# Chapter 1

# Theoretical Concepts for Describing a Replication-levels-based Uncertainty Analysis Approach

*Ander Zarketa-Astigarraga, Alain Martin-Mayor and Manex Martinez-Agirre*

## 1.1. Introduction

Delimiting the scope of uncertainty analysis, which places itself between the fields of mathematics and physical experimentalism, may turn tedious for the newbie who tries to take it to application for the first time. However, performing such an analysis is becoming an accepted standard on fields including any sort of experiment, and the subsequent results are being required to provide information on their degree of exactitude. This calls for a systematization when accounting for the uncertainties of measured magnitudes, and systematizing such an analysis, albeit possible and desirable, asks for a well-founded background on the notions that underpin the uncertainty theory. The fact, anyway, is that there seems to be no conclusive consensus regarding the basic concepts that are meant to constitute the building blocks of the theory; rather, those concepts are to be found on a number of canonical references [1, 3, 5, 7, 8] that, although compose a closed system of notions as a theory already liable to be applied, lack of a unified narrative necessary for constituting a holistic view of the subject.

Hence, the principal aim of the work presented herein is to perform an attempt to gather those disperse pieces of information and to put them in

Ander Zarketa-Astigarraga
Mondragon Unibertsitatea, Faculty of Engineering, Mechanical and Industrial Production, Mondragon, Spain

a format of a structured narrative; the secondary aim is to provide the theoretical background so that its application to practical case studies may be understood on a well-founded basis. As the purpose is to give a bottom-up description of the entities that come to interplay in the uncertainty analysis, the paper is structured as follows: Section 1.2 serves as a reminder of the foundational concepts of physical experimentalism; Section 1.3 introduces the notion of measurement chain, which is the starting point for properly describing the term uncertainty, detailed in Section 1.4. Section 1.5 verses on replication levels, which are central to the process of systematizing the analysis and, finally, Section 1.6 translates the previous concepts to a mathematical formulation, developing the tools that are to be applied on practical grounds.

## 1.2. On the Basic Definitions of Physical Experimentalism

The study of any physical phenomenon, from an empirical standpoint, is comparative. The baseline case is usually represented by a simplified model of a particular phenomenon. The purpose of a modelization is to enclose a minimal set of entities needed to reproduce a phenomenon; such a minimal set is known as a system. The term model, as it is employed herein, refers to the mathematical formulation that describes a physical process taking place within a defined system. The concept of physical phenomenon is related with the notion of change, and the questions to be answered are how, and why, a system changes (the constancy of a system can be understood as a lack of change). As such, a given model contains a set of descriptors that represent the variations of a system. Those descriptors are termed variables in the mathematical formulation, and their material counterparts are combinations of physical properties, or magnitudes. Intuitively, the lesser the number of magnitudes affecting a phenomenon, the less complex the relational analysis becomes. That’s why, as happens for the definition of a system, the formulation of a model seeks to describe a physical process with a set of variables that minimizes the magnitudes coming to interplay. That minimum amount need not be obtained by a unique collection of magnitudes; rather, it is left to the judgement of the experimentalist to choose from the set of potential properties the ones that best describe the different configurations to be found on a predefined system. The only constraint imposed upon the chosen magnitudes is that their discrete values must univocally reproduce a given configuration of a system.

Each of those configurations, represented by disjoint combinations of magnitudes, is known as a state of a system.

The definitions yielded above constitute a mere conceptual framework. A model is no more than a derivation coming from the abstract mathematical strata. Insofar a model addresses entities of different categorical sciences, as physics and mathematics, it remains theoretical unless corroborated by empirical evidence. From this conception, a proof seeks to ascertain the closeness between the physical reality of a phenomenon and its modelization. Accounting for that closeness lies at the core of the comparative nature of experimentalism. The primary act that experiments rely on is observation and a physically observed action constitutes a fact. The validity of an abstract, mathematical-empirical model depends on the assessment of relations between facts, or hypotheses. Hence, a hypothesis is a potentially true statement that links facts together. Thus, for a system undergoing a physical phenomenon~~,~~ a hypothesis constitutes a relation between the magnitudes that determine the states of that system. It follows that a physical fact is acknowledgeable as far as it is measurable, i.e. as far as the observed differences between magnitudes are quantified somehow. On experimental grounds, the magnitudes that constitute the variables of a model become measurands.

Any experimental effort is carried out within a given physical scenario, or experimental set-up. A set-up is projected and built so that a desired system may be reproduced at its states of interest. Inherently, any set-up comes with two main limitations. On the one hand, the experimentalist is supposed to operate on the system somehow, i.e. the system is requested to allow setting a number of measurands to known values. On the other hand, and in addition to the desired physical phenomenon, several other phenomena may be taking place. This implies the potential existence of magnitudes that are not measurands themselves. As such, even though hypotheses are stated in simple terms, yielding a one-to-one, causal relationships between measurands, it is practically unfeasible to incorporate the entire set of physical magnitudes in a set of hypotheses. Based on these considerations, the overall magnitudes of a set-up that interact with a given system may be divided into three categories:

- An independent magnitude is a measurand contained in a hypothesis, and settable by the experimentalist;

- A dependent magnitude is a measurand contained in a hypothesis, and determinable by the experimentalist;
- An extraneous magnitude is not contained in a set of hypotheses, and hence does not constitute a measurand itself. However, it may affect the values of the measurands of a system.

An experimental workflow leading to the determination of a set of dependent magnitudes is called an experiment. Due to the heuristic nature of physical experimentalism, the experiments are required to be normative; a way of restating this is to consider that the same abstract system is usually subjected to physical analysis repeatedly, either in a single set-up or in a number of different set-ups. Regardless of the possible different physical scenarios, for the experiments to be relevantly comparable they must comply with a reproducibility condition. This does not solely mean that the input measurands need to match among different trials, but that the experiments themselves are to be performed in a procedurally equivalent manner. That's why the design of an experiment is meant, ultimately, to define formal procedures, or protocols. These protocols contain the sets of rules for setting the independent magnitudes and determining the dependent ones; additionally, they should provide the experimental conditions under which the extraneous magnitudes are assumed unchanged, or frozen. In such circumstances, a theoretical model may be tested against a number of experiments to prove its validity.

