

---

# Probability in Metrology

Giovanni B. Rossi

Università degli Studi di Genova, DIMEC, Via All'Opera Pia 15 A, 16145 Genova, Italy

`gb.rossi@dimec.unige.it`

**Summary.** The relationship is investigated between probability and metrology, here intended as the science of measurement. Metrology is shown to have historically participated in the development of statistic–probabilistic disciplines, not only adopting principles and methods, but also contributing with new and influential ideas. Two mainstreams of studies are identified in the science of measurement. The former starts with the classical theory of errors and ends with the contemporary debate on uncertainty; the latter originates from the development of a formal theory of measurement and it has attained recent results that make a systematic use of probability as an appropriate logic for measurement. It is suggested that these two mainstreams may ultimately converge in a unique theory of measurement, formulated in a probabilistic language and applicable to all domains of science.

**Key words:** Metrology, measurement theory, probability, uncertainty, measurability

## 1 Probability, statistics, and measurement – An historical perspective

### 1.1 The origins: Gauss, Laplace, and the theory of errors

#### The Gauss problem

In his *Theoria motus corporum coelestium* (1809, [3]<sup>1</sup>) Carl Friedrich Gauss (1777–1855) discusses how to obtain estimates of the parameters of the orbits of heavenly bodies on the basis of a set of observations. In the third section of the second book of the treatise he considers the case of any number of observations and formulates what we here call the Gauss problem: *given  $N$  observations that depend upon  $n$  unknown parameters,  $n < N$ , according to*

---

<sup>1</sup> Note that the references in the bibliography at the end of the chapter have been listed in chronological order, in order to provide an overview of the historical development of the subject.

a known functional relation, and that are affected by measurement errors, estimate the unknown parameters.

In modern notation, we write <sup>2</sup>:

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) + \mathbf{v}, \quad (1)$$

where  $\mathbf{y}$  is a vector of observations,  $\mathbf{x}$  is a vector of the unknown parameters,  $\mathbf{v}$  is the vector of measurement errors, and  $\mathbf{f}$  is a vector function. This framework applies to many problems in metrology and thus it provides a good introduction to the role of probability in metrology.

We have started with this problem in order to present, in this first part of the chapter, a mainstream of studies centered on the problem of measurement uncertainty, starting with Gauss and ending with the current state of play. We analyze the historical development of this subject, from the classical theory of errors, through the contribution of orthodox statistics, up to the development of the *Guide to the Expression of Uncertainty in Measurement* [26] and to some of the main issues of the contemporary debate. In this context we also present our own view, which is based on a general approach to the probabilistic modelling of the measurement process.

In the second part of the chapter, we deal with another area of studies that concern the foundations of measurement and we attempt to establish a formal theory. We consider the contributions of Helmholtz and Campbell and the debate on the possibility of the measurement of ‘sensory events’, promoted by the British Association for the Advancement of Science in the 1930s, which has had consequences up to the present day. Then we present the representational approach to measurement and some criticism of it. We discuss the role of the measuring instrument in a formal theory and the benefit of a probabilistic approach. This part concludes with a brief outline of a probabilistic theory of measurement that we have recently proposed and published and also with an attempt to make some previsions on the possible future role of probability in the science of measurement. We do not attempt to deal with other approaches such as those concerned with fuzzy sets or with the theory of evidence, however.

But now let us go back to Gauss. To confront his problem, he adopts the following estimation criterion: choose the value of  $\mathbf{x}$  that has maximum probability, given the observations  $\mathbf{y}$ . So the estimate,  $\hat{\mathbf{x}}$ , must be such that

$$p(\hat{\mathbf{x}}|\mathbf{y}) = \max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}), \quad (2)$$

where  $p(\mathbf{x}|\mathbf{y})$  is the distribution of the parameters, given the observations and, conversely,  $p(\mathbf{y}|\mathbf{x})$  is the distribution of the observations, given the parameters.

---

<sup>2</sup> See Appendix for notation conventions and for a list of the main symbols used in this chapter.

