j)'),
                         paste0('u[r]','^','(j)'), paste0('u[R]','^','(j)'))
grid.newpage()
grid.table(TableHBB5, theme = tt)
TableD <- data.frame(c('3.04', '3.11'), c('Uniform', 'Gamma'),
                     c(LV1.sim$pD, LV1.sim3$pD), c(LV1.sim$DIC, LV1.sim3$DIC))
colnames(TableD) <- c('Model', 'Precision/prior', paste0('p[D]','^','(1)'),
                      paste0('DIC','^','(1)'))
grid.newpage()
grid.table(TableD, theme = tt)


### Other levels for the batch of material
bij <- c()
mBj <- c()
```

```
srj <- c()
sRj <- c()
Par <- matrix(nrow = q, ncol = 8)
myplots <- list()
for(z in 1:q){
   detach(data1)
   data1 <- data2[data2$Batch == j[z],c(1,4)]
   attach(data1)
   u <- tapply(xijk,list(Laboratory),length)
   u <- max(u)-tapply(xijk,list(Laboratory),length)
   a<-c()
   for(w in 1:p){
      a<-c(a,xijk[Laboratory==i[w]],rep(NA,u[w]))
   }
   xik <- a
   a<-max(tapply(xijk,list(Laboratory),length))
   data <- list (p=p, a=a, xik=t(structure(.Data=xik, .Dim=c(a,p))))
   inits <- function(){
     list(B = rnorm(p, 0, 100), mu.B = rnorm(1, 0, 1),
          sigma.R = runif(1, 0, 100), sigma.r = runif(1, 0, 100))
   }
   LV1.sim <- bugs(data, inits, model.file = "C:/PhD/Thesis/sulfur2a.txt",
                   parameters=c("B", "mu.B", "sigma.R", "sigma.r"), n.iter=1000)
   data1 <- data2[which(data2$Batch==j[z]),]
   n <- length(LV1.sim$sims.list$mu.B)
   myplots[[z]] <- ggplot(data=data.frame(yijk=as.vector(LV1.sim$sims.list$B),
                i = as.vector(t(matrix(rep(i, n), p, n)))), aes(x=i, y=yijk)) +

     geom_boxplot(aes(fill = 'lightgreen')) +
     theme(axis.text=element_text(size=12),
   axis.title=element_text(size=18,face="bold"),
   legend.text=element_text(size=12),
   legend.title=element_text(size=12, face="bold"),
   legend.position="bottom", plot.subtitle=element_text(size=15, hjust=0.5)) +
     geom_point(data=data.frame(i = data1$Laboratory, yijk=data1$xijk,
                   j=data1$Batch), aes(x=i, y=yijk, color='darkred'), size=3) +
     labs(x="",y=expression(paste(y[ik]^(j),'[%]')), subtitle=paste('j =',z)) +
     scale_fill_identity(name='Boxplot', guide='legend', labels=c('Bayesian'))+
     scale_colour_manual(name = 'Points', values =c('darkred'='darkred'),
   labels = c('Raw data'))
   bij <- c(bij, as.vector(LV1.sim$mean$B))
   mBj <- c(mBj, LV1.sim$sims.list$mu.B)
   Par[z, 1] <- LV1.sim$mean$mu.B
   Par[z, 2] <- LV1.sim$sd$mu.B
   Par[z, 3] <- (LV1.sim$mean$sigma.r/LV1.sim$sd$sigma.r)^2
   Par[z, 4] <- LV1.sim$mean$sigma.r/LV1.sim$sd$sigma.r^2
   Par[z, 5] <- (LV1.sim$mean$sigma.R/LV1.sim$sd$sigma.R)^2
   Par[z, 6] <- LV1.sim$mean$sigma.R/LV1.sim$sd$sigma.R^2
   Par[z, 7] <- LV1.sim$mean$sigma.r
   Par[z, 8] <- LV1.sim$mean$sigma.R
   srj <- c(srj, LV1.sim$sims.list$sigma.r)
   sRj <- c(sRj, LV1.sim$sims.list$sigma.R)
 }
multiplot(plotlist = myplots, cols = 2)

VS <- data.frame(bij = c(as.vector(FormB), bij), Lab_i = c(rep(i,q),rep(i,q)),
                 Model = c(rep("Sample", p*q), rep("Bayesian", p*q)),
                 LV = rep(as.vector(t(matrix(rep(paste('j =',1:q), p),
                          nrow = q))),2))

ggplot(VS, aes(Lab_i, bij)) +
   geom_point(aes(colour=factor(Model)), position=position_dodge(0.5), size=3)+
   labs(x = "", y = expression(paste(b[i]^(j), ' [%]'))) +
   scale_color_manual(name='Mean', breaks=levels(VS$Model), values=c("darkgreen
", "darkred")) +
   theme(axis.text=element_text(size=12),
axis.title=element_text(size=18, face="bold"),
legend.text=element_text(size=12), legend.position = 'bottom',
legend.title=element_text(size=12, face="bold"),
strip.text.x = element_text(size = 12, color = 'darkorange', face="bold")) +
```

```
   facet_wrap(~ LV, ncol = 2, scales="free_y") +
   geom_hline(data = data.frame(LV = paste('j =',1:q), bij = Par[,1]), aes(yint
ercept = bij), lty = 4, size = 1.05)
rm(VS)

gg_color_hue <- function(n) {
   hues = seq(15, 375, length = n + 1)
   hcl(h = hues, l = 65, c = 100)[1:n]
}

ggplot(data.frame(mB = mBj,
                  Batch=as.vector(t(matrix(rep(paste('j = ',1:q), n), nrow=q)))),
   aes(mB, fill = Batch, colour = Batch)) + geom_density(adjust=1, alpha=0.1) +

   labs(x = expression(paste(mu[B], ' [mm]'))) +
   stat_function(fun = dnorm, args = list(mean = Par[1,1], sd = Par[1,2]),
col = gg_color_hue(q)[1], lwd = 1.2, lty = 4) +
   stat_function(fun = dnorm, args = list(mean = Par[2,1], sd = Par[2,2]),
col = gg_color_hue(q)[2], lwd = 1.2, lty = 4) +
   stat_function(fun = dnorm, args = list(mean = Par[3,1], sd = Par[3,2]),
col = gg_color_hue(q)[3], lwd = 1.2, lty = 4) +
   stat_function(fun = dnorm, args = list(mean = Par[4,1], sd = Par[4,2]),
col = gg_color_hue(q)[4], lwd = 1.2, lty = 4) +
   theme(axis.text=element_text(size=12), axis.title=element_text(size=18,face=
"bold"),
legend.text=element_text(size=15),
legend.title=element_text(size=15, face = 'bold')) +
   geom_hline(yintercept = 0, size = 1.05)

p1 <- ggplot(data.frame(mB = srj,
             Batch = as.vector(t(matrix(rep(paste('j = ', 1:q),n), nrow=q)))),
      aes(mB, fill=Batch, colour=Batch)) + geom_density(adjust=1, alpha=0.1) +
   labs(x = expression(paste(sigma[r], ' [mm]'))) +
   stat_function(fun = dgamma, args = list(shape = Par[1,3], rate = Par[1,4]),
col = gg_color_hue(q)[1], lwd = 1.2, lty = 4) +
   stat_function(fun = dgamma, args = list(shape = Par[2,3], rate = Par[2,4]),
col = gg_color_hue(q)[2], lwd = 1.2, lty = 4) +
   stat_function(fun = dgamma, args = list(shape = Par[3,3], rate = Par[3,4]),
col = gg_color_hue(q)[3], lwd = 1.2, lty = 4) +
   stat_function(fun = dgamma, args = list(shape = Par[4,3], rate = Par[4,4]),
col = gg_color_hue(q)[4], lwd = 1.2, lty = 4) +
   theme(axis.text=element_text(size=12), axis.title=element_text(size=18,face=
"bold"),
legend.text=element_text(size=15),
legend.title=element_text(size=15, face = 'bold')) +
   geom_hline(yintercept = 0, size = 1.05) + ylim(0, 150) + xlim(0, 0.25)
p2 <- ggplot(data.frame(mB = sRj,
             Batch=as.vector(t(matrix(rep(paste('j = ', 1:q),n), nrow =q)))),
      aes(mB, fill=Batch, colour=Batch)) + geom_density(adjust=1, alpha=0.1) +
   labs(x = expression(paste(sigma[R], ' [mm]'))) +
   stat_function(fun = dgamma, args = list(shape = Par[1,5], rate = Par[1,6]),
col = gg_color_hue(q)[1], lwd = 1.2, lty = 4) +
   stat_function(fun = dgamma, args = list(shape = Par[2,5], rate = Par[2,6]),
col = gg_color_hue(q)[2], lwd = 1.2, lty = 4) +
   stat_function(fun = dgamma, args = list(shape = Par[3,5], rate = Par[3,6]),
col = gg_color_hue(q)[3], lwd = 1.2, lty = 4) +
   stat_function(fun = dgamma, args = list(shape = Par[4,5], rate = Par[4,6]),
col = gg_color_hue(q)[4], lwd = 1.2, lty = 4) +
   theme(axis.text=element_text(size=12), axis.title=element_text(size=18,face=
"bold"),
legend.text=element_text(size=15),
legend.title=element_text(size=15, face = 'bold')) +
   geom_hline(yintercept = 0, size = 1.05) + ylim(0, 150) + xlim(0, 0.25)
grid.arrange(p1,p2, ncol = 2)

TableHB5 <- data.frame(1:q, mB1 = format(round(Par[,1], 3), nsmall = 3),
                       sr1 = round(Par[,7], 3), sR1 = round(Par[,8], 3))
colnames(TableHB5) <- c('Level[j]', paste0('m[B]','^','(j)'), paste0('s[r]','^'
,'(j)'), paste0('s[R]','^','(j)'))
grid.newpage()
```

```
grid.table(TableHB5, theme = tt)
```

8 ANNEX C: R SCRIPT FOR THE VST ACCURACY EXPERIMENT

```
setwd("C:/PhD/Thesis ")
data3<-read.table("LVDT.txt",header = TRUE, dec = ",")
data3 <- data3[-which(data3$LVDT==data3$LVDT[length(data3$LVDT)]),]
data3$LVDT <- as.factor(data3$LVDT)
i<-levels(data3$LVDT)
p<-length(i)                          #nr. of testing equipments
j<-levels(data3$Jump)
q<-length(j)                          #nr. of testing levels
k<-levels(data3$round)
n<-length(k)                          #nr. of replicates
data3 <- data3[order(data3$LVDT, data3$Jump),]
attach(data3)
data3 <- cbind(data3,as.numeric(LVDT))
### ISO 5725-2 framework
FormB <- tapply(yijk,list(LVDT,Jump),mean)
round(FormB,3)                        #Cell means §7.2.9
#         3»2   4»3   5»4   6»5
#148857 1.002 1.002 0.996 1.003
#…       …     …     …     …
#155655 0.996 0.994 1.006 1.000
FormC <- tapply(yijk,list(LVDT,Jump),sd)
round(FormC,4)                        #Measure of cells spread §7.2.10
#          3»2    4»3    5»4    6»5
#148857 0.0013 0.0014 0.0005 0.0005
#…       …      …      …      …
#155655 0.0014 0.0075 0.0010 0.0006
FormN <- tapply(yijk,list(LVDT,Jump),length)
yj <- apply(FormB*FormN,2,sum)/apply(FormN,2,sum)
round(yj, 3)                          #General means
#  3»2   4»3   5»4   6»5
#1.002 0.999 0.999 1.000
library(ggplot2)
library(grid)
library(gridExtra)
mj <- 1:q
s2rj <- 1:q
s2dj <- 1:q
smj <- 1:q
pj <- 1:q
nj <- 1:q
for(w in 1:q){
  A <- data3$LVDT[Jump==j[w]]
  AOV1 <- aov(data3$yijk[Jump==j[w]]~A, contrasts = list(A = "contr.sum"))
  mj[w] <- AOV1$coefficients[1]
  s2dj[w] <- summary(AOV1)[[1]]$'Mean Sq'[1]
  s2rj[w] <- summary(AOV1)[[1]]$'Mean Sq'[2]
  smj[w] <- coef(summary.lm(AOV1))[1, "Std. Error"]
  pj[w] <- summary(AOV1)[[1]]$'Df'[1]+1
  nj[w] <- summary(AOV1)[[1]]$'Df'[2]/pj[w]+1
  }
s2Lj <- (s2dj-s2rj)/nj
s2Rj <- s2Lj + s2rj
round(mj, 3)                          #same result if compared with yj above
#[1] 1.002 0.999 0.999 1.000
mj <- round(mj,4)
srj <- round(sqrt(s2rj),4)
sRj <- round(sqrt(s2Rj),4)
sLj <- round(sqrt(s2Lj),4)
TableB <- data.frame(1:4, j, pj, mj, srj, sRj)
colnames(TableB) <- c('j',paste0('Level','^','(j)'), paste0('p','^','(j)'), pas
teO('hat(m)','^','(j)'), paste0('s[r]','^','(j)'), paste0('s[H]','^','(j)'))
```

```
tt <- ttheme_default(colhead=list(fg_params = list(parse=TRUE)))
grid.newpage()
grid.table(TableB, theme = tt)
c <- qt(.975, df=(n-1)*p)
TableB <- data.frame(TableB[,1:3], round(mj-c*srj/sqrt(n*p),4), round(mj+c*srj/
sqrt(n*p),4), TableB[,5:6])
colnames(TableB) <- c('j',paste0('Level','^','(j)'), paste0('p','^','(j)'), pas
te0('hat(m)[lower]','^','(j)'), paste0('hat(m)[upper]','^','(j)'), paste0('s[r]
','^','(j)'), paste0('s[H]','^','(j)'))
grid.newpage()
grid.table(TableB, theme = tt)
# Uncertainty budget
TableU <- data.frame(Source = c('Measurements', 'Resolution', 'Block[4]',
                                'Block[5]', 'Temperature'),
                pdf = c('T-student', 'Uniform', 'Normal', 'Normal', 'Uniform'),
                E = c(mj[3], 0, 4, 5, 0),
                U = c(srj[3], round(0.001/sqrt(12),4), 0.0001/2, 0.0001/2,
                      round(0.2/sqrt(12),4)),
                v = c(p*(n-1), 8, 60, 60, 8),
                c = c(1, 1, 1, 1, 10.8E-6))
TableU <- data.frame(i = 1:length(TableU[,1]), TableU,
                     cU2 = round((TableU$c*TableU$U)^2,15))
colnames(TableU) <- c('i', 'Source','pdf', 'bar(X[i])', 'u(x[i])', 'nu[i]', 'c[
i]', paste0('u[i]','^','2','~(y)'))
row.names(TableU) <- c(paste0('Y[r]','^','(3)'), paste0('Y[d]','^','(3)'), past
e0('T[4]','^','(3)'), paste0('T[5]','^','(3)'),paste0('Delta~T','^','(3)'))
grid.newpage()
tt <- ttheme_default(colhead=list(fg_params = list(parse=TRUE)), rowhead=list(f
g_params = list(parse=TRUE)))
grid.table(TableU, theme = tt)
print(uc <- round(sqrt(sum(TableU$`u[i]^2~(y)`)),5))
#[1] 0.00192                    #same magnitude of the repeatability precision
#Bias 95 % credibility interval and expanded precisions
TableU <- TableB
TableU$`hat(m)[lower]^(j)` <- TableU$`hat(m)[lower]^(j)`-1
TableU$`hat(m)[upper]^(j)` <- TableU$`hat(m)[upper]^(j)`-1
TableU$`s[r]^(j)` <- 2*TableU$`s[r]^(j)`
TableU$`s[H]^(j)` <- 2*TableU$`s[H]^(j)`
colnames(TableU) <- c('j',paste0('Level','^','(j)'), paste0('p','^','(j)'), pas
te0('hat(delta)[lower]','^','(j)'), paste0('hat(delta)[upper]','^','(j)'), past
e0('U[r]','^','(j)'), paste0('U[H]','^','(j)'))
grid.newpage()
tt <- ttheme_default(colhead=list(fg_params = list(parse=TRUE)))
grid.table(TableU, theme = tt)