## 1.3. On the Concept of Measurement Chain

When the previous methodology is applied to the study of a physical phenomenon, discrepancies may be found at two conceptual levels. The first one pertains the differences between the theoretical framework and the physical reality; a model may not reflect the overall changes that a system is undergoing, or it may be valid within a restricted range of physical magnitudes, or the effects of such magnitudes may not be acceptably imprinted on the model itself. Regardless of the specific cause, a decoupling between the two categorical strata, namely physics and mathematics, lies at the core of the divergences. Severe as it may sound, this kind of flaw has more to do with a lack of understanding of a phenomenon, rather than with mistakes committed at the execution level of the experiments.

The second source of discrepancies is found at the experimental stage. When a model is assumed valid enough so that its rejection is not considered an acceptable solution for the detected divergencies, explanations are to be found at possible experimental errors. Taking a model for granted is not an exception that raises few times; either if the mathematical expressions reflect basic conservation laws of physics (e.g. zero net production of energy on an isolated system) or refer to previously validated tests against universally accepted standards (e.g. calibration protocols), the underlying mathematical abstractions are not questioned. Instead a sensible and broadly accepted consideration in experimentalism is the existence of errors in measurements.

When the scope of error analysis is limited to experimental reasons, the basic concept serving as a starting point is that of measurement chain. This chain refers to the potential deviations that contribute to the mismatch between a measurand's value and its modeled counterpart. Notice that the comparison with an idealized system also lies at the conception of the measurement chain. According to [5], the disturbances introduced in a system by the mere act of measuring may be described sequentially, and yield a total of five potentially different values for a measurand, which are schematically summarized in Fig. 1.1.

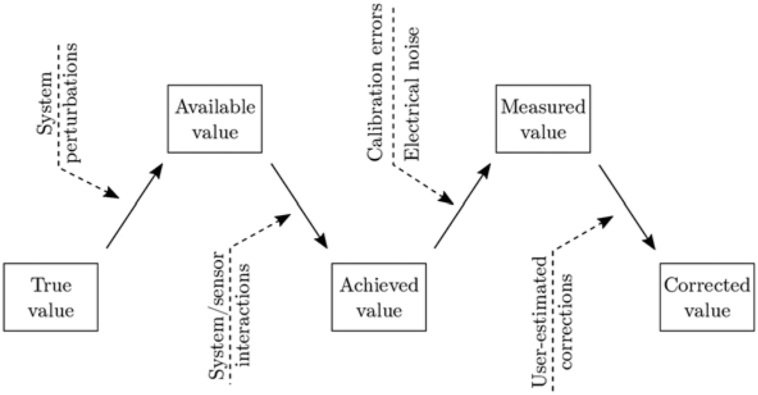

**Fig. 1.1.** Schematic depiction of the different potential values found on a generic measurement chain (adapted from [12]).

- The real value is the hypothetical value the measurand would have if the system were not affected by the measurement process.

- The available value is the value of the measurand in the system, at the measurement point, while the measurement is being taken. A device employed to measure a magnitude is termed sensor, or probe. Since any sensor is intrusive to some extent, it is considered unfeasible to leave a system intact when a probe is introduced. Put on formal terminology: as the sensor itself constitutes another system, what is being measured are the changes in the state of the sensor-system; those changes come at the expense of moving the original system to a different operating point, resulting in a further change of the measurand. Carefully designed sensors or non-intrusive techniques are aimed at minimizing these system disturbances as much as possible.

- The achieved value is the value that the measurand has in the sensor while the measurement is being made. If attention is focused on the system constituted by the sensor itself, it would be naïve to consider that the only magnitude affecting that sensor-system is the measurand. Extraneous magnitudes may further alter the value being recorded, which ultimately cause the sensor to equilibrate with the entire environment rather than with the measurand alone. These deviations are called system/sensor interactions.

- The measured value is the value attributed to the measurand when the output of the sensor is interpreted using the best estimate of the calibration of the sensor. The only potential source of error that stands between achieved and measured values is the measurement system. A proper calibration seeks to provide a complete map between the input and output of a measurement system in terms of a temporal parameter. Even in standardized calibration protocols, there exists a pre-established tolerance related to the acceptability of the measured values relative to the hypothetical real values. This tolerance can fall as small as the resolution of the sensor, which imposes a lower bound on the changes that the sensor itself is able to detect on a measurand, and is usually dictated by physical constraints. However, that tolerance may grow larger due to several reasons, as fabrication defects or above-resolution measurand fluctuations; in such cases of imperfect calibration, the achieved and measured values do not coincide, and the error coming from the calibration stage needs to be accounted for.

- Finally, the corrected value refers to the experimentalist's best estimate of the real value, once the system disturbances, system/sensor interactions and calibration errors are taken into account.

As far as measurands are descriptors of a system, if errors on measurement are assumed as an inherent part of the experiments, then a description of a system lacking information about those errors is incomplete by all means. An acceptable description requires defining certain bounds for the differences between real and corrected values and information must be provided on how close those values are estimated to lie. However, completing a system's description by error information is not to be understood, solely, as a formal good-practice exercise; in fact, another of its main purposes resides in constituting a checking baseline for other experiments. It tells experimentalists whether values measured in a hypothetically equivalent system's state, but under potentially different experimental conditions, is significant. In cases that those values differ more than the expected error range, it may be concluded that either the experimental conditions are not the same for the tests, or that unwanted effects are entering one of the measurement chains. The former cause is not treated herein as a formal experimental error, as it is fixable by modifying the operative point of the mistaken set-up. Accounting for measurement chain perturbations, besides, is related to the nature of two main factors: that of conducting the experiments and that of the error sources.

## 1.4. On the Concept of Uncertainty: Nature of Experiments and Error Sources

Previous considerations make clear that an error refers to a difference between real and corrected values, and that it constitutes a reliability descriptor of a measurand on an experiment. For a single observation on a test, the error is certainly a fixed number, computable by the experimentalist; when a finite number of measurements are performed, an error analysis may be carried out on the recorded data, or the stored values of the measurand. As those values are known, error calculations are based on statistical analysis.