Gauss uses a Bayesian argument to show that, assuming an indifference prior distribution for  $\mathbf{x}$ , this is equivalent to finding the value  $\hat{\mathbf{x}}$  that maximises the probability of the observations, given the parameters; that is,  $\hat{\mathbf{x}}$  may be equivalently characterised by the property

$$p(\mathbf{y}|\hat{\mathbf{x}}) = \max_{\mathbf{x}} p(\mathbf{y}|\mathbf{x}). \quad (3)$$

If we now assume that all the errors are independent and equally distributed, and if  $p_v(\cdot)$  is the distribution of each of them, we obtain

$$p(\mathbf{y}|\mathbf{x}) = \prod_i p_v(y_i - f_i(\mathbf{x})). \quad (4)$$

To proceed any further, it is necessary to adopt a proper distribution for the errors  $v_i$ . This is where the famous *normal* or *Gaussian* distribution comes into play.

### Gauss's probabilistic model for measurement errors

Gauss distinguishes between *systematic* and *random* errors. This distinction, just mentioned in his *Theoria motus*, is more clearly expressed in the successive *Theoria combinationis observationum erroribus minimis obnoxiae* (1823 [5]). Due to the importance of this topic in metrology, it is worthwhile reading the original text.

‘Certain causes of error,’ he writes

are such that their effect on any one observation depends on varying circumstances that seem to have no essential connection with the observation itself. Errors arising in this way are called *irregular* or *random*. . . . On the other hand, other sources of error by their nature *have a constant effect on all observations of the same class*. Or if the effect is not absolutely constant, its size varies regularly with circumstances that are essentially connected with the observations. These errors are called *constant* or *regular*.

Gauss further observes that ‘this distinction is to some extents relative and *depends on how broadly we take the notion of observations of the same class*.’ He explicitly excludes the consideration of systematic (regular, in his terminology) errors in his investigation and warns that ‘of course, it is up to the observer to ferret out all sources of constant error and remove them’.

This choice of neglecting systematic errors characterises the classical theory of errors and may be its main limitation. We show later that the need to overcome this limitation has been the driving force behind the studies on uncertainty in the second half of the 20th century. But for now let us stay with Gauss's approach and appreciate its merits. We thus come back to the *Theoria motus* to see how he deals with random errors.

He considers a special, but very important case of the general problem (1), the measurement of a single constant quantity  $x$  by repeated observations. In this case the model reads

$$\mathbf{y} = x + \mathbf{v} \quad (5)$$

and the probability distribution for the observations, given  $x$ , is

$$p(\mathbf{y}|x) = \prod_i p_v(y_i - x). \quad (6)$$

At this point, Gauss needs an explicit expression for the distribution of the errors  $p_v$ , and thus assumes some properties that correspond to the common understanding of measurement errors. He assumes that  $p_v$  is symmetric, maximum in its origin, and decreasing on each side of the origin. It may be either defined on a finite support, allowing for a maximum error, or rapidly tending to zero as the argument tends to infinity. Yet these assumptions are not enough to fully define the distribution  $p_v$ . Here is where Gauss makes a simple and genial move: he assumes that the most probable value for  $x$ , once the observations  $\mathbf{y}$  have been acquired, is the arithmetic mean of the observed values, because

it has been customary certainly to regard as an axiom the hypothesis that if any quantity has been determined by several direct observations, made under the same circumstances and with equal care, the arithmetic mean of the observed values affords the most probable value, if not rigorously, yet very nearly at least, so that it is always safe to adhere to it.

This key assumption may be explicated in this way:

$$\hat{x} = \bar{y} \triangleq N^{-1} \sum_i y_i. \quad (7)$$

On the basis of this assumption, Gauss is able to derive his celebrated *normal distribution*, which, in modern notation, reads

$$p(v) = (\sqrt{2\pi}\sigma)^{-1} \exp\left(-\frac{1}{2} \frac{v^2}{\sigma^2}\right), \quad (8)$$

where  $\sigma$  is the standard deviation<sup>3</sup>.

To sum up, Gauss was able to derive a *probabilistic model for random errors in measurement* which still maintains its validity [15]. During the same period, a similar result was reached, using a different route, by Laplace.