############ Bayesian Framework #############
library(R2OpenBUGS)
library(coda)
library(ggplot2)
detach(data3)
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
   library(grid)
   # Make a list from the ... arguments and plotlist
   plots <- c(list(...), plotlist)
   numPlots = length(plots)
   # If layout is NULL, then use 'cols' to determine layout
   if (is.null(layout)) {
     # Make the panel
     # ncol: Number of columns of plots
     # nrow: Number of rows needed, calculated from # of cols
     layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                      ncol = cols, nrow = ceiling(numPlots/cols))
   }
```

```
   if (numPlots==1) {
     print(plots[[1]])
   } else {
     # Set up the page
     grid.newpage()
     pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))
     # Make each plot, in the correct location
     for (i in 1:numPlots) {
       # Get the i,j matrix positions of the regions that contain this subplot
       matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))
       print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                       layout.pos.col = matchidx$col))
     }
   }
 }

data1 <- data3[data3$Jump == j[1],c(1,4)]
attach(data1)
inits <- function(){
   list(B = rnorm(p, 0, 100), mu.B = rnorm(1, 0, 1),
        sigma.R = runif(1, 0, 100), sigma.r = runif(1, 0, 100))
}
print(data <- list (p=p, n=n, yik=t(structure(.Data=yijk, .Dim=c(n,p)))))
#$p
#[1] 18
#
#$n
#[1] 4
#
#$yik
#       [,1]  [,2]  [,3]  [,4]
# [1,] 1.000 1.003 1.002 1.002
# [2,] 1.003 0.996 0.999 0.997
# [3,] 0.998 0.999 0.999 1.001
# [4,] 1.005 1.005 1.009 1.008
# [5,] 1.004 1.003 1.004 1.004
# [6,] 0.998 0.999 1.003 0.999
# [7,] 1.003 1.004 1.004 1.000
# [8,] 0.999 1.003 0.999 0.998
# [9,] 1.014 1.013 1.013 1.013
#[10,] 0.998 0.999 0.999 1.000
#[11,] 1.004 1.006 1.007 1.012
#[12,] 1.001 1.001 1.003 1.003
#[13,] 0.991 0.990 0.999 0.995
#[14,] 0.999 0.998 0.999 0.998
#[15,] 1.009 1.009 1.009 1.009
#[16,] 1.001 1.001 1.003 1.003
#[17,] 0.995 0.995 1.002 0.999
#[18,] 0.995 0.995 0.996 0.998

LV1.sim <- bugs(data, inits, model.file = "C:/PhD/Thesis/LVDT_model.txt",
                parameters = c("B", "mu.B", "sigma.R", "sigma.r"), n.iter=1500)
ggplot(data = data.frame(i = as.vector(t(matrix(rep(i, dim(LV1.sim$sims.list$B)
[1]), nrow = p))),
                    y1ik = as.vector(LV1.sim$sims.list$B)), aes(x=i, y=y1ik)) +
   geom_boxplot(aes(fill = 'lightgreen')) +
   labs(y = expression(paste(y[ik]^(1), '[%]'))) +
   theme(axis.text=element_text(size=12),
         axis.title=element_text(size=18,face="bold"),
         legend.text=element_text(size=12),
         legend.title=element_text(size=12, face="bold"),
         axis.title.x=element_blank()) +
   geom_point(data = data.frame(x = as.character(LVDT), y = yijk),
```

```
            aes(x=x, y=y, color = 'darkred'), size = 3) +
   scale_fill_identity(name='Boxplot', guide='legend',labels = c('Bayesian')) +
   scale_colour_manual(name='Points', values=c('darkred'='darkred'),
                       labels=c('Raw data')) +
   geom_hline(yintercept = 1, size = 1.05) +
   geom_hline(yintercept = c(0.99, 1.01), size = 1.05, lty = 4)
detach(data1)

sim.mB <- c()
sim.sr <- c()
sim.sR <- c()
LV <- c()
pD <- c()
DIC <- c()
Par <- matrix(nrow = q, ncol = 8)
TP <- matrix(nrow = q, ncol = 4)
B <- c()
myplots <- list()

for(w in 1:q){
  data1 <- data3[data3$Jump == j[w],c(1,4)]
  attach(data1)
  data <- list (p=p, n=n, yik=t(structure(.Data=(yijk-1), .Dim=c(n,p))))
  inits <- function(){
    list(B = rnorm(p, 0, 100), mu.B = rnorm(1, 0, 1),
    sigma.R = runif(1, 0, 100), sigma.r = runif(1, 0, 100))
  }
  LV1.sim <- bugs(data, inits, model.file = "C:/PhD/Thesis/LVDT_model.txt",
                  parameters=c("B", "mu.B", "sigma.R", "sigma.r"), n.iter=1500)
  l <- length(LV1.sim$sims.list$B[,1])
  LV <- c(LV, rep(j[w], l))
  sim.B <- data.frame(as.vector(1+LV1.sim$sims.list$B),
           as.vector(t(matrix(rep(i, l), nrow = p))), rep('Bayesian', l))
  myplots[[w]] <- ggplot(data = data.frame(i = as.vector(t(matrix(rep(i, dim(LV
1.sim$sims.list$B)[1]), nrow = p))),
y1ik = as.vector(LV1.sim$sims.list$B+1)), aes(x=i, y=y1ik)) +
           geom_boxplot(aes(fill = 'lightgreen')) +
           labs(y=expression(paste(y[ik]^(j),'[%]')), title=paste('j =', w)) +
           theme(axis.text=element_text(size=12),
           axis.title=element_text(size=18,face="bold"),
           legend.text=element_text(size=12),
           legend.title=element_text(size=12, face="bold"),
           axis.title.x=element_blank()) + geom_hline(yintercept=1,size=1.05) +

           geom_hline(yintercept = c(0.99, 1.01), size = 1.05, lty = 4) +
           geom_point(data = data.frame(x = as.character(LVDT), y = yijk),
                      aes(x=x, y=y, color = 'darkred'), size = 3) +
           scale_fill_identity(name = 'Boxplot', guide = 'legend',
                               labels = c('Bayesian')) +
           scale_colour_manual(name = 'Points', values =c('darkred'='darkred'),

                               labels = c('Raw data'))

  sim.mB <- c(sim.mB, LV1.sim$sims.list$mu.B+1)
  sim.sr <- c(sim.sr, LV1.sim$sims.list$sigma.r)
  sim.sR <- c(sim.sR, LV1.sim$sims.list$sigma.R)
  DIC <- c(DIC, LV1.sim$DIC)
  pD <- c(pD, LV1.sim$pD)
  Par[w, 1:2] <- c(LV1.sim$mean$mu.B+1, LV1.sim$sd$mu.B)
  Par[w, 3:4] <- c((LV1.sim$mean$sigma.r/LV1.sim$sd$sigma.r)^2,
                   LV1.sim$mean$sigma.r/LV1.sim$sd$sigma.r^2)
  Par[w, 5:6] <- c((LV1.sim$mean$sigma.R/LV1.sim$sd$sigma.R)^2,
                   LV1.sim$mean$sigma.R/LV1.sim$sd$sigma.R^2)
```

```
  Par[w, 7:8] <- c(LV1.sim$mean$sigma.r, LV1.sim$mean$sigma.R)
  #trueness:
  TP[w,1:2] <- round(quantile(LV1.sim$sims.list$mu.B-1, c(.025, .975)),3)
  #repeatability:
  TP[w,3] <- round(quantile(LV1.sim$sims.list$sigma.r, .95),4)
  #reproducibility
  TP[w,4] <- round(quantile(LV1.sim$sims.list$sigma.R, .95),4)
  #Equipments' bias
  B <- LV1.sim$mean$B
  detach(data1)
}
multiplot(plotlist = myplots[2:q], cols = 1)
gg_color_hue <- function(n) {
  hues = seq(15, 375, length = n + 1)
  hcl(h = hues, l = 65, c = 100)[1:n]
}
hue <- gg_color_hue(q)
ggplot(data = data.frame(mB = sim.mB,
                         Jump = rep(as.vector(t(matrix(rep(j, l), nrow=q))))),
       aes(mB, fill=Jump, colour=Jump)) + geom_density(adjust = 1, alpha=0.1) +

  labs(x = expression(paste(mu[B], ' [mm]'))) + xlim(0.990, 1.010) +
  stat_function(fun = dnorm, args = list(mean = Par[1,1], sd = Par[1,2]),
                col = hue[1], lwd = 1.2, lty = 4) +
  stat_function(fun = dnorm, args = list(mean = Par[2,1], sd = Par[2,2]),
                col = hue[2], lwd = 1.2, lty = 4) +
  stat_function(fun = dnorm, args = list(mean = Par[3,1], sd = Par[3,2]),
                col = hue[3], lwd = 1.2, lty = 4) +
  stat_function(fun = dnorm, args = list(mean = Par[4,1], sd = Par[4,2]),
                col = hue[4], lwd = 1.2, lty = 4) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=18,face="bold"),
        legend.text=element_text(size=12),legend.title=element_text(size=12)) +

  geom_hline(yintercept = 0, size = 1.05)
p1 <- ggplot(data = data.frame(sr = sim.sr,
                     Jump =rep(as.vector(t(matrix(rep(j, l), nrow = q))))),
             aes(sr, fill = Jump, colour = Jump)) +
  geom_density(adjust = 1, alpha = 0.1) +
  labs(x = expression(paste(sigma[r], ' [mm]'))) +
  stat_function(fun = dgamma, args = list(shape = Par[1,3], rate = Par[1,4]),
                col = hue[1], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[2,3], rate = Par[2,4]),
                col = hue[2], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[3,3], rate = Par[3,4]),
                col = hue[3], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[4,3], rate = Par[4,4]),
                col = hue[4], lwd = 1.2, lty = 4) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=18,face="bold"),
        legend.text=element_text(size=12),legend.title=element_text(size=12)) +
  ylim(0, 2250) + xlim(0, 0.012) + geom_hline(yintercept = 0, size = 1.05)
p2 <- ggplot(data = data.frame(sR = sim.sR,
                       Jump = rep(as.vector(t(matrix(rep(j, l), nrow = q))))),

             aes(sR, fill = Jump, colour = Jump)) +
  geom_density(adjust = 1, alpha = 0.1) +
  labs(x = expression(paste(sigma[H], ' [mm]'))) +
  stat_function(fun = dgamma, args = list(shape = Par[1,5], rate = Par[1,6]),
                col = hue[1], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[2,5], rate = Par[2,6]),
                col = hue[2], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[3,5], rate = Par[3,6]),
```

```
                col = hue[3], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[4,5], rate = Par[4,6]),
                col = hue[4], lwd = 1.2, lty = 4) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=18,face="bold"),
        legend.text=element_text(size=12),legend.title=element_text(size=12)) +
  ylim(0, 2250) + xlim(0, 0.012) + geom_hline(yintercept = 0, size = 1.05)
grid.arrange(p1,p2, ncol = 2)
TableHB <- data.frame(j =1:q, LV = j, pD =pD, DIC =DIC, mB =round(Par[,1],4),
                      sr = round(Par[,7],4), sH = round(Par[,8],4))
colnames(TableHB) <- c('j',paste0('Level','^','(j)'), paste0('pD','^','(j)'), p
aste0('DIC','^','(j)'), paste0('m[B]','^','(j)'), paste0('s[r]','^','(j)'), pas
te0('s[H]','^','(j)'))
grid.newpage()
grid.table(TableHB, theme = tt)

A <- data.frame(1:q, j, round(TableHB[, 5:7] - data.frame(mj, srj, sRj),5))
colnames(A) <- c('j', 'Level^(j)', paste0('Delta~m[B]','^','(j)'), paste0('Delt
a~s[r]','^','(j)'), paste0('Delta~s[H]','^','(j)'))
grid.newpage()
grid.table(A, theme = tt)

A <- data.frame(mj, srj, sRj)
colnames(A) <- c('m[B]', 's[r]', 's[H]')
A <- data.frame(t = c(as.vector(as.matrix(A)),
                      as.vector(as.matrix(TableHB[,5:7]))),
           theta = rep(as.vector(t(matrix(rep(colnames(A),q), ncol = q))),2),
           j = rep(j, 2*3), Model = c(rep("ANOVA", 3*q), rep("Bayesian", 3*q)))
A$theta <- ordered(A$theta, levels = c('m[B]', 's[r]', 's[H]'))
ggplot(A, aes(x = j)) +
  geom_point(aes(y =t, colour =Model), size =3, position =position_dodge(0.5))+

  scale_colour_manual(values = c("darkgreen", "darkred")) +
  facet_wrap(~theta, ncol = 3, scales="free_y", labeller = label_parsed) +
  labs(y = expression(E(theta)), x = "j", colour = "Model") +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=18,face="bold"),
        legend.text=element_text(size=12), legend.title=element_text(size=12),
        strip.text.x=element_text(size=15, color='darkorange', face="bold"))

colnames(TableHB)[dim(TableHB)[2]] <- paste0('s[L]','^','(j)')
grid.newpage()
grid.table(TableHB, theme = tt)
TableHB <- data.frame(TableHB,
                      round(sqrt(TableHB$`s[r]^(j)`^2+TableHB$`s[L]^(j)`^2),4))
colnames(TableHB) <- c('j',paste0('Level','^','(j)'),
                       paste0('p[D]','^','(j)'), paste0('DIC','^','(j)'),
                       paste0('m[B]','^','(j)'), paste0('s[r]','^','(j)'),
                       paste0('s[L]','^','(j)'), paste0('s[H]','^','(j)'))
grid.newpage()
grid.table(TableHB, theme = tt)

A <- data.frame(sLj, sRj)
colnames(A) <- c('s[L]', 's[H]')
A <- data.frame(t = c(as.vector(as.matrix(A)),
                      as.vector(as.matrix(TableHB[,7:8]))),
           theta = rep(as.vector(t(matrix(rep(colnames(A),q), ncol = q))),2),
           j = rep(j, 2*2), Model = c(rep("ANOVA", 2*q), rep("Bayesian", 2*q)))
A$theta <- ordered(A$theta, levels = c('s[L]', 's[H]'))
ggplot(A, aes(x = j)) +
  geom_point(aes(y =t, colour =Model), size =3, position =position_dodge(0.5))+

  scale_colour_manual(values = c("darkgreen", "darkred")) +
```

```
  facet_wrap(~theta, ncol = 3, scales="free_y", labeller = label_parsed) +
  labs(y = expression(E(theta)), x = "j", colour = "Model") +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=18,face="bold"),
        legend.text=element_text(size=12), legend.title=element_text(size=12),
        strip.text.x=element_text(size=15, color='darkorange', face="bold"))