However, the term uncertainty addresses a different concept, as the point is not to compute the error value from known data, but to ascertain a possible value that the error might have on future tests. From this standpoint, uncertainty analysis relies on statistical inference, and not statistical analysis. Uncertainty values are not descriptors in the way errors are, but estimators. The distinction is relevant in a twofold sense: on formal grounds, mathematical tools employed in statistical analysis are different from those used in statistical inference. From a practical point of view, a test owns an unknown error value *a priori*. At best, the error is expected to fall within the range predicted by the uncertainty value. The calculations to obtain this uncertainty may depend on the pool of error data coming from previous tests and, if so, its value changes due to the error resulting from the experiment itself. As such, error and uncertainty analyses may come to influence each other, but that they are defined in such a way so that uncertainties are meant to predict the possible errors committed in further tests.

Thus, uncertainty analysis is predictive by definition, but its relation to error analysis is not. The link between uncertainties and errors depends on the type of experiment conducted. Statistically, experimental tests may be classified in two main groups [3, 7], whose differences are subtle enough to require further explanation. The basic criterion is the availability to obtain independent data points on a given experimental configuration. From a statistical standpoint, achieving independent data points is equivalent to repeating the experiment a number of times. The convenience of repeatability has a well-founded rationale behind: ideally, if the same measurement were taken with acceptably large sets of different observers and instruments as to constitute a statistical population, then the reliability of the measurements could be assured by statistics. Multiple-sample experiments are those in which uncertainties are evaluated by such a repetition. On the other hand, single-sample experiments are constrained, for any reason whatsoever, to a limited number of observations, and their uncertainty evaluation is done by estimation, not statistical calculus.

Although the concept of statistical independence is easily understood, its experimental reality is not as evident. The difficulty seems to come from a number of factors that tend to lessen the repeatability condition [3], such as differences in the reading of the same measurand by several observers, or discrepancies on the same measurement by nominally equivalent, but individually different, sensors. Alternatively, inadequate measurement parameters may lead to single-sample observations [7]. Regardless of the number of samples taken, measurements are done with a given sampling frequency, and the physical phenomenon being studied shows characteristic changing rates, or frequency spectra. If the lapse between consecutive readings happens to be much larger than the lowest characteristic frequency of the system, then the samples may be considered independent. Otherwise, a single-sample experiment results.

The corollary is that the mentioned experimental categories are overlapping. A measurement may satisfy single-sample conditions for certain configurations, whereas it becomes multiple-sampled for others.

The distinction between the mentioned experimental categories arises when treating the recorded data. Statistical inference and statistical analysis are mathematical processing tools used when dealing with single- and multiple-sample experiments, respectively. Nevertheless, experiments are subjected to the same potential sources of error regardless of their statistical nature. Those sources are identified in Section 1.3 as being responsible for the different links that constitute the measurement chain, namely: system disturbances, system/sensor interactions and calibration errors (see Fig. 1.1). Although valid enough, that classification lacks the universality employed when defining the types of error that are found on experiments. Instead, error sources are better classified on a temporal basis alone. If time-dependency is taken as the primary criterion, it follows that errors fall into two main categories: those that change with time, and those that do not. Additionally, the ones that change with time may be predictable on a deterministic way, or be of a wholly random nature. A way of matching this temporal classification and the one introduced in Section 1.3 is to think of a time-marching structure, or a *past-present-future* sequence (which is an *ad hoc* classification made by the authors). *Past errors* are the ones not affected by time anymore, such as fabrication defects or calibration errors that may influence the measurements. *Future errors* refer to those that change with time, but whose effect is known in advance: system disturbances and system/sensor interactions can be regarded as deterministic as long as they follow well-established trends. *Present errors* are to be understood as purely random, either because system disturbances and system/sensor interactions are poorly understood or wrongly considered, or because there are a number of sources entirely contingent on experimental or operative conditions, such as scale- or display-reading actions performed by observers or natural equilibrium fluctuations, respectively.

This classification, based on deterministic grounds, allows performing a preciser denomination of errors. As the so-called past and future errors can be estimated, their effects are considered by the addition of a fixed correction factor to the measurements that compensates the bias introduced in the recorded value; hence, they constitute fixed or bias errors. On the other hand, present errors are not liable to corrections due to their random nature or precision scattering, and are assumed as random or precision errors. The double denomination has to do with historical differences on the terminological conventions regarding single- and multiple-sample uncertainty theories; whereas fixed and random errors are used on single-sample theory, bias and precision errors are typical of multiple-sample experiments.

Regardless of the type of experiment conducted, it is the experimentalist's duty to account for all the correctable errors, so that the tests are carried out with the only unavoidable sources at play, which are the random ones. Additionally, both fixed and random error values are meant to be reported with a justified uncertainty analysis performed on them. A systematic approach to account for those uncertainties lies on the concept of replication levels.

## 1.5. On the Concept of Replication Levels

Replication and repetition are close concepts, but refer to different experimental levels. When a test is said to be repeatable, it is meant that the recorded values of a measurand, on repeated trials, lie on an acceptably narrow range; in other words, that the committed errors are below a certain threshold. Section 1.4 links the concept of repeatability to the stage of data processing, whereby the statistical categories of experiments arise; similarly, it distinguishes between error values and error sources, showing them to be independent from each other. Keeping that distinction, replicability has to do with the conditions under which the repeated trials are assumed to take place; specifically, it points to the potential error sources that are supposed unchanged for all trials. As such, repeatibility is related to the error values resulting from a set of experiments, whereas replicability addresses the error sources of those tests.

Originally, the notion of replicability lies within the scope of single-sample experiments [5-8] and it owns an additional distinctive feature when compared to repeatability; as replicability addresses unchanged error sources, the experimentalist can make different assumptions regarding the sources that remain constant. This leads to defining different orders of replicability in accordance to the constraining level of those assumptions. The interest in considering constant error sources resides in the auxiliary information obtained from such analyses, which ultimately serves to diagnose the system from an uncertainty standpoint. Following the classification of error sources in Section 1.4, three main replication levels are defined [5, 8] (see Fig 1.2).