---

<sup>3</sup> Instead of considering the standard deviation, Gauss elicits a precision measure,  $h = \sqrt{2}\sigma^{-1}$  and discusses how interpercentile ranges depend upon it.

## Laplace's approach and the theory of errors

A near contemporary to the *Theoria motus* was Pierre-Simon Marquis de Laplace's (1749–1827) *Théorie analytique des probabilités*, published in 1812 [4]. He derived the normal distribution in another way [14]. Let us consider again the case of repeated measurement, as described by model (5). We still assume that the errors  $v_i$  are independent and equally distributed, and we also require that their distribution  $p(v)$  is symmetric about the origin and has a finite support. Let  $\hat{x} = \bar{y}$  be the selected estimate for  $x$  and

$$e = \hat{x} - x \quad (9)$$

the estimation error. Then Laplace shows that  $e$  is asymptotically normally distributed with a variance proportional to  $N^{-1}$ . So we find here another way of deriving the normal distribution: it is the distribution of the estimation error, suitable for long series of observations.

Still another viewpoint may be considered, offered by the central limit theorem [30], traceable again, in a basic formulation, to Laplace [14]. Informally, the basic idea is to consider the measurement error as resulting from the contribution of a large sum of small independent error sources; that is,

$$v = \sum_j w_j. \quad (10)$$

If none of them prevails over the others, the distribution of the resulting error tends to be normal as long as the number of the error sources tends to infinity.

In conclusion, the classical theory of measurement errors, which is due to the contributions of Gauss and Laplace in the main, concerns random errors only and results in a probabilistic model, the normal distribution, whose validity may be supported by different arguments.

- It results from assumptions about the nature of errors (symmetry about the origin, probability of large errors quickly decreasing) plus the axiom that the arithmetical mean of the observations provides the most reliable estimate, for a series of measurements ‘made under the same circumstances and with equal care’.
- It is asymptotically attained when estimating a quantity after a long series of observations of the same quality.
- It also results by assuming the error is the consequence of a large number of error sources, none of which prevails over the others.

We reconsider the theory of measurement errors later on and discuss its merits and limitations, and how to overcome them. But now we have to come back to the original Gauss problem and see how it can be solved.

## The origins of the least squares method

In order to solve the Gauss problem we have to find the value of the parameters  $\hat{\mathbf{x}}$  that maximise the probability of the observations (3). Formula (4) provides an explicit expression for  $p(\mathbf{y}|\mathbf{x})$ : if we substitute the normal distribution (8) in it, we obtain

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi)^{\frac{1}{2}N}\sigma^N} \exp\left\{-\frac{1}{2\sigma^2} \sum [y_i - f_i(\mathbf{x})]^2\right\}. \quad (11)$$

So the value of  $\mathbf{x}$  we are looking for is *the one that minimizes the sum of the squares of the errors*,

$$\sum [y_i - f_i(\mathbf{x})]^2 = [\mathbf{y} - \mathbf{f}(\mathbf{x})]^T [\mathbf{y} - \mathbf{f}(\mathbf{x})] \quad (12)$$

This is how the least squares method appears in the *Theoria motus*. At this point, Gauss considers the linear version of his problem, which in modern notation is

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v} \quad (13)$$

and provides the solution. We do not examine the original development, choosing simply to recall that, in modern notation, the solution may be obtained by pseudo-inversion

$$\hat{\mathbf{x}} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{y}. \quad (14)$$

Finally, if  $\sigma^2$  is the variance of the errors, the variance of the estimate will be

$$\text{Var}(\hat{\mathbf{x}}) = \sigma^2 (\mathbf{A}^T \mathbf{A})^{-1}. \quad (15)$$

Gauss reconsidered the least squares method in much more detail in the *Theoria combinationis*. There he provided another rationale for its use, which is no longer based on assuming a normal distribution for the errors, but rather on the minimisation of the expected mean-square error. In modern terminology, we would say that he presented the method as a way of obtaining a minimum-variance estimate. For our purposes the original derivation is more significant, because it is grounded in probability.