### Hierarchical model with repeatability, homogeneity and inter-equipments
### parameters.
TableHB2 <- TableHB
sim.mB <- c()
sim.sr <- c()
sim.sL <- c()
sim.sR <- c()
Par<-matrix(nrow = q, ncol = 8)

for(w in 1:q){
 data1 <- data3[data3$Jump == j[w],c(1,4)]
 attach(data1)
 data <- list (p=p, n=n, yik=t(structure(.Data=yijk-1, .Dim=c(n,p))))
 inits <- function(){
   list(B = rnorm(p, 0, 100), mu.B = rnorm(1, 0, 1),
        sigma.L = runif(1, 0, 100), sigma.r = runif(1, 0, 100))
 }
 R1.sim <- bugs(data, inits, model.file = "C:/PhD/Thesis/LVDT_repr.txt",
                parameters = c("B", "mu.B", "sigma.L", "sigma.r", "sigma.R"),
                n.iter = 1500)
 TableHB2[w,3] <- R1.sim$pD
 TableHB2[w,4] <- R1.sim$DIC
 TableHB2[w,5] <- round(R1.sim$mean$mu.B+1 ,4)
 TableHB2[w,6] <- round(R1.sim$mean$sigma.r,4)
 TableHB2[w,7] <- round(R1.sim$mean$sigma.L,4)
 TableHB2[w,8] <- round(R1.sim$mean$sigma.R,4)
 sim.mB <- c(sim.mB, R1.sim$sims.list$mu.B+1)
 sim.sr <- c(sim.sr, R1.sim$sims.list$sigma.r)
 sim.sL <- c(sim.sL, R1.sim$sims.list$sigma.L)
 sim.sR <- c(sim.sR, R1.sim$sims.list$sigma.R)
 Par[w, 1:2] <- c(R1.sim$mean$mu.B+1, R1.sim$sd$mu.B)
 Par[w, 3:4] <- c((R1.sim$mean$sigma.r/R1.sim$sd$sigma.r)^2,
                  R1.sim$mean$sigma.r/R1.sim$sd$sigma.r^2)
 Par[w, 5:6] <- c((R1.sim$mean$sigma.L/R1.sim$sd$sigma.L)^2,
                  R1.sim$mean$sigma.L/R1.sim$sd$sigma.L^2)
 Par[w, 7:8] <- c((R1.sim$mean$sigma.R/R1.sim$sd$sigma.R)^2,
                  R1.sim$mean$sigma.R/R1.sim$sd$sigma.R^2)
 detach(data1)
 }
TableHB                        #Homogeneity evaluated out of the Bayesian model
#  j Level^(j) p[D]^(j) DIC^(j) m[B]^(j) s[r]^(j) s[L]^(j) s[H]^(j)
#1 1       3»2    18.09  -666.4   1.0017   0.0021   0.0052   0.0056
#2 2       4»3    16.12  -600.2   0.9987   0.0034   0.0039   0.0052
#3 3       5»4    18.31  -679.3   0.9994   0.0019   0.0052   0.0055
#4 4       6»5    17.36  -616.6   1.0003   0.0030   0.0046   0.0055
TableHB2                       #Homogeneity evaluated within the Bayesian model
#  j Level^(j) p[D]^(j) DIC^(j) m[B]^(j) s[r]^(j) s[L]^(j) s[H]^(j)
#1 1       3»2    18.09  -666.4   1.0017   0.0021   0.0052   0.0056
#2 2       4»3    16.12  -600.2   0.9987   0.0034   0.0039   0.0052
#3 3       5»4    18.31  -679.3   0.9994   0.0019   0.0052   0.0055
#4 4       6»5    17.36  -616.6   1.0003   0.0030   0.0046   0.0055
grid.newpage()
grid.table(TableHB2, theme = tt)
l <- length(R1.sim$sims.list$B[,1])

sim.sr <- data.frame(sim.sr, rep(as.vector(t(matrix(rep(j, l), nrow = q)))))
```

```
colnames(sim.sr) <- c('sr', 'Jump')
p1 <- ggplot(sim.sr, aes(sr, fill = Jump, colour = Jump)) +
  geom_density(adjust = 1, alpha = 0.1) +
  labs(x = expression(paste(sigma[r], ' [mm]'))) +
  stat_function(fun = dgamma, args = list(shape = Par[1,3], rate = Par[1,4]),
                col = hue[1], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[2,3], rate = Par[2,4]),
                col = hue[2], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[3,3], rate = Par[3,4]),
                col = hue[3], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[4,3], rate = Par[4,4]),
                col = hue[4], lwd = 1.2, lty = 4) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=18, face="bold"),
        legend.text=element_text(size=12), legend.title=element_text(size=12),
        legend.position = 'bottom') + ylim(0, 2250) + xlim(0, 0.012) +
  geom_hline(yintercept = 0, size = 1.05)
sim.sL <- data.frame(sim.sL, rep(as.vector(t(matrix(rep(j, l), nrow = q)))))
colnames(sim.sL) <- c('sL', 'Jump')
p2 <- ggplot(sim.sL, aes(sL, fill = Jump, colour = Jump)) +
  geom_density(adjust = 1, alpha = 0.1) +
  labs(x = expression(paste(sigma[L], ' [mm]'))) +
  stat_function(fun = dgamma, args = list(shape = Par[1,5], rate = Par[1,6]),
                col = hue[1], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[2,5], rate = Par[2,6]),
                col = hue[2], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[3,5], rate = Par[3,6]),
                col = hue[3], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[4,5], rate = Par[4,6]),
                col = hue[4], lwd = 1.2, lty = 4) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=18, face="bold"),
        legend.text=element_text(size=12), legend.title=element_text(size=12),
        legend.position = 'bottom') + ylim(0, 2250) + xlim(0, 0.012) +
  geom_hline(yintercept = 0, size = 1.05)
sim.sR <- data.frame(sim.sR, rep(as.vector(t(matrix(rep(j, l), nrow = q)))))
> colnames(sim.sR) <- c('sR', 'Jump')
p3 <- ggplot(sim.sR, aes(sR, fill = Jump, colour = Jump)) +
  geom_density(adjust = 1, alpha = 0.1) +
  labs(x = expression(paste(sigma[H], ' [mm]'))) +
  stat_function(fun = dgamma, args = list(shape = Par[1,7], rate = Par[1,8]),
                col = hue[1], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[2,7], rate = Par[2,8]),
                col = hue[2], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[3,7], rate = Par[3,8]),
                col = hue[3], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[4,7], rate = Par[4,8]),
                col = hue[4], lwd = 1.2, lty = 4) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=18, face="bold"),
        legend.text=element_text(size=12), legend.title=element_text(size=12),
        legend.position = 'bottom') + ylim(0, 2250) + xlim(0, 0.012) +
  geom_hline(yintercept = 0, size = 1.05)
grid.arrange(p1,p2,p3, ncol = 3)

#Accuracy_table
TableHU2 <- data.frame(j = 1:q, j, pD = TableHB2[,3], DIC = TableHB2[,4],
+                       d = format(round(t(apply(X = matrix(sim.mB-1, ncol = q),
 2, function(x) quantile(x,c(.025, .975),na.rm=T))),3), nsmall = 3),
                       Ur = format(round(apply(X = matrix(sim.sr$sr, ncol = q),
 2, function(x) quantile(x,.95,na.rm=T)),4), nsmall = 4),
                       Ur = format(round(apply(X = matrix(sim.sR$sR, ncol = q),
 2, function(x) quantile(x,.95,na.rm=T)),4), nsmall = 4))
```

```
colnames(TableHU2) <- c('j',paste0('Level','^','(j)'),
                        paste0('p[D]','^','(j)'), paste0('DIC','^','(j)'),
                        paste0('hat(delta)[lower]','^','(j)'),
                        paste0('hat(delta)[upper]','^','(j)'),
                        paste0('U[r]','^','(j)'), paste0('U[H]','^','(j)'))
grid.newpage()
grid.table(TableHU2, theme = tt)
### Proper prior
TableHB3 <- TableHB2
sim.mB <- c()
sim.sr <- c()
sim.sL <- c()
sim.sR <- c()
Par<-matrix(nrow = q, ncol = 8)
sim.B <- c()

for(w in 1:q){
  data1 <- data3[data3$Jump == j[w],c(1,4)]
  attach(data1)
  data <- list (p=p, n=n, yik=t(structure(.Data=yijk-1, .Dim=c(n,p))))
  inits <- function(){
    list(B = rnorm(p, 0, 100), mu.B = runif(1, -0.003, 0.003),
    sigma.L = runif(1, 0, .004), sigma.r = runif(1, 0, .006))
  }
  R1.sim <- bugs(data, inits, model.file = "C:/PhD/Thesis/LVDT_proper.txt",
                 parameters = c("B", "mu.B", "sigma.L", "sigma.r", "sigma.R"),
                 n.iter = 1500)
  TableHB3[w,3] <- R1.sim$pD
  TableHB3[w,4] <- R1.sim$DIC
  TableHB3[w,5] <- round(R1.sim$mean$mu.B+1, 4)
  TableHB3[w,6] <- round(R1.sim$mean$sigma.r,4)
  TableHB3[w,7] <- round(R1.sim$mean$sigma.L,4)
  TableHB3[w,8] <- round(R1.sim$mean$sigma.R,4)
  sim.mB <- c(sim.mB, R1.sim$sims.list$mu.B+1)
  sim.sr <- c(sim.sr, R1.sim$sims.list$sigma.r)
  sim.sL <- c(sim.sL, R1.sim$sims.list$sigma.L)
  sim.sR <- c(sim.sR, R1.sim$sims.list$sigma.R)
  Par[w, 1:2] <- c(R1.sim$mean$mu.B+1, R1.sim$sd$mu.B)
  sim.B <- c(sim.B, as.vector(R1.sim$sims.list$B))
  Par[w, 3:4] <- c((R1.sim$mean$sigma.r/R1.sim$sd$sigma.r)^2,
                   R1.sim$mean$sigma.r/R1.sim$sd$sigma.r^2)
  Par[w, 5:6] <- c((R1.sim$mean$sigma.L/R1.sim$sd$sigma.L)^2,
                   R1.sim$mean$sigma.L/R1.sim$sd$sigma.L^2)
  Par[w, 7:8] <- c((R1.sim$mean$sigma.R/R1.sim$sd$sigma.R)^2,
                   R1.sim$mean$sigma.R/R1.sim$sd$sigma.R^2)
  detach(data1)
+ }
TableHB2                                #improper prior
#  j Level^(j) p[D]^(j) DIC^(j) m[B]^(j) s[r]^(j) s[L]^(j) s[H]^(j)
#1 1       3»2    18.09  -666.4   1.0017   0.0021   0.0052   0.0056
#2 2       4»3    16.12  -600.2   0.9987   0.0034   0.0039   0.0052
#3 3       5»4    18.31  -679.3   0.9994   0.0019   0.0052   0.0055
#4 4       6»5    17.36  -616.6   1.0003   0.0030   0.0046   0.0055
TableHB3                                #proper prior
#  j Level^(j) p[D]^(j) DIC^(j) m[B]^(j) s[r]^(j) s[L]^(j) s[H]^(j)
#1 1       3»2    18.15  -665.9   1.0016   0.0021   0.0051   0.0055
#2 2       4»3    16.08  -600.2   0.9987   0.0034   0.0039   0.0052
#3 3       5»4    18.09  -679.8   0.9994   0.0019   0.0052   0.0055
#4 4       6»5    17.30  -616.7   1.0003   0.0030   0.0045   0.0054
ggplot(data = data.frame(y = c(sim.B+1, sim.mB),
theta = c(rep(as.vector(t(matrix(rep(i, l), nrow = p))), q),
              as.vector(t(matrix(rep(paste0('m[B',1:q,']'), l),q)))),
```

```
j = c(as.vector(t(matrix(rep(j, l*p),q))), as.vector(t(matrix(rep(j, l),q)))),

  aes(y, fill = theta, colour = theta)) + geom_density(adjust =1, alpha =0.1) +
  facet_wrap(~j) + labs(x = expression(paste(y, ' [mm]'))) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=18, face="bold"),
        legend.position = 'bottom') + geom_hline(yintercept = 0, size = 1.05)

l <- length(R1.sim$sims.list$B[,1])
p1 <- ggplot(data = data.frame(mB = sim.mB,
                      Jump =rep(as.vector(t(matrix(rep(j, l), nrow = q))))),
             aes(mB, fill = Jump, colour = Jump)) +
  geom_density(adjust = 1, alpha = 0.1) +
  labs(x=expression(paste(mu[B],' [mm]'))) + xlim(0.994,1.006) + ylim(0,2300) +
  stat_function(fun = dnorm, args = list(mean = Par[1,1], sd = Par[1,2]),
                col = hue[1], lwd = 1.2, lty = 4) +
  stat_function(fun = dnorm, args = list(mean = Par[2,1], sd = Par[2,2]),
                col = hue[2], lwd = 1.2, lty = 4) +
  stat_function(fun = dnorm, args = list(mean = Par[3,1], sd = Par[3,2]),
                col = hue[3], lwd = 1.2, lty = 4) +
  stat_function(fun = dnorm, args = list(mean = Par[4,1], sd = Par[4,2]),
                col = hue[4], lwd = 1.2, lty = 4) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=18, face="bold"),
        legend.position = 'none') + geom_hline(yintercept = 0, size = 1.05)
sim.sr <- data.frame(sim.sr, rep(as.vector(t(matrix(rep(j, l), nrow = q)))))
colnames(sim.sr) <- c('sr', 'Jump')
p2 <- ggplot(sim.sr, aes(sr, fill = Jump, colour = Jump)) +
  geom_density(adjust = 1, alpha = 0.1) +
  labs(x = expression(paste(sigma[r], ' [mm]'))) +
  stat_function(fun = dgamma, args = list(shape = Par[1,3], rate = Par[1,4]),
                col = hue[1], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[2,3], rate = Par[2,4]),
                col = hue[2], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[3,3], rate = Par[3,4]),
                col = hue[3], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[4,3], rate = Par[4,4]),
                col = hue[4], lwd = 1.2, lty = 4) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=18,face="bold"),
        legend.text=element_text(size=12), legend.title=element_text(size=12),
        axis.text.y = element_blank(), axis.title.y = element_blank(),
        axis.ticks.y = element_blank()) + ylim(0, 2300) + xlim(0, 0.012) +
  geom_hline(yintercept = 0, size = 1.05)
sim.sL <- data.frame(sim.sL, rep(as.vector(t(matrix(rep(j, l), nrow = q)))))
colnames(sim.sL) <- c('sL', 'Jump')
p3 <- ggplot(sim.sL, aes(sL, fill = Jump, colour = Jump)) +
  geom_density(adjust = 1, alpha = 0.1) +
  labs(x =expression(paste(sigma[L],' [mm]'))) + xlim(0,0.012) + ylim(0,2300) +

  stat_function(fun = dgamma, args = list(shape = Par[1,5], rate = Par[1,6]),
                col = hue[1], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[2,5], rate = Par[2,6]),
                col = hue[2], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[3,5], rate = Par[3,6]),
                col = hue[3], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[4,5], rate = Par[4,6]),
                col = hue[4], lwd = 1.2, lty = 4) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=18, face="bold"),
        legend.position = 'none') + geom_hline(yintercept = 0, size = 1.05)
sim.sR <- data.frame(sim.sR, rep(as.vector(t(matrix(rep(j, l), nrow = q)))))
colnames(sim.sR) <- c('sR', 'Jump')
```

```
p4 <- ggplot(sim.sR, aes(sR, fill = Jump, colour = Jump)) +
  geom_density(adjust = 1, alpha = 0.1) +
  labs(x = expression(paste(sigma[H], ' [mm]'))) +
  stat_function(fun = dgamma, args = list(shape = Par[1,7], rate = Par[1,8]),
                col = hue[1], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[2,7], rate = Par[2,8]),
                col = hue[2], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[3,7], rate = Par[3,8]),
                col = hue[3], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[4,7], rate = Par[4,8]),
                col = hue[4], lwd = 1.2, lty = 4) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=18,face="bold"),
        legend.text=element_text(size=12), legend.title=element_text(size=12),
        axis.text.y = element_blank(), axis.title.y = element_blank(),
        axis.ticks.y = element_blank()) + ylim(0, 2300) + xlim(0, 0.012) +
  geom_hline(yintercept = 0, size = 1.05)
grid.arrange(p1,p2,p3,p4, ncol = 2)

#Accuracy description of the whole product line
p1 <- ggplot(data = data.frame(mB = sim.mB-1,
             Jump =rep(as.vector(t(matrix(rep(j, 1), nrow = q))))),
             aes(mB, fill = Jump, colour = Jump)) +
  geom_density(adjust = 1, alpha = 0.1) +
  labs(x =expression(paste(delta,' [mm]'))) +xlim(-0.006,0.006) +ylim(0,2300) +
  stat_function(fun = dnorm, args = list(mean = Par[1,1]-1, sd = Par[1,2]),
                col = hue[1], lwd = 1.2, lty = 4) +
  stat_function(fun = dnorm, args = list(mean = Par[2,1]-1, sd = Par[2,2]),
                col = hue[2], lwd = 1.2, lty = 4) +
  stat_function(fun = dnorm, args = list(mean = Par[3,1]-1, sd = Par[3,2]),
                col = hue[3], lwd = 1.2, lty = 4) +
  stat_function(fun = dnorm, args = list(mean = Par[4,1]-1, sd = Par[4,2]),
                col = hue[4], lwd = 1.2, lty = 4) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=18, face="bold"),
        legend.position = 'none') + geom_hline(yintercept = 0, size = 1.05)
p2 <- ggplot(sim.sr, aes(sr, fill = Jump, colour = Jump)) +
  geom_density(adjust = 1, alpha = 0.1) +
  labs(x = expression(paste(sigma[r], ' [mm]'))) +
  stat_function(fun = dgamma, args = list(shape = Par[1,3], rate = Par[1,4]),
                col = hue[1], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[2,3], rate = Par[2,4]),
                col = hue[2], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[3,3], rate = Par[3,4]),
                col = hue[3], lwd = 1.2, lty = 4) +
  stat_function(fun = dgamma, args = list(shape = Par[4,3], rate = Par[4,4]),
                col = hue[4], lwd = 1.2, lty = 4) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=18,face="bold"),
        legend.position = 'none', axis.text.y = element_blank(),
        axis.title.y = element_blank(), axis.ticks.y = element_blank()) +
  ylim(0, 2300) + xlim(0, 0.012) + geom_hline(yintercept = 0, size = 1.05)
grid.arrange(p1,p2,p4, ncol = 3)