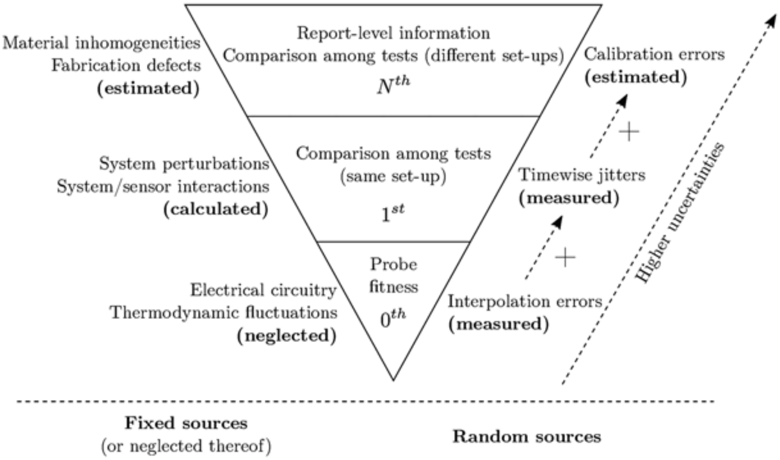

**Fig. 1.2.** Schematic depiction of the replication level philosophy.

- $0^{th}$ order replication level: at this level, time itself is considered frozen. A way of thinking of a time-independent replication pattern is to assume that no error sources are allowed to change, just the reading that different observers make of a certain measurand on the same display. If a picture of that display is taken and shown to a number of observers, the only error source available is the scale-reading interpolation resulting from the resolution of the sensor. The utility of this particular replication pattern is found on the planning stage of an experiment, when a survey is taken over a number of sensors to check their suitability for a given measurement. If the precision of that measurement is required to lie below a predefined value, sensors that own larger interpolation scattering can be discarded in advance. A generalized case should consider the contributions of electrical white noise [2] and thermodynamic equilibrium fluctuations [4]. However, the resolution interval of the device usually subsumes white noise, as it happens to be several orders of magnitude smaller than the detectable electrical output. A way of characterizing the relevance of thermodynamic equilibrium fluctuations is to estimate whether the medium within which measurements are taken behaves as a continuum; as those fluctuations happen at a molecular scale, the comparison between that scale and the probe's representative dimension is used as an indicator. That is precisely the meaning of the Knudsen number, Kn = $\lambda / d$ , where $\lambda$ stands for the molecular mean free path and $d$ is the characteristic dimension of the probe. If the resulting Kn is small, the sensor is large compared to the scale at which thermodynamic fluctuations take place, and their effect is, usually, averaged out over the measurement volume. As such, for typical cases in which white noise and thermodynamic fluctuations are neglected, the $0^{\text{th}}$ order replication level provides the random uncertainty value of the measurement system.

- $1^{\text{st}}$ order replication level: when the temporal parameter is allowed to change, the inherent unsteadiness of the process enters the uncertainty calculation. This timewise jitter, as it is termed, is captured if a number of subsequent measurements are taken while the test is running; from previous considerations, it is important to ensure the independence of those measurements, so that a proper statistical calculation can follow. Additionally, for this replication level to be representative, the measurements must cover a characteristic time-lapse of the measurand's change rate during the experiment. The lower bound of that time-lapse is dictated by the response time of the probe, but the upper bound is specifically test-dependent, and corresponds to the experimentalist to choose a sensible value. As the set of subsequent measurements is obtained while the experiment is running, both timewise jitter and interpolation errors are present on the recorded data. The timewise jitter, measured this way, only accounts for random errors of system disturbance and system/sensor interaction types. Usually, the $1^{\text{st}}$ order replication test is run separately at the debugging stage of the experimental set-up, and its value stored to diagnose further measurements.

- $N^{\text{th}}$ order replication level: it is the broadest replication level conceivable, and may be considered by thinking that, for each measurement, the probe is changed by a similar one coming from the same manufacturer. The potential errors committed at the calibration level are added to the uncertainty value, together with the experimental unsteadiness and interpolation scattering. Typically, tests performed on a single experimental set-up do not face the scattering problems associated with sensor identity, as the measurements are carried out with the same equipment. However, information about calibration uncertainty is necessary when comparing experimental data from different facilities that undertake the same tests. When random calibration errors are considered together with the value coming from the 1st order replication test, the resulting uncertainty is meant to be part of the final report. Unlike the available tests for the lower replication levels, there is not an auxiliary test or a generalized rule of thumb for obtaining the calibration error of a probe. The manufacturer may provide this information in the form of a calibration sheet, but the usual case is not to do so, either because the information itself has not been obtained or because it is concealed on behalf of confidentiality. Instead, the experimentalist has to interpret the specification sheet and retrieve the necessary information, making the correspondent assumptions to justify the data.

The concept of replication levels comprises a property that determines the relation between single- and multiple-sample experiments. By definition, all orders of replication show a common feature, not mentioned so far: they require the repeatability of tests for their determination. Further, they need those repetitions to be both large and independent so that statistical analysis applies on the recorded data. It is precisely the purpose of multiple-sample theory the evaluation of uncertainties by repetitive, independent tests. Hence, it follows that replication orders aim at calculating random uncertainty terms of single-sample experiments by undertaking individual, multiple-sample tests on each of those random terms. If it is not possible to perform multiple-sample tests on a term, which is the case, for example, of calibration-related errors, then statistical inference is used to estimate its contribution. With all, the notion of replication level provides a hinge between the statistical categories of experimental tests.

However, replication levels only account for random errors. For a complete description of the uncertainties, it is necessary to add the correspondent fixed errors at each level. The resultant values are named 0th, 1st and *N*th order uncertainties, respectively. Fixed errors coming from the measurement system are usually due to ground loops or flawed connections, and should not show up if the electrical circuitry is properly designed; hence, 0th order uncertainty is usually equal to the random uncertainty of the measurement system. 1st order fixed errors come from deterministic system disturbances and system/sensor interactions; the typical way of treating them is by analytical-empirical correlations that model the expected effects. If those models quantify the error in terms of the measurand alone, the resultant value is directly added as a fixed error. Instead, if the correction models depend on additional parameters, the uncertainty associated with each of them affects the fixed error value, and needs to be estimated. $N^{\text{th}}$ order fixed errors coming from the calibration stage are as elusive to detect as their random counterparts. Their ultimate sources are fabrication defects or unaccounted calibration biases that unavoidably leak into the tests. On measurement argot, it is said that these errors are *fossilized*, in the sense that they can neither be detected nor removed, but simply assumed. If the calibration process is taken for granted, the fossilized error may be neglected without further concerns. With all, the corollary to be drawn from previous lines is that fixed and bias errors are equivalent terms, in the sense that they both need to be estimated regardless of the type of experiment conducted. Random and precision errors differ on the statistical treatment, which ultimately depends on how large the population of measurements grows.