The theory of errors and the method of the least squares provided a great start for the theory of measurement and were the major results of the 19th century. At the beginning of the 20th century new ideas and methods became available to experimenters thanks to the contribution of ‘orthodox’ statistics [36].

## 1.2 Orthodox statistics

### Experiments in metrology

Orthodox or classic is the name given to the statistics developed in the first part of the 20th century and whose principal exponent was Ronald Aylmer

Fisher (1890–1962) [8, 11, 23, 36]<sup>4</sup>. He was a geneticist and had the merit of explicitly addressing some of the main problems that experimenters encounter in their work. This is probably a reason for the success his approach encountered among experimenters, including metrologists. On the other hand the effectiveness of some of his methods, for example, his approach to the design of experiments, may have led to an overestimation of the value of other methods of orthodox statistics, such as their approach to point or interval estimation. A book published in 1964 by John Mandel [12], a statistics consultant of the National Bureau of Standards (NBS), provides a nice synthesis of the statistical tools and instruments available to the metrologists near the middle 1900s, which mainly refer to this school. Orthodox statistics promoted the development of probabilistic–statistical models by providing a store of methods for their use in conjunction with experimentation. Such methods include

- Criteria for the design of experiments, in order to optimise the information obtainable in a finite number of trials
- Methods for the estimation of parameters involved in the models
- Criteria for assessing the validity of the models

We have no room here for dealing with the design of experiments, which anyway is less central to our subject. We do, however, use a very simple example to illustrate the other two points. Consider the measurement of a single constant quantity by a series of  $n$  repeated observations as described by model (5). This model assumes that systematic effects are negligible, as generally admitted in the classic theory of errors. Suppose now that we have a set of  $m$  measuring instruments of the same type, independently calibrated. If we want to apply model (5) to them, we should consider whether, for example, the residual calibration error, which could give rise to a systematic effect, is really negligible. So we may perform a simple experiment that consists in measuring the same fixed quantity  $x$  with all the instruments at our disposal, and repeating the measurement  $n$  times for each instrument, thus collecting a total of  $N = n \cdot m$  observations. We may wish to estimate the variance  $\sigma_v^2$ , to check whether the hypothesis of negligible systematic effect is justified and, if not, to provide a quantitative estimate of the systematic effect. In order to do that, we have to consider a more general model than (5), that is,

$$y_{ij} = x + \theta_i + v_{ij}, \quad (16)$$

where

- $i = 1, \dots, m$  is the index denoting the instruments,
- $j = 1, \dots, n$  is the index denoting the repetitions,
- $\theta_i$  is a random variable representing the residual calibration error of each instrument,
- $v_{ij}$  is an array of random variables representing independent realizations of the same normal variable  $v$ , the random error.

---

<sup>4</sup> We use the term ‘orthodox’, because we prefer to reserve the term ‘classic’ for authors such as Gauss and Laplace.

Note that, as already pointed out by Gauss, the same phenomenon, the residual calibration error  $\theta$ , gives rise to a systematic error if we consider as ‘observations of the same class’ the indications of a single instrument (index  $i$  fixed to, say,  $i_0$ ), whilst it becomes a random variation if we sample instruments from the class of all the instrument of the same type (index  $i$  varying from 1 to  $m$ ). Consider the following averages.

- Grand average,  $\bar{y} = 1/N \sum_{ij} y_{ij}$ , which is an estimate of  $x$ .
- Average per instrument,  $\bar{y}_i = 1/n \sum_j y_{ij}$ .
- Instrument deviations,  $(\bar{y}_i - \bar{y})$ , which is an estimate of  $\theta_i$ .