#Accuracy_table
TableHU3 <- data.frame(j = 1:q, j, pD = TableHB2[,3], DIC = TableHB2[,4],
                       d = format(round(t(apply(X = matrix(sim.mB-1, ncol = q),
 2, function(x) quantile(x,c(.025, .975),na.rm=T))),3), nsmall = 3),
                       Ur = format(round(apply(X = matrix(sim.sr$sr, ncol = q),
 2, function(x) quantile(x,.95,na.rm=T)),4), nsmall = 4),
                       Ur = format(round(apply(X = matrix(sim.sR$sR, ncol = q),
 2, function(x) quantile(x,.95,na.rm=T)),4), nsmall = 4))
colnames(TableHU3) <- colnames(TableHU2)
```

```
grid.newpage()
grid.table(TableHU3, theme = tt)
```

```
### 3 levels hierarchical model

data1 <- data3[ ,c(1,2,4)]
data1 <- data1[order(data1$Jump, data1$LVDT),]
attach(data1)
data <- list (p=p, q=q, n=n, xkij = structure(.Data = yijk, .Dim=c(n,p,q)))
data
#$p
#[1] 18
#
#$q
#[1] 4
#
#$n
#[1] 4
#
#$xkij
#, , 1
#
#      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  …  [,14] [,15] [,16] [,17] [,18]
#[1,] 1.000 1.003 0.998 1.005 1.004 0.998  …  0.999 1.009 1.001 0.995 0.995
#[2,] 1.003 0.996 0.999 1.005 1.003 0.999  …  0.998 1.009 1.001 0.995 0.995
#[3,] 1.002 0.999 0.999 1.009 1.004 1.003  …  0.999 1.009 1.003 1.002 0.996
#[4,] 1.002 0.997 1.001 1.008 1.004 0.999  …  0.998 1.009 1.003 0.999 0.998
#
#, , 2
#
#      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  …  [,14] [,15] [,16] [,17] [,18]
#[1,] 1.001 0.994 0.999 0.997 0.994 0.999  …  0.992 1.005 1.001 0.999 0.991
#[2,] 1.001 1.001 1.000 0.992 0.994 0.999  …  1.003 1.006 1.004 1.000 1.005
#[3,] 1.002 0.998 1.000 0.999 0.996 1.003  …  1.001 1.006 1.001 1.000 0.989
#[4,] 1.004 1.001 1.000 0.994 0.996 1.000  …  1.000 1.006 1.004 1.003 0.990
#
#, , 3
#
#      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  …  [,14] [,15] [,16] [,17] [,18]
#[1,] 0.996 1.001 0.995 0.996 0.993 0.995  …  0.992 0.991 1.000 0.999 1.006
#[2,] 0.997 1.004 0.999 0.998 0.993 0.998  …  0.992 0.992 0.992 1.000 1.006
#[3,] 0.996 1.003 0.999 0.996 0.993 0.998  …  0.990 0.994 0.996 0.997 1.008
#[4,] 0.996 1.006 0.999 0.996 0.990 0.996  …  0.990 0.993 0.995 0.997 1.006
#
#, , 4
#
#      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  …  [,14] [,15] [,16] [,17] [,18]
#[1,] 1.003 0.995 0.999 1.005 1.003 1.001  …  1.001 0.989 0.992 1.001 1.000
#[2,] 1.004 1.003 0.999 1.007 1.001 1.000  …  1.003 0.990 0.994 1.000 1.000
#[3,] 1.003 1.001 1.000 1.005 1.004 1.003  …  0.998 0.994 0.991 1.003 1.001
#[4,] 1.003 0.994 1.000 1.011 1.004 1.001  …  0.995 0.992 0.994 1.006 1.001

# Testing the data input with just one parameter for the reference value
inits <- function(){
  list(B = rnorm(p, 0, .01), mu.B = rnorm(1, 0, .01),
       sigma.R = runif(1, 0, 10), sigma.r = runif(1, 0, 10))
}
LV.sim <- bugs(data, inits, model.file = "C:/PhD/Thesis/LVDT_full1.txt",
               parameters = c("B", "mu.B", "sigma.R","sigma.r"), n.iter = 1500)
LV.sim$mean$mu.B                  #same results
#[1] 1.001711
LV.sim$mean$sigma.r               #same result
#[1] 0.002105952
#LV.sim$mean$sigma.R               #same result
[1] 0.005169028
```

```
# Full model

data <- list (p=p, q=q, n=n, xkij = structure(.Data = yijk-1, .Dim=c(n,p,q)))
inits <- function(){
  list(B = matrix(rnorm(p*q, 0, 1.0E-6), nrow = p, ncol = q),
       mu.B = rnorm(q, 0, 1.0E-6), m = runif(1, -.003, 0.003),
       sigma.L = runif(1, 0, 0.006), sigma.r = runif(1, 0, 0.004))
 }
FB.sim <- bugs(data, inits, model.file = "C:/PhD/Thesis/LVDT_full.txt",
               parameters = c("B","mu.B","m","sigma.r","sigma.L","sigma.R"),
               n.iter = 1000)
mBj <- round(FB.sim$mean$mu.B+1,4)
rownames(mBj) <- paste0('m[B]','[',1:4,']')
grid.newpage()
grid.table(t(mBj), theme = tt)

l <- length(FB.sim$sims.list$m)
l <- data.frame(theta =c(FB.sim$sims.list$m+1,
                         as.vector(FB.sim$sims.list$mu.B+1)),
                Parameters =c(rep('m[P]', l),
                         as.vector(t(matrix(rep(paste0('m[B',1:4,']'),l),q)))))
ggplot(data = l, aes(theta, fill = Parameters, colour = Parameters)) +
  geom_density(adjust = 1, alpha = 0.3) + labs(x = expression(theta)) +
  theme(axis.text=element_text(size=15),
        axis.title=element_text(size=18, face="bold"),
        legend.text=element_text(size=18), legend.position = 'bottom') +
  scale_colour_discrete(name = "",
                        labels = c(expression(mu[B1]), expression(mu[B2]),
                                   expression(mu[B3]), expression(mu[B4]),
                                   expression(mu[P]))) +
  scale_fill_discrete(name = "",
                      labels = c(expression(mu[B1]), expression(mu[B2]),
                                 expression(mu[B3]), expression(mu[B4]),
                                 expression(mu[P]))) +
  stat_function(fun = dnorm, args = list(mean = FB.sim$mean$m+1,
     sd = FB.sim$sd$m), col = gg_color_hue(q+1)[5], lwd = 1.2, lty = 4) +
  stat_function(fun = dnorm, args = list(mean = FB.sim$mean$mu.B[1]+1,
     sd = FB.sim$sd$mu.B[1]), col = gg_color_hue(q+1)[1], lwd = 1.2, lty = 4) +
  stat_function(fun = dnorm, args = list(mean = FB.sim$mean$mu.B[2]+1,
     sd = FB.sim$sd$mu.B[2]), col = gg_color_hue(q+1)[2], lwd = 1.2, lty = 4) +

  stat_function(fun = dnorm, args = list(mean = FB.sim$mean$mu.B[3]+1,
     sd = FB.sim$sd$mu.B[3]), col = gg_color_hue(q+1)[3], lwd = 1.2, lty = 4) +

  stat_function(fun = dnorm, args = list(mean = FB.sim$mean$mu.B[4]+1,
     sd = FB.sim$sd$mu.B[4]), col = gg_color_hue(q+1)[4], lwd = 1.2, lty = 4) +
  geom_hline(yintercept = 0, lty = 4, size = 1.05) + xlim(.994, 1.006)
p1 <- ggplot(data = data.frame(delta = FB.sim$sims.list$m), aes(delta)) +
  geom_density(adjust = 1, alpha = 0.3, fill = hue[1]) + ylim(0, 3000) +
  labs(x = expression(paste(delta, ' [mm]'))) + xlim(-0.0075, 0.0075) +
  stat_function(fun = dnorm, args = list(mean = FB.sim$mean$m,
                sd = FB.sim$sd$m), col = hue[1], lwd = 1.2, lty = 4) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=18,face="bold")) +
  geom_hline(yintercept = 0, size = 1.05)
a <- (FB.sim$mean$sigma.r/FB.sim$sd$sigma.r)^2
b <- FB.sim$mean$sigma.r/FB.sim$sd$sigma.r^2
p2 <- ggplot(data = data.frame(s.r = FB.sim$sims.list$sigma.r), aes(s.r)) +
  geom_density(adjust = 1, alpha = 0.3, fill = hue[2]) + ylim(0, 3000) +
  labs(x = expression(paste(sigma[r], ' [mm]'))) + xlim(0, 0.015) +
  stat_function(fun = dgamma, args = list(shape = a, rate = b),
                col = hue[2], lwd = 1.2, lty = 4) +
  theme(axis.text=element_text(size=12),
```

```
        axis.title=element_text(size=18,face="bold")) +
  geom_hline(yintercept = 0, size = 1.05)
a <- (FB.sim$mean$sigma.R/FB.sim$sd$sigma.R)^2
b <- FB.sim$mean$sigma.R/FB.sim$sd$sigma.R^2
p3 <- ggplot(data = data.frame(s.H = FB.sim$sims.list$sigma.R), aes(s.H)) +
+   geom_density(adjust = 1, alpha = 0.3, fill = hue[3]) + ylim(0, 3000) +
  labs(x = expression(paste(sigma[H], ' [mm]'))) + xlim(0, 0.015) +
  stat_function(fun = dgamma, args = list(shape = a, rate = b),
                col = hue[3], lwd = 1.2, lty = 4) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=18,face="bold")) +
  geom_hline(yintercept = 0, size = 1.05)
grid.arrange(p1,p2,p3, ncol = 3)

#Accuracy_table
TableFB <- data.frame(pD = FB.sim$pD, DIC = FB.sim$DIC,
 d = format(round(t(quantile(FB.sim$sims.list$m, c(.025, .975))),4), nsmall=4),
 Ur = format(round(quantile(FB.sim$sims.list$sigma.r, .95), 4), nsmall = 4),
 UH = format(round(quantile(FB.sim$sims.list$sigma.R, .95), 4), nsmall = 4))
colnames(TableFB) <- c('p[D]', 'DIC', 'hat(delta)[lower]', 'hat(delta)[upper]',

                       'U[r]', 'U[H]')
grid.newpage()
grid.table(TableFB, theme = tt)

#Empirical Cumulative Distribution Function
P = ecdf(FB.sim$sims.list$m)
l <- round(P(0.0039)-P(-0.0039), 3)
P = ecdf(FB.sim$sims.list$sigma.R)
l <- c(l, round(P(0.006), 3))
l <- data.frame(A = c('Trueness', 'Precision homogeneity'),
                theta = paste(c('±0.004', '<0.006'), 'mm'),
                C.P = paste(100*l, '%'))
colnames(l) <- c('Accuracy', 'hat(theta)', 'C.P.')
grid.newpage()
grid.table(l, theme = tt)
```

9 ANNEX D: R SCRIPT FOR THE PROFICIENCY TESTING

```
library(R2OpenBUGS)
library(coda)
library(ggplot2)
library(ggrepel)
library(grid)
library(gridExtra)

data3<-read.table("LVDT.txt",header = TRUE, dec = ",")
data3 <- data3[-which(data3$LVDT==data3$LVDT[length(data3$LVDT)]),]

str(data3)
data3 <- data3[order(data3$LVDT, data3$Jump),]

i<-levels(as.factor(data3$LVDT))
print(p<-length(i))       #nr. of "Laboratories" = nr. of LVDTs
j<-levels(data3$Jump)
k<-levels(data3$round)
print(n<-length(k))       #nr. of replicates

proficiency <- rbind(data3[c(data3$Jump==j[1]) ,c(1,2,4)], data3[c(data3$Jump==
j[2]) ,c(1,2,4)])
proficiency <- as.matrix(proficiency[order(proficiency$Jump, proficiency$LVDT),
])
rm(data3)
proficiency <- data.frame(proficiency, stringsAsFactors = FALSE)
proficiency$yijk <- as.numeric(proficiency$yijk)
proficiency$LVDT <- as.numeric(proficiency$LVDT)

j<-levels(as.factor(proficiency$Jump))
print(q<-length(j))       #nr. of testing levels



### Drawn out data of 2nd brand from a population spreading similarly than the
Instron one
set.seed(1)
tm1  <- 1.000
tsL1 <-  .005
tsr1 <-  .003
rmB1 <- round(rnorm(q, tm1, tsL1),4)
print(B1 <- c(round(rnorm(p, rmB1[1], tsL1), 3), round(rnorm(p, rmB1[2], tsL1),
 3)))
print(yijkz <- matrix(t(matrix(rep(B1, n), ncol = n)), ncol = 1))
t <- p*q*n
for(w in 1:t){
  yijkz[w] <- rnorm(1, yijkz[w], tsr1)
}
print(yijk <- c(proficiency$yijk, round(yijkz, 3)))

tm2  <- 1.006
tsL2 <-  .004
tsr2 <-  tsr1
rmB2 <- round(rnorm(q, tm2, tsL2),4)
print(B2 <- c(round(rnorm(p, rmB2[1], tsL2), 3), round(rnorm(p, rmB2[2], tsL2),
 3)))
print(yijkz <- matrix(t(matrix(rep(B2, n), ncol = n)), ncol = 1))
for(w in 1:t){
  yijkz[w] <- rnorm(1, yijkz[w], tsr2)
}
print(yijk <- c(yijk, round(yijkz, 3)))
```