The outputs generated by single- and multiple-sample analyses are not strictly equivalent. As mentioned, the concept of replication level is linked to the theory of single-sample experiments; as such, the notions of replication order or uncertainty order are defined within that scope. For any reason whatsoever, multiple-sample theory has been, originally, more concerned with analyzing standard procedures [8]. The outcomes of multiple-sample experiments are defined to be the bias limit and the precision index [1]; the former is equivalent to the $N^{\text{th}}$ fixed error of single-sample theory, whereas the second is related (but not equal) to the $1^{\text{st}}$ order random error. When combined, they yield the overall uncertainty, which matches the $N^{\text{th}}$ order uncertainty of single-sample experiments.

With the core concepts laid, what is left is to translate the key definitions to mathematical notation for laying the proper formulation of the theory.

## 1.6. On the Mathematical Basis of Uncertainty Analysis

The primary question posed by the uncertainty analysis is how to process the measured data so that potential errors committed during a test get reflected on the final results. So far, the term *result* has not been formally introduced on its uncertainty-related meaning. Instead, the term *measurand* has been used to address the primary quantity, or magnitude, coming from a measurement and the term *data* has served to refer to the stored values of a given measurand. A result, however, is obtained after processing the data. This processing stage may not only include converting the electrical output into a physical magnitude or applying signal treatment techniques, but making operations with different blocks of data. As such, the distinctive feature of a result, from the uncertainty standpoint, is that it comes from calculations performed by a number of different measurands. Of course, there may be measurands that constitute results themselves, i.e. they need not be combined with other measurands in order to yield a proper descriptor of a system.

Such a definition of result involves considering how the errors that affect each measurand on the calculation are propagated to the derived values. Although the former definition of propagation, due to Kline & McClintock [3], calls uncertainties to the errors carried by each of the measurands to the result, updated versions of the theory state that uncertainties are meant to be specifically result-related [8]; measurands used to derive results are said to own fixed and random errors, if single-sampled, or bias limits and precision indices, if multiple-sampled. The distinction is important insofar the information provided by uncertainties and errors is different; in fact, as uncertainties are obtained by combining errors statistically, the contributions of individual error terms are no longer discernible in an overall uncertainty value.

The analysis begins by stating the functional relation between a generic result $R$ and a number of magnitudes $x_1,\ldots,x_n$:

$$R = f\left(x_1,\ldots,x_n\right) \tag{1.1}$$

The term *magnitude* is used to address each of the $x_i$-s present on Eq. (1.1). From those, the set containing either dependent or independent magnitudes constitutes, by definition, the observed measurands. The rest are extraneous magnitudes that do not get measured, but that may affect the outcome if they are not properly considered; in case the experiment is designed and conducted correctly, those extraneous magnitudes remain unchanged or frozen. Letting $N$ be the number of measurands, and $m$ the number of extraneous magnitudes that, for the sake of simplicity, are assumed frozen, Eq. (1.1) may be reformulated thusly:

$$R = f\left(X_1,\ldots,X_N; x_1,\ldots,x_m\right) \underset{\text{frozen}}{\overset{x1,\ldots xm}{\Rightarrow}} R = f\left(X_1,\ldots,X_N\right), \tag{1.2}$$

where the right hand-side expression is to be understood as the frozen magnitudes not affecting the result.

If the measurands are considered independent and their respective errors small enough, the overall error may be assumed to follow a linearized expression:

$$\delta R = \sum_{i=1}^{i=N} \frac{\partial R}{\partial X_i} \delta X_i \tag{1.3}$$

In Eq. (1.3), the terms $\partial R / \partial X_i$ constitute the error sensitivities, i.e. each of the partial contributions to the overall error due to a unit error in a measurand [6]. The terms $\delta X_i$ represent measurand uncertainties, in case measurands were considered results themselves. However, being terms that enter a result calculation, these factors are named *variation intervals* further on, thus complying with formal terminological conventions.

Further derivations of Eq. (1.3) require assumptions regarding those variation intervals. As mentioned in Section 1.4, uncertainties refer to the probabilities of errors falling between certain values. The procedures for modeling errors and calculating uncertainties, hence, require the usage of probability spaces. Mathematically, when an experiment is conducted, all possible events comprising error observations constitute a set $\Omega$; such a set is interpreted as the potentially observable errors that can be obtained when randomly sampling $\Omega$, which is why $\Omega$ is termed a sample space. An event is the actual observation of a particular error on a given trial; intuitively, it is assumed that certain events are more probable to happen than others, which is expressed by a functional relation $P$ that assigns probabilities to events. The triplet $(\Omega, \mathcal{F}, P)$ is called a probability space; the variable $\mathcal{F}$ refers to a mathematical entity, namely a $\sigma$-algebra structure, that is formally necessary to define a probability space with the required mathematical rigor, but which is not of a major concern for the derivation that follows.

A probability space links events and probabilities together, but it is defined in terms of error observations instead of error values. Providing values to events is accomplished by introducing the concept of random variable that, within the scope of uncertainty analysis, is a functional relation assigning real numbers to error observations. When a random variable $\Phi$ is defined so that it maps each error observation to the actual error value coming from a test, it becomes possible to relate $\Phi$ and the probability assigning function $P$ to yield relevant information about the experiments. If an infinite number of trials were conducted, different relations between $\Phi$ and $P$ would constitute continuous functions upon which the tools of differential calculus would apply. Two of such functions that provide experimental information are the cumulative distribution function and the probability density function, respectively [9]:

$$\begin{aligned} F_{\Phi}: \quad \mathbb{R} \quad &\mapsto \quad \left[0,1\right] \\ \Phi \quad &\mapsto \quad F_{\Phi}(\phi) := P\left(\Phi \leq \phi\right), \end{aligned} \tag{1.4}$$

$$\begin{aligned} f_{\Phi}: \quad \mathbb{R} \quad &\mapsto \quad \mathbb{R} \\ \phi \quad &\mapsto \quad f_{\Phi}(\phi) := \frac{\mathrm{d}F_{\Phi}\left(\phi\right)}{\mathrm{d}\phi} \end{aligned} \tag{1.5}$$

The cumulative distribution function, Eq. (1.4), provides the probability of an error acquiring a value of $\phi$ or less, which is why it is mapped to the interval $\left[0,1\right]$ (the probability of committing any error is unity, whereas that of no error is zero). Its derivative with respect to the random variable is the probability density function (PDF onwards), Eq. (1.5), also named the frequency distribution function. Intuitively, such a concept of derivative may be understood as $f_{\Phi}(\phi)\mathrm{d}\phi$ being the probability of $\Phi$ falling in the interval $\left[\phi, \phi + \mathrm{d}\phi\right]$. Although any of the distribution functions contains the necessary information for mathematically acknowledging a random variable, the PDF provides a straightforward relation to certain experimental data, as this function can be constructed from that data.

So far, no distinction has been made on the particular nature of the errors being treated. Building the PDF of a given experimental error is done by following the definition of Eq. (1.5); for a set of subsequent trials, the number of times a particular error value is measured is plotted against that same value. When normalizing the number of events by the total number of trials, and on the limit of taking the number of trials to infinity, this procedure results on a continuous function that represents the probabilities of measuring those error values, which is precisely the definition of $f_{\Phi}\left(\phi\right)$. However, requiring subsequent trials for the construction of $f_{\Phi}\left(\phi\right)$ discards any possibility of treating fixed errors in such a way, as fixed errors do not manifest themselves on the scattering of a set of consecutive measurements. Instead, $f_\Phi(\phi)$ addresses random errors only, whose specific nature will depend on the order of replication considered in the experiments.

The fact that random errors may be represented by PDFs is acknowledged on early reports treating the subject [10, 3]. The choice of a particular distribution for describing a random error is done according to additional assumptions regarding that randomness. Usually, those assumptions lead to three typical random error distributions [9]:

- A Gaussian or normal distribution (see Fig. 1.3) is used when the scattering of the measurements is considered to be of a random nature itself:

$$f_\Phi\left(\phi \mid \mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}exp\left(-\frac{(\phi-\mu)^2}{2\sigma^2}\right) \tag{1.6}$$

White noise or equilibrium thermodynamic fluctuation measurements, if feasible, would enter this group. So would fluctuations of a measurand above the resolution of the measuring device. The standard uncertainty interval, $\delta$, corresponding to a Gaussian distribution is given by [9]:

$$\delta=\frac{\sigma}{N}, \tag{1.7}$$

being $N$ the number of measurements, or sampling population.

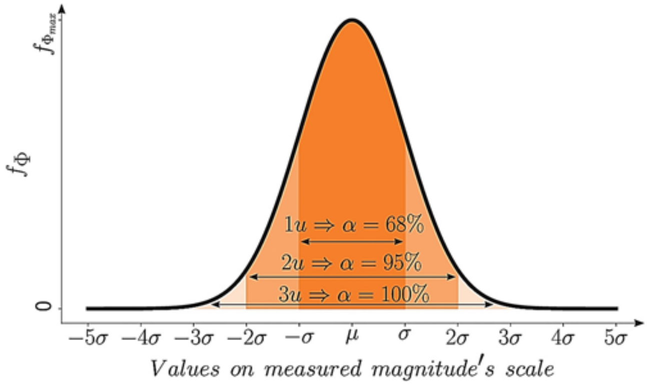

**Fig. 1.3.** Schematic of a Gaussian distribution (may not be properly scaled).

- A uniform distribution (see Fig. 1.4) is used when the interval of possible values is known, but the measurement provides little additional information. Being $[a,b]$ the interval of possible values for $\phi$:

$$f_{\Phi}(\phi|\sigma)=\begin{cases} 0 & \phi<a, \\ \dfrac{1}{2\sqrt{3}\sigma} & a\leq\phi\leq b, \\ 0 & b<\phi \end{cases} \tag{1.8}$$

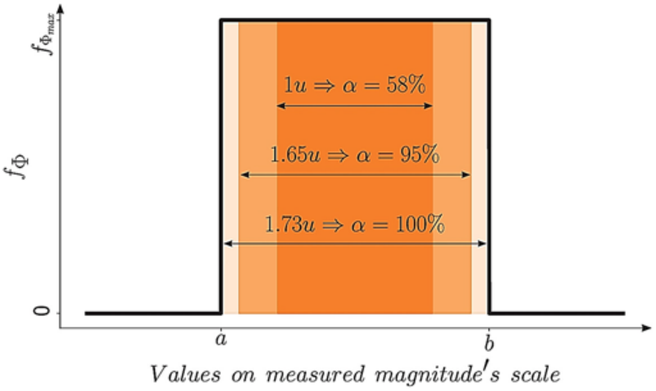

**Fig. 1.4.** Schematic of a uniform distribution (may not be properly scaled).

Again, from [9], the standard uncertainty of a uniform distribution is given by:

$$\delta=\frac{b-a}{2\sqrt{3}} \tag{1.9}$$

Digital-display equipment is considered to follow this distribution at the resolution level of the device. The resolution represents the range $(b-a)$ of possible values, within which the measured value is known to lie. As it is unknown how the device performs the round-off operation, the actual value is equiprobable within the range $(b-a)$, from which the uniform distribution results.

- A triangular distribution (see Fig. 1.5) is used when, in addition to the range $(b-a)$ of possible values of $\phi$, the measurand is considered to lie closer to the central value of that range:

$$f_\Phi(\phi|a,b)=\begin{cases} 0 & \phi<a \\ \dfrac{\phi-a}{\left((b-a)/2\right)^2} & a\le\phi<\dfrac{a+b}{2}, \\ \dfrac{2}{b-a} & \phi=\dfrac{a+b}{2}, \\ \dfrac{b-\phi}{\left((b-a)/2\right)^2} & \dfrac{a+b}{2}<\phi\le b, \\ 0 & b<\phi \end{cases} \tag{1.10}$$

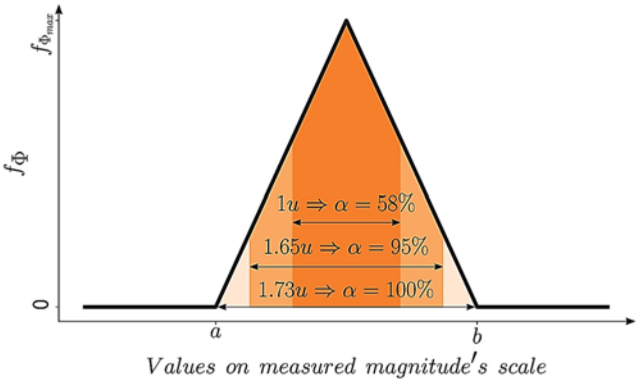

**Fig. 1.5.** Schematic of a triangular distribution (may not be properly scaled).

As for the previous cases, [9] provides the standard uncertainty for a triangular distribution:

$$\delta=\frac{b-a}{2\sqrt{6}}, \tag{1.11}$$

where $(b-a)$ stands for the range of possible values, as before. Triangular distributions are typical for analog-display devices such as calipers or manometers. The actual reading allows discerning the indicator's closeness to the center value, which affects the subsequent round-off operation, unlike for digital displays.