```
print(LVDT <- max(proficiency$LVDT)+1)
LVDT <- matrix(t(matrix(rep(LVDT:(LVDT+p-1), n), ncol = n)), ncol = 1)
LVDT <- rep(LVDT, q)
print(L <- max(LVDT)+1)
L <- matrix(t(matrix(rep(L:(L+p-1), n), ncol = n)), ncol = 1)
L <- rep(L, q)
print(LVDT <- c(proficiency$LVDT, LVDT, L))
rm(L)
Jump <- rep(c(rep(j[1], n*p), rep(j[2], n*p)), q)
print(Jump <- c(proficiency$Jump, Jump))
Brand <- c(rep("Instron", p*q*n), rep("Best", p*q*n), rep("Biased", p*q*n))
str(proficiency <- data.frame(yijk, LVDT, Jump, Brand))
proficiency$LVDT <- as.factor(proficiency$LVDT)
str(proficiency)

i <- levels(proficiency$LVDT)
z <- levels(proficiency$Brand)
print(r <- length(z))

data<-proficiency[proficiency$Jump==j[1],]
ggplot(data=data, aes(yijk)) + geom_histogram(
  bins = round(sqrt(length(data$yijk)),0), aes(x=yijk, y=..density..),
  alpha = 0.5) + geom_density(aes(x=yijk, y=..density..), adjust = 1,
  alpha = 0, size = 1.05) + theme(axis.text = element_text(size=15),  title = e
lement_text(size=24),
  axis.title=element_text(size=18,face="bold"), legend.text=element_text(size=1
5), legend.title=element_text(size=15)
  ) + labs(x =expression(x[kiz]), title = paste0('Histogram (level: ', j[1],')'
)
  ) + xlim(.98, 1.02)
ggplot(data=data, aes(yijk, colour = Brand, fill = Brand)) + geom_histogram(
  bins = round(sqrt(length(yijk)/r),0), aes(x=yijk, y=..density..), alpha = 0.3
  ) + geom_density(aes(x=yijk, y=..density.., colour = Brand), alpha = 0, size
= 1.05
  ) + theme(axis.text = element_text(size=15), axis.title=element_text(size=18,
face="bold"
  ), legend.text=element_text(size=15), legend.title=element_text(size=15),  ti
tle = element_text(size=15)
  ) + labs(x =expression(x[kiz]), title = paste0('Histogram (level: ', j[1],')'
)) + xlim(.98, 1.02)
y <- qnorm(c(0.25, 0.75), mean = mean(data$yijk), sd = sd(data$yijk))
x <- qnorm(c(0.25, 0.75))
slope <- diff(y)/diff(x)
int <- y[1] - slope * x[1]
#dev.off()
ggplot(data) +stat_qq(aes(sample = yijk)) + geom_abline(slope = slope,
  intercept = int, lty = 4, size = 1.05) + theme(axis.text = element_text(
  size=15), axis.title=element_text(size=18,face="bold"),  title = element_text
(size=15)
  ) + labs(y =expression(x[kiz]), title = paste0('QQ-plot (level: ', j[1],')'))
 + ylim(.98, 1.02)
g <- ggplot(data) + stat_qq(aes(sample = yijk, colour = Brand)) + theme(
  axis.text = element_text(size=15),  title = element_text(size=15), axis.title
=element_text(size=18,face="bold"
  ), legend.text=element_text(size=15), legend.title=element_text(size=15)) + l
abs(
  y =expression(x[kiz]), title = paste0('QQ-plot (level: ', j[1],')')) + ylim(.
98, 1.02)
gg_color_hue <- function(n) {
  hues = seq(15, 375, length = n + 1)
  hcl(h = hues, l = 65, c = 100)[1:n]
}
```

```
for(w in 1:r){
  y <- qnorm(c(0.25, 0.75), mean = mean(data$yijk[data$Brand==z[w]]), sd = sd(d
ata$yijk[data$Brand==z[w]]))
  slope <- diff(y)/diff(x)
  int <- y[1] - slope * x[1]
  g <- g + geom_abline(slope = slope, intercept = int, colour = gg_color_hue(r)
[w])
}
g
rm(x, y, slope, int)

data<-proficiency[proficiency$Jump==j[2],]
ggplot(data=data, aes(yijk)) + geom_histogram(
  bins = round(sqrt(length(data$yijk)),0), aes(x=yijk, y=..density..), alpha =
0.5
  ) + geom_density(aes(x=yijk, y=..density..), adjust = 1, alpha = 0, size = 1.
05
  ) + theme(axis.text = element_text(size=15),  title = element_text(size=15),
  axis.title=element_text(size=18,face="bold"), legend.text=element_text(size=1
5), legend.title=element_text(size=15)
  ) + labs(x =expression(x[kiz]), title = paste0('Histogram (level: ', j[2],')'
)
  ) + xlim(.98, 1.02)
ggplot(data=data, aes(yijk, colour = Brand, fill = Brand)) + geom_histogram(
  bins = round(sqrt(length(yijk)/r),0), aes(x=yijk, y=..density..), alpha = 0.3
  ) + geom_density(aes(x=yijk, y=..density.., colour = Brand), alpha = 0, size
= 1.05
  ) + theme(axis.text = element_text(size=15), axis.title=element_text(size=18,
face="bold"
  ), legend.text=element_text(size=15), legend.title=element_text(size=15),  ti
tle = element_text(size=15)
  ) + labs(x =expression(x[kiz]), title = paste0('Histogram (level: ', j[2],')'
)) + xlim(.98, 1.02)
y <- qnorm(c(0.25, 0.75), mean = mean(data$yijk), sd = sd(data$yijk))
x <- qnorm(c(0.25, 0.75))
slope <- diff(y)/diff(x)
int <- y[1] - slope * x[1]
#dev.off()
ggplot(data) +stat_qq(aes(sample = yijk)) + geom_abline(slope = slope,
 intercept = int, lty = 4, size = 1.05) + theme(axis.text = element_text(
 size=15), axis.title=element_text(size=18,face="bold"),  title = element_text(
size=15)
 ) + labs(y =expression(x[kiz]), title = paste0('QQ-plot (level: ', j[2],')'))
+ ylim(.98, 1.02)
g <- ggplot(data) + stat_qq(aes(sample = yijk, colour = Brand)) + theme(
  axis.text = element_text(size=15),  title = element_text(size=15), axis.title
=element_text(size=18,face="bold"
  ), legend.text=element_text(size=15), legend.title=element_text(size=15)) + l
abs(
  y =expression(x[kiz]), title = paste0('QQ-plot (level: ', j[2],')')) + ylim(.
98, 1.02)
for(w in 1:r){
  y <- qnorm(c(0.25, 0.75), mean = mean(data$yijk[data$Brand==z[w]]), sd = sd(d
ata$yijk[data$Brand==z[w]]))
  slope <- diff(y)/diff(x)
  int <- y[1] - slope * x[1]
  g <- g + geom_abline(slope = slope, intercept = int, colour = gg_color_hue(r)
[w])
}
g
rm(x, y, slope, int)
```

```
### Proficiency ISO 13528

A <- proficiency$LVDT[Jump==j[1]]
AOV1 <- aov(proficiency$yijk[Jump==j[1]]~A, contrasts = list(A = "contr.sum"))
print(m1 <- AOV1$coefficients[1])
print(s2d1 <- summary(AOV1)[[1]]$'Mean Sq'[1])
print(s2r1 <- summary(AOV1)[[1]]$'Mean Sq'[2])
print(p1 <- summary(AOV1)[[1]]$'Df'[1]+1)
print(n1 <- summary(AOV1)[[1]]$'Df'[2]/p1+1)
print(sL1 <- round(sqrt((s2d1-s2r1)/n1),3))
A <- proficiency$LVDT[Jump==j[2]]
AOV2 <- aov(proficiency$yijk[Jump==j[2]]~A, contrasts = list(A = "contr.sum"))
print(m2 <- AOV2$coefficients[1])
print(s2d2 <- summary(AOV2)[[1]]$'Mean Sq'[1])
print(s2r2 <- summary(AOV2)[[1]]$'Mean Sq'[2])
print(p2 <- summary(AOV2)[[1]]$'Df'[1]+1)
print(n2 <- summary(AOV2)[[1]]$'Df'[2]/p2+1)
print(sL2 <- round(sqrt((s2d2-s2r2)/n2),3))

y1 <- AOV1$coefficients+AOV1$coefficients[1]
print(y1 <- as.vector(c(y1[2:(r*p)], -sum(AOV1$coefficients[2:length(AOV1$coeff
icients)])+AOV1$coefficients[1])))
### Proof
y1 - as.vector(tapply(proficiency$yijk, list(proficiency$LVDT, proficiency$Jump
), mean))[1:(r*p)]
sp1 <- tsL1
z1 <- (y1 - 1)/sp1

y2 <- AOV2$coefficients+AOV2$coefficients[1]
print(y2 <- as.vector(c(y2[2:(r*p)], -sum(AOV2$coefficients[2:length(AOV2$coeff
icients)])+AOV2$coefficients[1])))
### Proof
y2 - as.vector(tapply(proficiency$yijk, list(proficiency$LVDT, proficiency$Jump
), mean))[(r*p+1):(2*r*p)]
sp2 <- tsL1
z2 <- (y2 - 1)/sp2

Z<-data.frame(z1, z2, i, Brand[c(1:p,((q*n*p)+1):((q*n+1)*p),((2*q*n*p)+1):((2*
q*n+1)*p))])
colnames(Z)[4] <- "Brand"
gg_circle <- function(r, xc, yc, color, type, fill=NA, ...) {
  x <- xc + r*cos(seq(0, pi, length.out=100))
  ymax <- yc + r*sin(seq(0, pi, length.out=100))
  ymin <- yc + r*sin(seq(0, -pi, length.out=100))
  annotate("ribbon", x=x, ymin=ymin, ymax=ymax, size = 1.2, color=color, linety
pe=type, fill=fill, ...)
}
ggplot(Z, aes(x = z1, y = z2)) + geom_point(mapping = aes(color = Brand),
  shape = 1, size = 4.5, stroke = 2) + geom_text_repel(aes(label=i), size = 6
  ) + gg_circle(r=2, xc=0, yc=0, color='gray30', type = 'dashed'
  ) + gg_circle(r=3, xc=0, yc=0, color='gray30', type = 'solid'
  ) + theme(axis.text = element_text(size=18), axis.title=element_text(size=21,
face="bold"),
  legend.text=element_text(size=18), legend.title=element_text(size=18)
  ) + labs(y = expression(z^("2")), x =expression(z^("1"))) + geom_hline(
  yintercept = 0, linetype="dashed") + geom_vline(xintercept = 0, linetype="das
hed")  # + scale_color_brewer(palette="Set1")

sigma1 <- round(as.vector(tapply(proficiency$yijk, list(proficiency$LVDT, profi
ciency$Jump), sd))/sqrt(n-1),4)
sigma2 <- sigma1[(p*r+1):(p*r*q)]
sigma1 <- sigma1[1:(p*r)]
X <- Z
```

```
X$z1 <- X$z1*sp1+1
X$z2 <- X$z2*sp2+1
X <- data.frame(X, sigma1, sigma2)
colnames(X)[1:2]<-c('x1','x2')

ggplot(X, aes(i, x1, colour=Brand)) + geom_errorbar(aes(ymin = x1-sigma1, ymax
= x1+sigma1), size = 1.05) + theme(title=element_text(size=18,face="bold"),
  axis.text.x = element_text(angle = 90, hjust = 1), axis.text = element_text(s
ize=18), axis.title=element_text(size=18,face="bold"),
  legend.text=element_text(size=18), legend.title=element_text(size=18)) + labs
(
  y = expression(x[i]^("1")), x = expression(LVDT[i]), title = paste0('level: '
, j[1])
  ) + geom_point(size=3, shape=21, fill="white", stroke = 1.5) + geom_hline(
  yintercept = c(1-2*sp1, 1, 1+2*sp1), color='gray30', lty = 4, size = 1.05) +
geom_hline(
  yintercept = c(1-3*sp1, 1+3*sp1), color='gray30', size = 1.05) + scale_y_cont
inuous(limits = c(.98, 1.02))
ggplot(X, aes(i, x2, colour=Brand)) + geom_errorbar(aes(ymin = x2-sigma2, ymax
= x2+sigma2), size = 1.05) + theme(title=element_text(size=18,face="bold"),
  axis.text.x = element_text(angle = 90, hjust = 1), axis.text = element_text(s
ize=18), axis.title=element_text(size=18,face="bold"),
  legend.text=element_text(size=18), legend.title=element_text(size=18)) + labs
(
  y = expression(x[i]^("2")), x = expression(LVDT[i]), title = paste0('level: '
, j[2])
  ) + geom_point(size=3, shape=21, fill="white", stroke = 1.5) + geom_hline(
  yintercept = c(1-2*sp1, 1, 1+2*sp1), color='gray30', lty = 4, size = 1.05) +
geom_hline(
  yintercept = c(1-3*sp1, 1+3*sp1), color='gray30', size = 1.05) + scale_y_cont
inuous(limits = c(.98, 1.02))



### Proficiency Hierarchical Bayesian model
#   2-hierarchical levels model
t=p*r
print(data <- list (t=t, n=n, xki = structure(.Data = -1+proficiency$yijk[profi
ciency$Jump==j[1]], .Dim=c(n,t))))
inits <- function(){
  list(B = rnorm(t, 0, .003), mu = rnorm(1, 0, .003), sigma.L = runif(1, 0, 0.0
05), sigma.r = runif(1, 0, 0.003))
}
Pr1.sim <- bugs(data, inits, model.file = "D:/Giuse/Istruzione/Università/Dotto
rato di ricerca - Matematica/Progetti/05 PhD/LVDT_proficiency-2.txt",
                parameters = c("B", "mu", "sigma.r", "sigma.L", "sigma.R"), n.i
ter = 1000)
rm(data)
print(Pr1.sim, digits.summary = 3)

print(data <- list (t=t, n=n, xki = structure(.Data = -1+proficiency$yijk[profi
ciency$Jump==j[2]], .Dim=c(n,t))))
Pr2.sim <- bugs(data, inits, model.file = "D:/Giuse/Istruzione/Università/Dotto
rato di ricerca - Matematica/Progetti/05 PhD/LVDT_proficiency-2.txt",
                parameters = c("B", "mu", "sigma.r", "sigma.L", "sigma.R"), n.i
ter = 1000)
rm(data)
print(Pr2.sim, digits.summary = 3)

X$x1 <- round(Pr1.sim$mean$B, 4)+1
X$x2 <- round(Pr2.sim$mean$B, 4)+1
X$sigma1 <- round(Pr1.sim$sd$B, 4)
X$sigma2 <- round(Pr2.sim$sd$B, 4)
```

```
ggplot(X, aes(i, x1, colour=Brand)) + geom_errorbar(aes(ymin = x1-sigma1, ymax
= x1+sigma1), size = 1.05) + theme(title=element_text(size=18,face="bold"),
  axis.text.x = element_text(angle = 90, hjust = 1), axis.text = element_text(s
ize=18), axis.title=element_text(size=24,face="bold"),
  legend.text=element_text(size=18), legend.title=element_text(size=18)) + labs
(
  y = expression(x["·i"]^("1")), x = expression(LVDT[i]), title = paste0('level
: ', j[1])
  ) + geom_point(size=3, shape=21, fill="white", stroke = 1.5) + geom_hline(
  yintercept = c(1-2*sp1, 1, 1+2*sp1), color='gray30', lty = 4, size = 1.05) +
geom_hline(
  yintercept = c(1-3*sp1, 1+3*sp1), color='gray30', size = 1.05) + scale_y_cont
inuous(limits = c(.98, 1.02))
ggplot(X, aes(i, x2, colour=Brand)) + geom_errorbar(aes(ymin = x2-sigma2, ymax
= x2+sigma2), size = 1.05) + theme(title=element_text(size=18,face="bold"),
  axis.text.x = element_text(angle = 90, hjust = 1), axis.text = element_text(s
ize=18), axis.title=element_text(size=24,face="bold"),
  legend.text=element_text(size=18), legend.title=element_text(size=18)) + labs
(
  y = expression(x["·i"]^("2")), x = expression(LVDT[i]), title = paste0('level
: ', j[2])
  ) + geom_point(size=3, shape=21, fill="white", stroke = 1.5) + geom_hline(
  yintercept = c(1-2*sp1, 1, 1+2*sp1), color='gray30', lty = 4, size = 1.05) +
geom_hline(
  yintercept = c(1-3*sp1, 1+3*sp1), color='gray30', size = 1.05) + scale_y_cont
inuous(limits = c(.98, 1.02))

Z$z1 <- (X$x1-1)/sp1
Z$z2 <- (X$x2-1)/sp2
ggplot(Z, aes(x = z1, y = z2)) + geom_point(mapping = aes(color = Brand),
  shape = 1, size = 4.5, stroke = 2) + geom_text_repel(aes(label=i), size = 6)
+
  gg_circle(r=2, xc=0, yc=0, color='gray30', type = 'dashed') +
  gg_circle(r=3, xc=0, yc=0, color='gray30', type = 'solid') +
  theme(title=element_text(size=18,face="bold"), axis.text = element_text(size=
18),
  axis.title=element_text(size=21,face="bold"), legend.text=element_text(size=1
8), legend.title=element_text(size=18)) +
  labs(y = expression(z^("2")), x =expression(z^("1")), title = 'Bayesian model
', caption = paste0("Solid line: 3z","\n","Dot-dashed line: 2z")) +
  geom_hline(yintercept = 0, linetype="dashed") + geom_vline(xintercept = 0, li
netype="dashed")  # + scale_color_brewer(palette="Set1")