The definition of typical PDFs entails additional statistical concepts not introduced so far, such as the parameters $\mu$ or $\sigma$ in Eqs. (1.6), (1.8). Mathematically, those parameters are grouped under the notion of moment, which serves to quantitatively describe the shape of a function. Formally, the $k$-th moment of a continuous PDF about a point $\hat{\phi}$ is defined as [9]:

$$\mu_k = \int_{-\infty}^{\infty} \left(\phi - \hat{\phi}\right)^k f_\Phi(\phi)\,\mathrm{d}\phi \tag{1.12}$$

The parameter $\mu$ corresponds to the centered ($\hat{\phi} = 0$) first moment ($k = 1$) of a PDF:

$$\mu = \int_{-\infty}^{\infty} \phi f_\Phi(\phi)\,\mathrm{d}\phi\,, \tag{1.13}$$

and it represents the average value of a given distribution, or the mean. The parameter $\sigma^2$ equates to the mean-centered ($\hat{\phi} = \mu$) second moment ($k = 2$) of a PDF:

$$\sigma^2 = \int_{-\infty}^{\infty} (\phi - \mu)^2 f_\Phi(\phi)\,\mathrm{d}\phi, \tag{1.14}$$

and is known as the variance of the distribution ($\sigma^2 = Var$). The parameter $\sigma$ itself, which is unit-consistent with $\mu$, is termed standard deviation, and quantifies the scattering of a distribution around its mean. Although higher-order moments provide additional morphological information on the distribution, the mean and the standard deviation suffice for the purpose of uncertainty analysis.

Running an infinite number of experiments constitutes an idealization, and on practical ground the PDFs are not continuous, but discrete. They are obtained not from the entire population that would result from the idealized set of infinite experiments, but from the finite sampling space that represents the actually undertaken ones. As such, the moments obtained from that finite sampling space do not describe the shape of the theoretical PDF faithfully but, rather, constitute estimators of its moments. If a sufficiently large number of samples is taken, as may be the case for random errors, the estimators will lie close enough to the theoretical moments so that the differences can be regarded as negligible. Otherwise, as happens for fixed errors, the experimentalist is forced to guess the PDF that would result from a hypothetical repetitive pattern of those biases. The main difference between single- and multiple-sample experiments, at a mathematical level, is to be found here: multiple-sample experiments are liable to follow a practically discrete, but conceptually continuous, PDF, whereas single-sample tests are not. However, the point on describing the different replication levels in Section 1.5 is to justify that the distinction between fixed and random errors is formally relative; in fact, fixed errors may be thought of as random errors not been sampled the necessary amount of times to obtain a statistically relevant PDF. Written in simpler terms [3]: fixed errors are also accepted to own a theoretical PDF similar to random errors.

With all, the discrete estimators of the mean and the (squared) standard deviation are written as follows [9]:

$$\bar{\phi} = \frac{1}{N}\sum_{i=1}^{N}\phi_i, \tag{1.15}$$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\phi_i - \bar{\phi}\right)^2 \tag{1.16}$$

The aim of the estimators is to provide the best estimate of a measurand, and the magnitude of the error on that estimation. Employing Eq. (1.15) as the best estimate is straightforward; agreeing on an estimator for the errors, which ultimately stand for each of the $\delta X$ mentioned in Eq. (1.3), is not. The problem with considering the standard deviation as the basic uncertainty estimator is that it is inconsistent for errors coming from different distributions [3]. That inconsistency is found at the definition level. As a data scattering quantifier, the standard deviation is to be understood as follows: a value of a unit standard deviation around the mean encloses a certain area of the underlying PDF; that area is the probability of an error value lying on the range $[\mu - \sigma, \mu + \sigma]$. For the typical PDFs in Eqs. (1.6), (1.8) and (1.10), those probabilities are of 68 %, 58 % and 65 %, respectively (see Figs. 1.3 to 1.5).

Usually, the experimentalist asks for higher probability rates. A confidence level of 95 % is commonplace when reporting experiments, which is to say that, regardless of the PDF, the scattering is quantified with a 95 % of probability of finding future error values lying on the resulting confidence interval around the mean. The advantages of considering uncertainties with such a method are that it homogenizes the differences coming from different distributions, allows propagating the uncertainties with constant probability to the results, and provides the experimentalist with a flexible parameter to play with, namely the settable value of the confidence level. Bearing such considerations in mind, measurands are meant to be reported on the following manner [3]:

$$\phi = \bar{\phi} \pm S\left(b \text{ to } 1\right), \tag{1.17}$$

where $\bar{\phi}$ and $S$ are the arithmetic mean and standard deviation estimators, respectively, and the expression $\left(b \text{ to } 1\right)$ stands for the odds that the experimentalist would be willing to bet that the error is less than $S$. The parameter $b$ and the confidence level are related as $b = 1/\left(1 - \frac{\alpha}{100}\right)$, with $\alpha$ being the chosen level. For the case $\alpha = 95\%$, the previous expression yields $b = 20$, which provides the odds $\left(20 \text{ to } 1\right)$ for the measurand-reporting expression. In addition to the interpretations given so far, such an expression tells that, for consecutive trials of an experiment, the error committed in the measurand value, 19 times out of 20, is supposed to lie below $S$. For the typical Gaussian, uniform and triangular distributions, the $95\%$ confidence level uncertainties are related to the standard uncertainties expressed in Eqs. (1.7), (1.9) and (1.11) by constants, such that $\delta\big|_{\alpha=0.95} = 2\delta$, $1.65\delta$ and $1.81\delta$, respectively.