### Graphical concistency technique and outliers detection test
## Mandel's h statistics
# ISO 5725-2
library(metRology)
library(outliers)
H <- data.frame(h1 = (y1-m1)/sd(y1-m1), h2 = (y2-m2)/sd(y2-m2),
                LVDT = as.factor(levels(proficiency$LVDT)),
                Brand = c(rep('Instron', p), rep('Best', p), rep('Biased', p)))
S <- qmandelh(c(.025,.975), p1) #stragglers
O <- qmandelh(c(.005,.995), p1) #outliers
Consistency <- rep('Medium', p1)
Consistency[which(H$h1<O[1]|H$h1>O[2])] <- rep('Low', length(which(H$h1<O[1]|H$
h1>O[2])))
Consistency[which(H$h1>=S[1]&H$h1<=S[2])] <- rep('High', length(which(H$h1<O[1]
|H$h1>O[2])))
H <- data.frame(H, Consistency)
p1 <- ggplot(H, aes(x=LVDT, y=h1, label=round(h1,2))) +
  geom_point(stat='identity', aes(col=Consistency), size=9) +
```

```
  geom_segment(aes(x=LVDT, xend=LVDT, y=0, yend=h1)) +
  geom_text(color="black", size=4) + ylim(-3.5, 3.5) + facet_wrap(~Brand,scales
="free_x") +
  theme(title=element_text(size=15,face="bold"), axis.text.x = element_text(ang
le = 90, hjust = 1
  ), axis.text = element_text(size=15), axis.title=element_text(size=15,face="b
old"),
  legend.text=element_text(size=15), legend.title=element_text(size=15), strip.
text.x = element_text(size = 15)) +
  labs(y = expression(h[i]^("1")), x = expression(LVDT[i]), title = paste0('lev
el: ', j[1])) +
  scale_color_manual(values=gg_color_hue(r)[c(2,1,3)]) +
  geom_hline(yintercept = c(S,0), color='gray30', lty = 4, size = 1.05) +
  geom_hline(yintercept = O, color='gray30', lty = 1, size = 1.05)
rm(Consistency)
Consistency <- rep('Medium', p2)
Consistency[which(H$h2<O[1]|H$h2>O[2])] <- rep('Low', length(which(H$h2<O[1]|H$
h2>O[2])))
Consistency[which(H$h2>=S[1]&H$h2<=S[2])] <- rep('High', length(which(H$h2>=S[1
]&H$h2<=S[2])))
H$Consistency <- Consistency
p2 <- ggplot(H, aes(x=LVDT, y=h2, label=round(h2,2))) +
  geom_point(stat='identity', aes(col=Consistency), size=9) +
  geom_segment(aes(x=LVDT, xend=LVDT, y=0, yend=h2)) +
  geom_text(color="black", size=4) + ylim(-3.5, 3.5) + facet_wrap(~Brand,scales
="free_x") +
  theme(title=element_text(size=15,face="bold"), axis.text.x = element_text(ang
le = 90, hjust = 1
  ), axis.text = element_text(size=15), axis.title=element_text(size=15,face="b
old"),
  legend.text=element_text(size=15), legend.title=element_text(size=15), strip.
text.x = element_text(size = 15)) +
  labs(y = expression(h[i]^("2")), x = expression(LVDT[i]), title = paste0('lev
el: ', j[2])) +
  scale_color_manual(values=gg_color_hue(r)[c(2,1,3)]) +
  geom_hline(yintercept = c(S,0), color='gray30', lty = 4, size = 1.05) +
  geom_hline(yintercept = O, color='gray30', lty = 1, size = 1.05)
grid.arrange(p1,p2, nrow = 2)
#Youden plot
H <- H[,1:4]
ggplot(H, aes(x = h1, y = h2)) +
  geom_point(mapping = aes(color = Brand), shape = 1, size = 4.5, stroke = 2) +

  geom_text_repel(aes(label=i), size = 6) +
  gg_circle(r=S[2], xc=0, yc=0, color='gray30', type = 'dashed') +
  gg_circle(r=O[2], xc=0, yc=0, color='gray30', type = 'solid') +
  theme(axis.text = element_text(size=18), axis.title=element_text(size=21,face
="bold"), legend.text=element_text(size=18), legend.title=element_text(size=18)
) +
  labs(y = expression(h[i]^("2")), x =expression(h[i]^("1")), title = 'ISO 5725
-2', caption = paste0("Solid line: 3z","\n","Dot-dashed line: 2z")) +
  geom_hline(yintercept = 0, linetype="dashed") +
  geom_vline(xintercept = 0, linetype="dashed")



### Proficiency testing according to Bayesian formulation
H <- data.frame(h1 = (Pr1.sim$mean$B-mean(Pr1.sim$mean$B))/sd(Pr1.sim$mean$B-me
an(Pr1.sim$mean$B)),
                h2 = (Pr2.sim$mean$B-mean(Pr2.sim$mean$B))/sd(Pr2.sim$mean$B-me
an(Pr2.sim$mean$B)),
                LVDT = as.factor(levels(proficiency$LVDT)),
                Brand = c(rep('Instron', p), rep('Best', p), rep('Biased', p)))
```

```
l <- dim(Pr1.sim$sims.list$B)[1]
L <- as.vector(matrix(rep(H$LVDT,l), ncol = p*r, byrow = TRUE))
L <- data.frame(x = as.vector(Pr1.sim$sims.list$B)+1, LVDT = L, Brand = c(rep('
Instron',l*p), rep('Best',l*p), rep('Biased',l*p)))
p1 <- ggplot(L, aes(x, colour = LVDT)) + geom_density(adjust = 4.5, alpha = 0.1
) +
  labs(x = expression(paste(chi[i]^("1"), ' [mm]')), title = paste0('level: ',
j[1])) + xlim(0.980, 1.020) +
  facet_wrap(~Brand,scales="free_x") + stat_function( fun = dnorm, args = list(
mean = Pr1.sim$mean$mu+1, sd = Pr1.sim$mean$sigma.L), col = 'black', lty = 4, l
wd = 1.05) +
  theme(title=element_text(size=18,face="bold"), axis.text = element_text(size=
18), axis.title=element_text(size=18,face="bold"),
  legend.text=element_text(size=18), legend.title=element_text(size=18), legend
.position="null", strip.text.x = element_text(size = 18))
y <- qnorm(c(0.25,0.75), mean=Pr1.sim$mean$mu+1, sd=sqrt(mean(Pr1.sim$sd$B^2)))
x <- qnorm(c(0.25, 0.75))
slope <- diff(y)/diff(x)
int <- y[1] - slope * x[1]
p3 <- ggplot(L, aes(col = LVDT)) + stat_qq(aes(sample = x)) + facet_wrap(~Brand
,scales="free_x") +
  geom_abline(slope = slope, intercept = int, lty = 4, size = 1.05) +
  theme(axis.text = element_text(size=15), axis.title=element_text(size=18,face
="bold"),  title = element_text(size=15), legend.position="null", strip.text.x
= element_text(size = 18)) +
  labs(y =expression(paste(x[ki]^("1"), ' [mm]')), title = paste0('QQ-plot (lev
el: ', j[1],')')) + ylim(.98, 1.02)

l <- dim(Pr2.sim$sims.list$B)[1]
L <- as.vector(matrix(rep(H$LVDT,l), ncol = p*r, byrow = TRUE))
L <- data.frame(x = as.vector(Pr2.sim$sims.list$B)+1, LVDT = L, Brand = c(rep('
Instron',l*p), rep('Best',l*p), rep('Biased',l*p)))
p2 <- ggplot(L, aes(x, colour = LVDT)) + geom_density(adjust = 4.5, alpha = 0.1
) +
  labs(x = expression(paste(chi[i]^("2"), ' [mm]')), title = paste0('level: ',
j[2])) + xlim(0.980, 1.020) +
  facet_wrap(~Brand,scales="free_x") + stat_function( fun = dnorm, args = list(
mean = Pr2.sim$mean$mu+1, sd = Pr2.sim$mean$sigma.L), col = 'black', lty = 4, l
wd = 1.05) +
  theme(title=element_text(size=18,face="bold"), axis.text = element_text(size=
18), axis.title=element_text(size=18,face="bold"),
  legend.text=element_text(size=18), legend.title=element_text(size=18), legend
.position="null", strip.text.x = element_text(size = 18))
grid.arrange(p1,p2, nrow = 2)

y <- qnorm(c(0.25,0.75), mean=Pr2.sim$mean$mu+1, sd=sqrt(mean(Pr2.sim$sd$B^2)))
x <- qnorm(c(0.25, 0.75))
slope <- diff(y)/diff(x)
int <- y[1] - slope * x[1]
p4 <- ggplot(L, aes(col = LVDT)) + stat_qq(aes(sample = x)) + facet_wrap(~Brand
,scales="free_x") +
  geom_abline(slope = slope, intercept = int, lty = 4, size = 1.05) +
  theme(axis.text = element_text(size=15), axis.title=element_text(size=18,face
="bold"),  title = element_text(size=15), legend.position="null", strip.text.x
= element_text(size = 18)) +
  labs(y =expression(paste(x[ki]^("2"), ' [mm]')), title = paste0('QQ-plot (lev
el: ', j[2],')')) + ylim(.98, 1.02)
grid.arrange(p3,p4, nrow = 2)

L<-L[1:t,]
L$x<-Pr1.sim$mean$B+1
L$LVDT<-levels(L$LVDT)
L$Brand <- c(rep('Instron',p), rep('Best',p), rep('Biased',p))
```

```
y <- qnorm(c(0.25, 0.75), mean = Pr1.sim$mean$mu+1, sd = Pr1.sim$mean$sigma.L)
x <- qnorm(c(0.25, 0.75))
slope <- diff(y)/diff(x)
int <- y[1] - slope * x[1]
p3 <- ggplot(L) + stat_qq(aes(sample = x)) +
  geom_abline(slope = slope, intercept = int, lty = 4, size = 1.05, col = 'dark
red') +
  theme(axis.text = element_text(size=15), axis.title=element_text(size=18,face
="bold"),  title = element_text(size=15), legend.position="null", strip.text.x
= element_text(size = 18)) +
  labs(y =expression(paste(x['·i']^("1"), ' [mm]')), title = paste0('QQ-plot (l
evel: ', j[1],')')) + ylim(.98, 1.02)
y <- qnorm(c(0.25, 0.75), mean = Pr2.sim$mean$mu+1, sd = Pr2.sim$mean$sigma.L)
L$x<-Pr2.sim$mean$B+1
slope <- diff(y)/diff(x)
int <- y[1] - slope * x[1]
p4 <- ggplot(L) + stat_qq(aes(sample = x)) +
  geom_abline(slope = slope, intercept = int, lty = 4, size = 1.05, col = 'dark
red') +
  theme(axis.text = element_text(size=15), axis.title=element_text(size=18,face
="bold"),  title = element_text(size=15), legend.position="null", strip.text.x
= element_text(size = 18)) +
  labs(y =expression(paste(x['·i']^("2"), ' [mm]')), title = paste0('QQ-plot (l
evel: ', j[2],')')) + ylim(.98, 1.02)
grid.arrange(p3,p4, ncol = 2)

mean(Pr1.sim$mean$B)+1
Pr1.sim$mean$mu+1
mean(Pr2.sim$mean$B)+1
Pr2.sim$mean$mu+1

ggplot(H, aes(x = h1, y = h2)) +
  geom_point(mapping = aes(color = Brand), shape = 1, size = 4.5, stroke = 2) +

  geom_text_repel(aes(label=i), size = 6) +
  gg_circle(r=S[2], xc=0, yc=0, color='gray30', type = 'dashed') +
  gg_circle(r=O[2], xc=0, yc=0, color='gray30', type = 'solid') +
  theme(axis.text = element_text(size=18), axis.title=element_text(size=21,face
="bold"), legend.text=element_text(size=18), legend.title=element_text(size=18)
) +
  labs(y = expression(h[i]^("2")), x =expression(h[i]^("1"))) +
  geom_hline(yintercept = 0, linetype="dashed") +
  geom_vline(xintercept = 0, linetype="dashed")
Consistency <- rep('Medium', p*r)
Consistency[which(H$h1<O[1]|H$h1>O[2])] <- rep('Low', length(which(H$h1<O[1]|H$
h1>O[2])))
Consistency[which(H$h1>=S[1]&H$h1<=S[2])] <- rep('High', length(which(H$h1<O[1]
|H$h1>O[2])))
H <- data.frame(H, Consistency)
p1 <- ggplot(H, aes(x=LVDT, y=h1, label=round(h1,2))) +
  geom_point(stat='identity', aes(col=Consistency), size=9) +
  geom_segment(aes(x=LVDT, xend=LVDT, y=0, yend=h1)) +
  geom_text(color="black", size=4) + ylim(-3.5, 3.5) + facet_wrap(~Brand,scales
="free_x") +
  theme(title=element_text(size=15,face="bold"), axis.text.x = element_text(ang
le = 90, hjust = 1
  ), axis.text = element_text(size=15), axis.title=element_text(size=15,face="b
old"),
  legend.text=element_text(size=15), legend.title=element_text(size=15), strip.
text.x = element_text(size = 15)) +
  labs(y = expression(h[i]^("1")), x = expression(LVDT[i]), title = paste0('lev
el: ', j[1])) +
  scale_color_manual(values=gg_color_hue(r)[c(2,1,3)]) +
```

```
  geom_hline(yintercept = c(S,0), color='gray30', lty = 4, size = 1.05) +
  geom_hline(yintercept = O, color='gray30', lty = 1, size = 1.05)
rm(Consistency)
Consistency <- rep('Medium', p*r)
Consistency[which(H$h2<O[1]|H$h2>O[2])] <- rep('Low', length(which(H$h2<O[1]|H$
h2>O[2])))
Consistency[which(H$h2>=S[1]&H$h2<=S[2])] <- rep('High', length(which(H$h2>=S[1
]&H$h2<=S[2])))
H$Consistency <- Consistency
p2 <- ggplot(H, aes(x=LVDT, y=h2, label=round(h2,2))) +
  geom_point(stat='identity', aes(col=Consistency), size=9) +
  geom_segment(aes(x=LVDT, xend=LVDT, y=0, yend=h2)) +
  geom_text(color="black", size=4) + ylim(-3.5, 3.5) + facet_wrap(~Brand,scales
="free_x") +
  theme(title=element_text(size=15,face="bold"), axis.text.x = element_text(ang
le = 90, hjust = 1
  ), axis.text = element_text(size=15), axis.title=element_text(size=15,face="b
old"),
  legend.text=element_text(size=15), legend.title=element_text(size=15), strip.
text.x = element_text(size = 15)) +
  labs(y = expression(h[i]^("2")), x = expression(LVDT[i]), title = paste0('lev
el: ', j[2])) +
  scale_color_manual(values=gg_color_hue(r)[c(2,1,3)]) +
  geom_hline(yintercept = c(S,0), color='gray30', lty = 4, size = 1.05) +
  geom_hline(yintercept = O, color='gray30', lty = 1, size = 1.05)
grid.arrange(p1,p2, nrow = 2)

H <- data.frame(H, Dh1 = H$h1-(y1-m1)/sd(y1-m1), Dh2 = H$h2-(y2-m2)/sd(y2-m2))

p1 <- ggplot(H, aes(x=LVDT, y=Dh1, label=round(Dh1,2))) +
  geom_point(stat='identity', aes(col=Consistency), size=9) +
  geom_text(color="black", size=4) + ylim(-0.03, 0.03) + facet_wrap(~Brand,scal
es="free_x") +
  theme(title=element_text(size=15,face="bold"), axis.text.x = element_text(ang
le = 90, hjust = 1
  ), axis.text = element_text(size=15), axis.title=element_text(size=15,face="b
old"),
  legend.text=element_text(size=15), legend.title=element_text(size=15), strip.
text.x = element_text(size = 15)) +
  labs(y = expression(Delta~h[i]^("1")), x = expression(LVDT[i]), title = paste
0('level: ', j[1])) +
  scale_color_manual(values=gg_color_hue(r)[c(2,1,3)])
p2 <- ggplot(H, aes(x=LVDT, y=Dh2, label=round(Dh2,2))) +
  geom_point(stat='identity', aes(col=Consistency), size=9) +
  geom_text(color="black", size=4) + ylim(-0.03, 0.03) + facet_wrap(~Brand,scal
es="free_x") +
  theme(title=element_text(size=15,face="bold"), axis.text.x = element_text(ang
le = 90, hjust = 1
  ), axis.text = element_text(size=15), axis.title=element_text(size=15,face="b
old"),
  legend.text=element_text(size=15), legend.title=element_text(size=15), strip.
text.x = element_text(size = 15)) +
  labs(y = expression(Delta~h[i]^("2")), x = expression(LVDT[i]), title = paste
0('level: ', j[2])) +
  scale_color_manual(values=gg_color_hue(r)[c(2,1,3)])
grid.arrange(p1,p2, nrow = 2)
#Youden plot
H <- H[,1:4]
ggplot(H, aes(x = h1, y = h2)) +
  geom_point(mapping = aes(color = Brand), shape = 1, size = 4.5, stroke = 2) +

  geom_text_repel(aes(label=i), size = 6) +
  gg_circle(r=S[2], xc=0, yc=0, color='gray30', type = 'dashed') +
```