Reports [3, 8] show that, if a result can be linearized as in Eq. (1.3), with each of the measurands on the expression being independent and expressed as in Eq. (1.17), then a quadratic combination of uncertainty intervals (with $\delta X_i = S_i$) allows propagating the uncertainties to the result with constant probability:

$$S_R = \sqrt{\sum_{i=1}^{i=N} \left( \frac{\partial R}{\partial X_i}^{2} \right) S_{i|R}^{2}} \tag{1.18}$$

With such an approach, fixed and random errors of measurands are kept separated until the last step of computing the uncertainty of a result is undertaken. Thus, Eq. (1.18) is applied separately for fixed and random terms, namely $S_{R|fixed}$ and $S_{R|random}$. The overall uncertainty in a result is given as [1]:

$$\delta X_R = \sqrt{S^2_{R|fixed} + \left(tS_{R|random}\right)^2}\,, \tag{1.19}$$

where $S_{R|fixed}$ and $S_{R|random}$ are calculated from Eq. (1.18). In the original formulation, the parameter $t$ is the Student-t value, which depends on the degrees of freedom used in estimating each of the $S$ factors. For relatively large samples ($N > 30$), $t$ may be assumed to be $\approx 2$; otherwise, the Welch-Satterthwaite formula is to be applied [11, 1]. The purpose of the Student-t parameter is to account for the difference between the statistical parameter (coming from a continuous PDF) and its estimator (discrete), as mentioned before. As the 95 % confidence intervals already account for that difference, the approach undertaken herein does not employ the Student-t parameter, and the formulation changes slightly by already introducing the $S_{R|random}$ terms with their respective 95 % intervals $u_{R|random}$:

$$S_{R|random} = \sqrt{\sum_{i=1}^{i=N}\left(u_{R|random}S_{R|random}\right)^2_i} \tag{1.20}$$

As the errors are propagated with constant probability, the Student-t parameter is absent in the final expression:

$$\delta X_R = \sqrt{S^2_{R|fixed} + S^2_{R|random}} \tag{1.21}$$

The treatment of errors developed so far serves to systematically account for different sources and their propagation to results. When doing so, it is common practice to provide the broadest uncertainty intervals in order to keep the measurements as conservative as possible. However, such a practice hinders the original purpose of uncertainty analysis: if results from different set-ups or facilities are to be compared, shorter uncertainty intervals allow for preciser comparisons, easily detecting differences among tests and capturing mislead measurements. Thus, the claim made herein is to employ the shortest uncertainty intervals available when reporting the measured data.

## 1.7. Concluding Remarks

It has been shown that the growing importance that experimental uncertainty analysis has had historically can be tackled by a systematic procedure lying on the concept of replication levels. Those levels are delimited by the nature of the error sources that enter the experimental test. The $0^{th}$ replication level considers, merely, the reading-interpolation error committed at the measuring device's scope. The $1^{st}$ one adds potential system/sensor interactions, which are accounted for by analytical-empirical correlations, and timewise jitters. The broadest level, namely the $N^{th}$ one, encompasses the uncertainties coming from possible manufacturing defects or calibration errors. The former two levels are calculated by tools coming from the field of statistical inference, whereas the latter can only be estimated based on manufacturers' specification sheets. Anyhow, the mathematical strategy that backs up the mentioned procedure assumes that the combination of different uncertainty intervals is to be performed on a probability-preserving manner. This means that, when certain measurands are combined to yield a derived magnitude, the uncertainty interval of such a magnitude is meant to own a confidence level that matches the one that the original measurands have. This approach provides a way for building a hierarchical classification of magnitudes depending on their functional relation to the basic measurands, with the overall set of magnitudes owning the same confidence level regarding their particular uncertainty intervals. These intervals can be pieced down into a number of contributors that, if traced back, correspond to the basic measurands.

## Acknowledgements

The authors gratefully acknowledge the financial support from the Department of Education of the Basque Government for the Research Grant [PRE_2017_1_0178].

## References

[1]. R. B. Abernethy, R. P. Benedict, R. B. Dowdell, ASME measurement uncertainty, *J. Fluids Eng.*, Vol. 107, 1985, pp. 161-164.

[2]. J. B. Johnson, Thermal agitation of electricity in conductors, *Phys. Rev.*, Vol. 32, 1928, pp. 97-109.

[3]. J. S. Kline, F. A. McClintock, Uncertainties in single-sample experiments, *Mech. Eng.* , 1953, pp. 3-8.

[4]. Y. Mishin, Thermodynamic theory of equilibrium fluctuations, *Annals of Physics*, Vol. 363, 2015, pp. 48-97.

[5]. R. J. Moffat, The measurement chain and validation of experimental measurements, in *Proceedings of the 6th Congress of the International Measurement Confederation (ACTA IMEKO'73)*, Vol. 1, Dresden, Germany, 1973, pp. 45-53.

[6]. R. J. Moffat, Contributions to the theory of single-sample uncertainty analysis, *J. Fluids Eng.*, Vol. 104, 1982, pp. 250-258.

[7]. R. J. Moffat, Using uncertainty analysis in the planning of an experiment, *J. Fluids Eng.*, Vol. 107, 1985, pp. 173-178.

[8]. R. J. Moffat, Describing the uncertainties in experimental results, *Exp. Therm. Fluid Sci.*, Vol. 1, 1988, pp. 3-17.

[9]. J. Olarrea Busto, M. Cordero Gracia, Estadística, 45 Problemas Útiles, 1st Edition, *Garcia Maroto Editores*, Madrid, 2009.

[10]. K. Pearson, On the mathematical theory of errors of judgment, with special reference to the personal equation, *Philosophical Transactions of the Royal Society of London Series A*, Vol. 198, 1902, pp. 235-299.

[11]. B. L. Welch, The generalization of student's problem when several different population variances are involved, *Biometrika*, Vol. 34, 1947, pp. 28-35.

[12]. T. Arts, J.-M. Buchlin, Temperature measurements, Chapter 4, in Measurement Techniques in Fluid Dynamics, 3rd Edition, *von Karman Institute for Fluid Dynamics*, Brussels, 2009.