```
  gg_circle(r=O[2], xc=0, yc=0, color='gray30', type = 'solid') +
  theme(axis.text = element_text(size=18), axis.title=element_text(size=21,face
="bold"), legend.text=element_text(size=18), legend.title=element_text(size=18)
) +
  labs(y = expression(h[i]^("2")), x =expression(h[i]^("1")), title = 'Bayesian
 framework', caption = paste0("Solid line: 3z","\n","Dot-dashed line: 2z")) +
  geom_hline(yintercept = 0, linetype="dashed") +
  geom_vline(xintercept = 0, linetype="dashed")
rm(S,O, Consistency, p1, p2, p3, p4)


## Grubbs test
round(qmandelh(c(0, 1) - c(-1, 1)*.05/t, t),3) # critical values for stragglers
round(qmandelh(c(0, 1) - c(-1, 1)*.01/t, t),3) # critical values for outliers
#p.value for Single Grubbs test
round(grubbs.test(Pr1.sim$mean$B+1, type = 10)$p.value,4)
round(grubbs.test(y1, type = 10)$p.value,4)
round(grubbs.test(Pr2.sim$mean$B+1, type = 10)$p.value,4)
round(grubbs.test(y2, type = 10)$p.value,4)
#proof
(1-pmandelh((max(Pr2.sim$mean$B+1)-mean(Pr2.sim$mean$B+1))/sd(Pr2.sim$mean$B),t
))*t
#p.value for Double Grubbs test
round(grubbs.test(Pr1.sim$mean$B+1, type = 11)$p.value,4)
round(grubbs.test(y1, type = 11)$p.value,4)
grubbs.test(Pr1.sim$mean$B+1, type = 11)
grubbs.test(y1, type = 11)
round(grubbs.test(Pr2.sim$mean$B+1, type = 11)$p.value,4)
round(grubbs.test(y2, type = 11)$p.value,4)
grubbs.test(Pr2.sim$mean$B+1, type = 11)
grubbs.test(y2, type = 11)



###  3-hierarchical levels model

data <- list(p=p, r=r, n=n,
          xkiz = structure(.Data = -1+proficiency$yijk[proficiency$Jump==j[1]],

          .Dim=c(n,p,r)))
inits <- function(){
  list(B = matrix(rnorm(p*r, 0, 1.0E-6), nrow = p, ncol = r),
       mu.Bz = rnorm(r, 0, 1.0E-6), sigma.Lz = runif(r, 0, 0.006),
       m = runif(1, -.002, 0.002), sigma.rz = runif(r, 0, 0.004))
}
Pr1.sim <- bugs(data, inits, model.file="C:/PhD/Thesis/LVDT_proficiency-3.txt",
                parameters = c("B", "mu.Bz", "sigma.Lz", "sigma.Rz", "m",
                               "sigma.rz"), n.iter = 1000)
rm(data)
data <- list (p=p, r=r, n=n,
          xkiz = structure(.Data = -1+proficiency$yijk[proficiency$Jump==j[2]],

          .Dim=c(n,p,r)))
Pr2.sim <- bugs(data, inits, model.file="C:/PhD/Thesis/LVDT_proficiency-3.txt",
                parameters = c("B", "mu.Bz", "sigma.Lz", "sigma.Rz", "m",
                               "sigma.rz"), n.iter = 1000)
rm(data)
x1 <- c(y1, 1+as.vector(round(Pr1.sim$mean$B, 4)))
x2 <- c(y2, 1+as.vector(round(Pr2.sim$mean$B, 4)))
sigma1 <- c(sigma1, as.vector(round(Pr1.sim$sd$B, 4)))
sigma2 <- c(sigma2, as.vector(round(Pr2.sim$sd$B, 4)))
Model <- c(rep("Sample mean", p*r), rep("Bayesian mean", p*r))
LVDT <- c(i, i)
```

```
Brand <- rep(c(rep('Instron', p), rep('Best', p), rep('Biased', p)), 2)
x1 <- data.frame(x1, x1-sigma1, x1+sigma1, LVDT, Brand, Model)
x2 <- data.frame(x2, x2-sigma2, x2+sigma2, LVDT, Brand, Model)
colnames(x1)[1:3] <- c('x', 'xmin', 'xmax')
colnames(x2)[1:3] <- c('x', 'xmin', 'xmax')

p1 <- ggplot(x1,aes(x = LVDT, y = x, ymin = xmin, ymax = xmax, color = Model, s
hape = Model)
  ) + geom_point(position=position_dodge(width=1), size = 4.5) + geom_errorbar(
position=position_dodge(width=1), size =1.2
  ) + theme(title=element_text(size=18,face="bold"), axis.text.x = element_text
(angle = 90, hjust = 1
  ), axis.text = element_text(size=18), axis.title=element_text(size=21,face="b
old"),
  legend.text=element_text(size=18), legend.title=element_text(size=18), legend
.position="bottom", strip.text.x = element_text(size = 18)) + labs(
  y = expression(x[i]^("1")), x = expression(LVDT[i]), title = paste0('level: '
, j[1])
  ) + geom_hline(yintercept = c(1-2*sp1, 1, 1+2*sp1), color='darkred', lty = 4,
 size = 1.2) + geom_hline(
  yintercept = c(1-3*sp1, 1+3*sp1), color='darkred', size = 1.2) + scale_y_cont
inuous(
  limits = c(.98, 1.02)) + facet_wrap(~Brand,scales="free_x")
hline.data <- data.frame(z = c(rmB1[1], rmB2[1], mean(x1$x[x1$Brand==z[3]])), B
rand = z)
p1 + geom_hline(aes(yintercept = z), hline.data, color='gray30', lty = 4, size
= 1.2)
#+ scale_color_brewer(palette="Dark2")

p2 <- ggplot(x2,aes(x = LVDT, y = x, ymin = xmin, ymax = xmax, color = Model, s
hape = Model)) +
  geom_point(position=position_dodge(width=.5), size = 4.5) +
  geom_errorbar(position=position_dodge(width=.5), size =1.2) +
  theme(title=element_text(size=18,face="bold"), axis.text.x = element_text(ang
le = 90, hjust = 1
  ), axis.text = element_text(size=18), axis.title=element_text(size=21,face="b
old"),
legend.text=element_text(size=18), legend.title=element_text(size=18), legend.p
osition="bottom", strip.text.x = element_text(size = 18)) + labs(
  y = expression(x[i]^("2")), x = expression(LVDT[i]), title = paste0('level: '
, j[2])
) + geom_hline(yintercept = c(1-2*sp1, 1, 1+2*sp1), color='darkred', lty = 4, s
ize = 1.2) + geom_hline(
  yintercept = c(1-3*sp1, 1+3*sp1), color='darkred', size = 1.2) + scale_y_cont
inuous(
    limits = c(.98, 1.02)) + facet_wrap(~Brand,scales="free_x")
hline.data <- data.frame(z = c(rmB1[2], rmB2[2], mean(x2$x[x2$Brand==z[3]])), B
rand = z)
p2 + geom_hline(aes(yintercept = z), hline.data, color='gray30', lty = 4, size
= 1.2)

Z$z1 <- (as.vector(1+Pr1.sim$mean$B)-1)/sp1
Z$z2 <- (as.vector(1+Pr2.sim$mean$B)-1)/sp2

ggplot(Z, aes(x = z1, y = z2)) +
  geom_point(mapping = aes(color = Brand), shape = 1, size = 4.5, stroke = 2) +

  geom_text_repel(aes(label=i), size = 6) +
  gg_circle(r=2, xc=0, yc=0, color='gray30', type = 'dashed') +
  gg_circle(r=3, xc=0, yc=0, color='gray30', type = 'solid') +
  theme(title=element_text(size=18,face="bold"), axis.text = element_text(size=
18),
```

```
  axis.title=element_text(size=21,face="bold"), legend.text=element_text(size=1
8), legend.title=element_text(size=18)) +
  labs(y = expression(z^("2")), x =expression(z^("1")), title = 'Bayesian model
 4.16', caption = paste0("Solid line: 3z","\n","Dot-dashed line: 2z")) +
  geom_hline(yintercept = 0, linetype="dashed") +
  geom_vline(xintercept = 0, linetype="dashed")  # + scale_color_brewer(palette
="Set1")

Instron <- data.frame(mu1 = c(round(1+Pr1.sim$mean$mu.Bz[1],3),1.002), sr1 = c(
round(Pr1.sim$mean$sigma.rz[1],4), 0.0021), sL1 = c(round(Pr1.sim$mean$sigma.Lz
[1],4), 0.0052), sR1 = c(round(Pr1.sim$mean$sigma.Rz[1],4), 0.0056), mu2 = c(ro
und(1+Pr2.sim$mean$mu.Bz[1],3), 0.999), sr2 = c(round(Pr2.sim$mean$sigma.rz[1],
4), 0.0034), sL2 = c(round(Pr2.sim$mean$sigma.Lz[1],4), 0.0039), sR2 = c(round(
Pr2.sim$mean$sigma.Rz[1],4), 0.0052))
colnames(Instron) <- c(paste0('x[1]','^','(1)'), paste0('sigma[r[1]]','^','(1)'
), paste0('sigma[L[1]]','^','(1)'), paste0('sigma[R[1]]','^','(1)'), paste0('x[
1]','^','(2)'), paste0('sigma[r[1]]','^','(2)'), paste0('sigma[L[1]]','^','(2)'
), paste0('sigma[R[1]]','^','(2)'))
row.names(Instron) <- c('Model 4.16','Model 3.15')
tt <- ttheme_default(colhead=list(fg_params = list(parse=TRUE)))
grid.newpage()
grid.table(Instron, theme = tt)

Bz <- c(as.vector(1+Pr1.sim$sims.list$mu.Bz), as.vector(1+Pr2.sim$sims.list$mu.
Bz))
l <- length(Pr1.sim$sims.list$mu.Bz[,1])
Brand <- c(rep('Instron', l), rep('Best', l), rep('Biased', l))
Level <- c(rep(j[1], l*r), rep(j[2], l*r))
Bz <- data.frame(Bz, Brand, Level)

ggplot(Bz, aes(Bz, fill = Brand, colour = Brand)) + geom_density(adjust = 1, al
pha = 0.1) +
  labs(x = expression(paste(chi['··z'], ' [mm]'))) + xlim(0.990, 1.010) + ylim(
0, 450) +
  facet_wrap(~Level,scales="free_x") +
  theme(title=element_text(size=18,face="bold"), axis.text = element_text(size=
18), axis.title=element_text(size=18,face="bold"),
  legend.text=element_text(size=18), legend.title=element_text(size=18), legend
.position="bottom", strip.text.x = element_text(size = 18))
#+  stat_function(fun = dnorm, args=list(mean=Pr1.sim$mean$mu.Bz[1], sd=Pr1.sim
$sd$mu.Bz[1]))

Bz$Bz <- c(Pr1.sim$sims.list$sigma.Lz, Pr2.sim$sims.list$sigma.Lz)
p1 <- ggplot(Bz[Bz$Level==j[1],], aes(Bz, fill = Brand, colour = Brand)) + geom
_density(adjust = 1, alpha = 0.1) +
      labs(x = expression(paste(sigma[L], ' [mm]')), title = paste0('level: ',
j[1])) + ylim(0, 2250) + xlim(0.0015, 0.0105) +
      theme(title=element_text(size=18,face="bold"), axis.text = element_text(s
ize=18), axis.title=element_text(size=18,face="bold"),
      legend.text=element_text(size=18), legend.title=element_text(size=18), le
gend.position="bottom", strip.text.x = element_text(size = 18))
p2 <- ggplot(Bz[Bz$Level==j[2],], aes(Bz, fill = Brand, colour = Brand)) + geom
_density(adjust = 1, alpha = 0.1) + ylim(0, 2250) + xlim(0.0015, 0.0105) +
  labs(x = expression(paste(sigma[L], ' [mm]')), title = paste0('level: ', j[2]
)) +
  theme(title=element_text(size=18,face="bold"), axis.text = element_text(size=
18), axis.title=element_text(size=18,face="bold"),
  legend.text=element_text(size=18), legend.title=element_text(size=18), legend
.position="bottom", strip.text.x = element_text(size = 18))
a1 <- (Pr1.sim$mean$sigma.Lz/Pr1.sim$sd$sigma.Lz)^2
b1 <- Pr1.sim$mean$sigma.Lz/Pr1.sim$sd$sigma.Lz^2
a2 <- (Pr2.sim$mean$sigma.Lz/Pr2.sim$sd$sigma.Lz)^2
b2 <- Pr2.sim$mean$sigma.Lz/Pr2.sim$sd$sigma.Lz^2
```

```
C <- gg_color_hue(r)
C <- C[c(3,1,2)]
for(w in 3:1){
  p1 <- p1 + stat_function(fun = dgamma, args = list(shape = a1[w], rate = b1[w
]), col = C[w], lwd = 1.05, lty = 4)
  p2 <- p2 + stat_function(fun = dgamma, args = list(shape = a2[w], rate = b2[w
]), col = C[w], lwd = 1.05, lty = 4)
  }
grid.arrange(p1,p2, ncol = 2)

Bz$Bz <- c(as.vector(Pr1.sim$sims.list$sigma.rz), as.vector(Pr2.sim$sims.list$s
igma.rz))
p1 <- ggplot(Bz[Bz$Level==j[1],], aes(Bz, fill = Brand, colour = Brand)) + geom
_density(adjust = 1, alpha = 0.1) +
  labs(x = expression(paste(sigma[r], ' [mm]')), title = paste0('level: ', j[1]
)) + ylim(0, 2250) + xlim(0.0015, 0.0105) +
  theme(title=element_text(size=18,face="bold"), axis.text = element_text(size=
18), axis.title=element_text(size=18,face="bold"),
  legend.text=element_text(size=18), legend.title=element_text(size=18), legend
.position="bottom", strip.text.x = element_text(size = 18))
p2 <- ggplot(Bz[Bz$Level==j[2],], aes(Bz, fill = Brand, colour = Brand)) + geom
_density(adjust = 1, alpha = 0.1) + ylim(0, 2250) + xlim(0.0015, 0.0105) +
  labs(x = expression(paste(sigma[r], ' [mm]')), title = paste0('level: ', j[2]
)) +
  theme(title=element_text(size=18,face="bold"), axis.text = element_text(size=
18), axis.title=element_text(size=18,face="bold"),
  legend.text=element_text(size=18), legend.title=element_text(size=18), legend
.position="bottom", strip.text.x = element_text(size = 18))
a1 <- (Pr1.sim$mean$sigma.rz/Pr1.sim$sd$sigma.rz)^2
b1 <- Pr1.sim$mean$sigma.rz/Pr1.sim$sd$sigma.rz^2
a2 <- (Pr2.sim$mean$sigma.rz/Pr2.sim$sd$sigma.rz)^2
b2 <- Pr2.sim$mean$sigma.rz/Pr2.sim$sd$sigma.rz^2
for(w in 3:1){
  p1 <- p1 + stat_function(fun = dgamma, args = list(shape = a1[w], rate = b1[w
]), col = C[w], lwd = 1.05, lty = 4)
  p2 <- p2 + stat_function(fun = dgamma, args = list(shape = a2[w], rate = b2[w
]), col = C[w], lwd = 1.05, lty = 4)
}
grid.arrange(p1,p2, ncol = 2)

Bz$Bz <- c(as.vector(Pr1.sim$sims.list$sigma.Rz), as.vector(Pr2.sim$sims.list$s
igma.Rz))
p1 <- ggplot(Bz[Bz$Level==j[1],], aes(Bz, fill = Brand, colour = Brand)) + geom
_density(adjust = 1, alpha = 0.1) +
  labs(x = expression(paste(sigma[H], ' [mm]')), title = paste0('level: ', j[1]
)) + ylim(0, 2250) + xlim(0.0015, 0.0105) +
  theme(title=element_text(size=18,face="bold"), axis.text = element_text(size=
18), axis.title=element_text(size=18,face="bold"),
  legend.text=element_text(size=18), legend.title=element_text(size=18), legend
.position="bottom", strip.text.x = element_text(size = 18))
p2 <- ggplot(Bz[Bz$Level==j[2],], aes(Bz, fill = Brand, colour = Brand)) + geom
_density(adjust = 1, alpha = 0.1) + ylim(0, 2250) + xlim(0.0015, 0.0105) +
  labs(x = expression(paste(sigma[H], ' [mm]')), title = paste0('level: ', j[2]
)) +
  theme(title=element_text(size=18,face="bold"), axis.text = element_text(size=
18), axis.title=element_text(size=18,face="bold"),
  legend.text=element_text(size=18), legend.title=element_text(size=18), legend
.position="bottom", strip.text.x = element_text(size = 18))
a1 <- (Pr1.sim$mean$sigma.Rz/Pr1.sim$sd$sigma.Rz)^2
b1 <- Pr1.sim$mean$sigma.Rz/Pr1.sim$sd$sigma.Rz^2
a2 <- (Pr2.sim$mean$sigma.Rz/Pr2.sim$sd$sigma.Rz)^2
b2 <- Pr2.sim$mean$sigma.Rz/Pr2.sim$sd$sigma.Rz^2
for(w in 3:1){
```

```
  p1 <- p1 + stat_function(fun = dgamma, args = list(shape = a1[w], rate = b1[w
]), col = C[w], lwd = 1.05, lty = 4)
  p2 <- p2 + stat_function(fun = dgamma, args = list(shape = a2[w], rate = b2[w
]), col = C[w], lwd = 1.05, lty = 4)
}
grid.arrange(p1,p2, ncol = 2)

Bz <- Bz[1:(2*l),c(1,3)]
Bz$Level <- c(rep(j[1],l), rep(j[2],l))
Bz$Bz <- c(1+Pr1.sim$sims.list$m, 1+Pr2.sim$sims.list$m)
p1 <- ggplot(Bz[Bz$Level==j[1],], aes(Bz)) + geom_density(adjust = 1, alpha = 0
.1) +
  labs(x = expression(paste(mu, ' [mm]')), title = paste0('level: ', j[1])) + x
lim(.990, 1.010) + ylim(0, 450) +
  theme(title=element_text(size=18,face="bold"), axis.text = element_text(size=
18), axis.title=element_text(size=18,face="bold"),
  legend.text=element_text(size=18), legend.title=element_text(size=18), legend
.position="bottom", strip.text.x = element_text(size = 18)) +
  stat_function(fun = dnorm, args = list(mean = 1+Pr1.sim$mean$m, sd = Pr1.sim$
sd$m), col = 'darkred', lwd = 1.2, lty = 4)
p2 <- ggplot(Bz[Bz$Level==j[2],], aes(Bz)) + geom_density(adjust = 1, alpha = 0
.1) +
  labs(x = expression(paste(mu, ' [mm]')), title = paste0('level: ', j[2])) + x
lim(.990, 1.010) + ylim(0, 450) +
  theme(title=element_text(size=18,face="bold"), axis.text = element_text(size=
18), axis.title=element_text(size=18,face="bold"),
  legend.text=element_text(size=18), legend.title=element_text(size=18), legend
.position="bottom", strip.text.x = element_text(size = 18)) +
  stat_function(fun = dnorm, args = list(mean = 1+Pr2.sim$mean$m, sd = Pr2.sim$
sd$m), col = 'darkred', lwd = 1.2, lty = 4)
grid.arrange(p1,p2, ncol = 2)
rm(l, Brand, Level, C)

p1 <- data.frame(matrix(nrow = 3, ncol = 4))
row.names(p1) <- c('Instron','Best', 'Biased')
p2 <- p1
alfa <- .05
for(w in 1:r){
p1[w,1:2] <- format(round(quantile(Pr1.sim$sims.list$mu.Bz[,w], (c(0,1)-c(-1,1)
*alfa/2)), 3), nsmall = 3)
p2[w,1:2] <- format(round(quantile(Pr2.sim$sims.list$mu.Bz[,w], (c(0,1)-c(-1,1)
*alfa/2)), 3), nsmall = 3)
p1[w,3] <- format(round(quantile(Pr1.sim$sims.list$sigma.rz[,w], (1-alfa)), 4),
 nsmall = 4)
p2[w,3] <- format(round(quantile(Pr2.sim$sims.list$sigma.rz[,w], (1-alfa)), 4),
 nsmall = 4)
p1[w,4] <- format(round(quantile(Pr1.sim$sims.list$sigma.Rz[,w], (1-alfa)), 4),
 nsmall = 4)
p2[w,4] <- format(round(quantile(Pr2.sim$sims.list$sigma.Rz[,w], (1-alfa)), 4),
 nsmall = 4)
}
colnames(p1) <- c(paste0('hat(delta)[z[lower]]','^','(1)'), paste0('hat(delta)[
z[upper]]','^','(1)'), paste0('U[r[z]]','^','(1)'), paste0('U[R[z]]','^','(1)')
)
colnames(p2) <- c(paste0('hat(delta)[z[lower]]','^','(2)'), paste0('hat(delta)[
z[upper]]','^','(2)'), paste0('U[r[z]]','^','(2)'), paste0('U[R[z]]','^','(2)')
)
grid.newpage()
grid.table(p1, theme = tt)
grid.newpage()
grid.table(p2, theme = tt)

USL <- +0.01
```

```
LSL <- -0.01
for(w in 1:r){
  s3 <- quantile(Pr1.sim$sims.list$sigma.Rz[,w], round(1-(1-pnorm(3))*2,4))
  p1[w,1] <- format(round((USL-LSL)/(2*s3), 3), nsmall = 3)
  p1[w,2] <- format(round(min(c((USL-Pr1.sim$mean$mu.Bz[w])/s3, (Pr1.sim$mean$m
u.Bz[w]-LSL)/s3)), 3), nsmall = 3)
  rm(s3)
  s3 <- quantile(Pr2.sim$sims.list$sigma.Rz[,w], round(1-(1-pnorm(3))*2,4))
  p1[w,3] <- format(round((USL-LSL)/(2*s3), 3), nsmall = 3)
  p1[w,4] <- format(round(min(c((USL-Pr2.sim$mean$mu.Bz[w])/s3, (Pr2.sim$mean$m
u.Bz[w]-LSL)/s3)), 3), nsmall = 3)
  rm(s3)
  }
colnames(p1) <- c(paste0('hat(R)[p[z]]','^','(1)'), paste0('hat(R)[pk[z]]','^',
'(1)'),
                  paste0('hat(R)[p[z]]','^','(2)'), paste0('hat(R)[pk[z]]','^',
'(2)'))
grid.newpage()
grid.table(p1, theme = tt)
```

10 BIBLIOGRAPHY

[1] A. M. H. van der Veen, «Bayesian analysis of homogeneity studies in the production of reference materials», *Accreditation Qual. Assur.*, vol. 22, n. 6, pagg. 307–319, dic. 2017.
[2] «News from the JCGM-WG1 2017», pag. 3.
[3] «JCGM_100_2008_E.pdf».
[4] «JCGM_200_2012».
[5] «UNI EN ISO 7500-1:2016». feb-2016.
[6] «ISO 306:2013(en), Plastics — Thermoplastic materials — Determination of Vicat softening temperature (VST)». Available at: https://www.iso.org/obp/ui/#iso:std:iso:306:ed-5:v1:en.
[7] «ISO 5725-3:1994(en), Accuracy (trueness and precision) of measurement methods and results — Part 3: Intermediate measures of the precision of a standard measurement method». Available at: https://www.iso.org/obp/ui/#iso:std:iso:5725:-3:ed-1:v1:en.
[8] M. Cox e K. Shirono, «Informative Bayesian Type A uncertainty evaluation, especially applicable to a small number of observations», *Metrologia*, vol. 54, n. 5, pag. 642, 2017.
[9] «ISO 13528:2015(en), Statistical methods for use in proficiency testing by interlaboratory comparison». Available at: https://www.iso.org/obp/ui/#iso:std:iso:13528:ed-2:v2:en.
[10] D. Lunn, C. Jackson, N. Best, A. Thomas, e D. Spiegelhalter, *The BUGS Book: A Practical Introduction to Bayesian Analysis*. CRC Press, 2012.
[11] «ISO 5725-1:1994(en), Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions».
Available at: https://www.iso.org/obp/ui/#iso:std:iso:5725:-1:ed-1:v1:en.
[12] «ISO 5725-2:1994(en), Accuracy (trueness and precision) of measurement methods and results — Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method».
Available at: https://www.iso.org/obp/ui/#iso:std:iso:5725:-2:ed-1:v1:en.
[13] «ISO 5725-4:1994(en), Accuracy (trueness and precision) of measurement methods and results — Part 4: Basic methods for the determination of the trueness of a standard measurement method».
Available at: https://www.iso.org/obp/ui/#iso:std:iso:5725:-4:ed-1:v1:en.
[14] «ISO 5725-5:1998(en), Accuracy (trueness and precision) of measurement methods and results — Part 5: Alternative methods for the determination of the precision of a standard measurement method».
Available at: https://www.iso.org/obp/ui/#iso:std:iso:5725:-5:ed-1:v1:en.
[15] «ISO 5725-6:1994(en), Accuracy (trueness and precision) of measurement methods and results — Part 6: Use in practice of accuracy values».
Available at: https://www.iso.org/obp/ui/#iso:std:iso:5725:-6:ed-1:v1:en.
[16] M. Gasparini, *Modelli probabilistici e statistici con temi d'esame*. CLUT, 2015.
[17] J. L. Devore e K. N. Berk, *Modern Mathematical Statistics With Applications*, 2 edizione. New York ; London: Springer Verlag, 2011.
[18] S. R. Searle, G. Casella, e C. E. McCulloch, *Variance Components*. John Wiley & Sons, 2009.
[19] M. J. Crawley, *The R Book*, 2 edizione. Wiley, 2012.
[20] T. Hothorn e B. S. Everitt, *A Handbook of Statistical Analyses Using R*. CRC Press, 2006.
[21] E. O. Doebelin, A. Cigada, e M. Gasparetto, *Strumenti e metodi di misura. Materiali didattici on-line*, 2 edizione. Milano: McGraw-Hill Education, 2008.
[22] P. McCullagh e J. A. Nelder, *Generalized Linear Models, Second Edition*, 2 edition. Boca Raton: Chapman and Hall/CRC, 1989.
[23] L. Deldossi e D. Zappa, «ISO 5725 and GUM: comparison and comments», *Accreditation Qual. Assur.*, vol. 14, n. 3, pagg. 159–166, mar. 2009.
[24] T. Bayes, «LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S», *Philos. Trans.*, vol. 53, pagg. 370–418, gen. 1763.
[25] «JCGM_101_2008_E.pdf». .

[26] C. Elster e B. Toman, «Bayesian uncertainty analysis for a regression model versus application of GUM Supplement 1 to the least-squares estimate», *Metrologia*, vol. 48, n. 5, pag. 233, 2011.
[27] C. Elster, «Bayesian uncertainty analysis compared with the application of the GUM and its supplements», *Metrologia*, vol. 51, n. 4, pag. S159, 2014.
[28] C. Elster e B. Toman, «Bayesian uncertainty analysis under prior ignorance of the measurand versus analysis using the Supplement 1 to the Guide : a comparison», *Metrologia*, vol. 46, n. 3, pag. 261, 2009.
[29] I. Lira e D. Grientschnig, «Equivalence of alternative Bayesian procedures for evaluating measurement uncertainty», *Metrologia*, vol. 47, n. 3, pag. 334, 2010.
[30] J. M. Bernardo e A. F. M. Smith, *Bayesian Theory*. John Wiley & Sons, 2009.
[31] D. Spiegelhalter, A. Thomas, N. Best, e D. Lunn, *WinBUGS user manual*. version, 2003.
[32] A. E. Gelfand e A. F. M. Smith, «Sampling-Based Approaches to Calculating Marginal Densities», *J. Am. Stat. Assoc.*, vol. 85, n. 410, pagg. 398–409, giu. 1990.
[33] G. Casella e E. I. George, «Explaining the Gibbs Sampler», *Am. Stat.*, vol. 46, n. 3, pagg. 167–174, ago. 1992.
[34] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, e E. Teller, «Equation of State Calculations by Fast Computing Machines», *J. Chem. Phys.*, vol. 21, n. 6, pagg. 1087–1092, giu. 1953.
[35] S. OpenBUGS e M. N. Thomas, «Package 'R2OpenBUGS'», 2013.
[36] D. B. Rubin, «Inference and missing data», *Biometrika*, vol. 63, n. 3, pagg. 581–592, 1976.
[37] «JCGM_102_2011_E.pdf». .
[38] L. Breiman e P. Spector, «Submodel selection and evaluation in regression. The X-random case», *Int. Stat. Rev. Int. Stat.*, pagg. 291–319, 1992.
[39] A. Gelman, «Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)», *Bayesian Anal.*, vol. 1, n. 3, pagg. 515–534, 2006.
[40] D. W. Tholen, «ISO/IEC 17043: the new International Standard for proficiency testing», *Accreditation Qual. Assur.*, vol. 13, n. 12, pagg. 727–730, dic. 2008.
[41] M. Thompson, S. L. Ellison, e R. Wood, «The international harmonized protocol for the proficiency testing of analytical chemistry laboratories (IUPAC Technical Report)», *Pure Appl. Chem.*, vol. 78, n. 1, pagg. 145–196, 2006.
[42] W. R. Thompson, «On a Criterion for the Rejection of Observations and the Distribution of the Ratio of Deviation to Sample Standard Deviation», *Ann. Math. Stat.*, vol. 6, n. 4, pagg. 214–219, 1935.
[43] P.-T. Wilrich, «Critical values of Mandel's h and k, the Grubbs and the Cochran test statistic», *AStA Adv. Stat. Anal.*, vol. 97, n. 1, pagg. 1–10, gen. 2013.
[44] F. E. Grubbs, «Sample Criteria for Testing Outlying Observations», *Ann. Math. Stat.*, vol. 21, n. 1, pagg. 27–58, mar. 1950.
[45] G. Barbato, E. M. Barini, G. Genta, e R. Levi, «Features and performance of some outlier detection methods», *J. Appl. Stat.*, vol. 38, n. 10, pagg. 2133–2149, ott. 2011.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Roberto Fontana for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance and his passion for Statistics helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

My sincere thanks also goes to Prof. Giulio Barbato and Dr. Francesca Pennecchi, for their insightful comments and encouragement, but also for their hard questions which incented me to widen my research from various perspectives.

I thank to INSTRON, the company I work with, that gave me the opportunity of realizing this crazy dream: being a PhD researcher while I was working on company projects. In particular, I am grateful to the general manager Stefano Vergano who always put his trust on me, encouraging and supporting my creativity and to my colleague Andrea Calzolari for the stimulating discussions and his availability on guiding me in the entrenched world of the international standards.

I would also like to thank my dear friend Andrea, for providing me with stead fast support and continuous encouragement throughout my years of study.

Finally, I must express my very profound gratitude to my parents that taught me to drop that craze for foundation-stones, and put the finishing touch to each one of my projects and my brothers for supporting me spiritually throughout writing this thesis.

Last but not the least, a special thanks goes to my wife Martina for standing beside me throughout my career and especially while I was writing this thesis. She has been my inspiration and motivation for continuing to improve my knowledge and move my career forward. She is my rock, and I dedicate this book to her.