The variance of  $v$  may be estimated by

$$\hat{\sigma}_v^2 = \frac{1}{N-m} \sum_{ij} (y_{ij} - \bar{y}_i)^2. \quad (17)$$

We now want to check whether the influence of the calibration errors  $\theta_i$  is negligible. To do so, suppose that all the  $\theta_i$  are null: we call this the *null hypothesis* and denote it by  $H_0$ . If  $H_0$  is true, we may estimate the variance of  $v$  also by

$$\hat{\sigma}_v^{2'} = \frac{1}{N-1} \sum_{ij} (y_{ij} - \bar{y})^2 \quad (18)$$

and the result will be, more or less, the same as obtained by (17). We may then check whether the difference between  $\hat{\sigma}_v^2$  and  $\hat{\sigma}_v^{2'}$  is *significant*. If  $H_0$  is true, the difference between  $\hat{\sigma}_v^2$  and  $\hat{\sigma}_v^{2'}$  is only due to the different degrees of freedom of the two estimates, namely,  $\nu_1 = N - m$  for the former and  $\nu_2 = N - 1$  for the latter. It may be proved that the ratio  $\rho = \hat{\sigma}_v^2 / \hat{\sigma}_v^{2'}$ , considered as a random variable, has an  $F$ -Fisher distribution, with parameters  $\nu_1$  and  $\nu_2$ . So a *significance test* may be performed. To do this, we first divide the space of the possible values of the ratio  $\rho$  into two *regions*, one with high probability or highly likely and the other with low probability or unlikely. Note that this is possible because, if  $H_0$  holds, the distribution of  $\rho$  is *known*. Then we compute the value  $\hat{\rho}$  that is the outcome of the data and we check in which region it falls.

- If it falls in the ‘unlikely’ region, we reject  $H_0$  and thus conclude that accounting for  $\theta$  makes a difference and thus the calibration error is not negligible.
- If it falls in the ‘likely’ region, we conclude that the difference between the two estimates  $\hat{\sigma}_v^2$  and  $\hat{\sigma}_v^{2'}$  may be due to their different degrees of freedom and so we may neglect the calibration error  $\theta$  and adopt model (5)<sup>5</sup>.

In the case of the calibration error not being negligible, we may quantitatively evaluate its influence by estimating its variance through

---

<sup>5</sup> We have presented significance testing very informally here, but we discuss it more fully later on.



$$\hat{\sigma}_\theta^2 = \frac{1}{m-1} \sum_i (\bar{y}_i - \bar{y})^2. \quad (19)$$

In conclusion, we think that this example, extremely simple in our opinion, may give an idea of what may be achieved through orthodox statistics in the development of experiments apt to characterise measuring systems. A much more elaborate example is presented by Mandel under the heading ‘systematic evaluation of measuring process’ [12], to which the reader is referred for further details on this subject. Yet before trying to formulate some conclusions on the contribution of orthodox statistics to metrology, we have to discuss two additional key points, namely *estimation* and *statistical testing*.

### Estimation in Fisher’s view

Consider again Gauss’s problem, for simplicity, in the special case of measurement based on repeated observations, as in model (5). Gauss looked for *the value of  $\hat{x}$  having maximum probability, given the (vector) observation  $\mathbf{y}$* , that is, for  $\hat{x}$  such that

$$p(\hat{x}|\mathbf{y}) = \max_x p(x|\mathbf{y}). \quad (20)$$

For calculating  $\hat{x}$  he used a Bayesian argument that is essentially equivalent to the following considerations. Applying the Bayes–Laplace rule<sup>6</sup>, we see that

$$p(x|\mathbf{y}) \propto p(\mathbf{y}|x)p(x). \quad (21)$$

If we assume a uniform distribution for  $x$ <sup>7</sup>, we obtain

$$p(x|\mathbf{y}) \propto p(\mathbf{y}|x). \quad (22)$$

So maximizing  $p(x|\mathbf{y})$  with respect to  $x$ , in view of (22), is computationally equivalent to maximizing  $p(\mathbf{y}|x)$ , that is, to searching  $\hat{x}$  such that

$$p(\mathbf{y}|\hat{x}) = \max_x p(\mathbf{y}|x). \quad (23)$$

But, for Gauss, *if formula (23) may be used for computing  $\hat{x}$ , the meaning of  $\hat{x}$  is still established by formula (20)*. Fisher changes this perspective completely, although obtaining, in this case, the same final result. He argues that the Bayes–Laplace rule (21) can not be applied, unless it is possible to determine an ‘objective’ prior distribution,  $p(x)$ , for  $x$ . So he applies directly formula (23), without deriving it from formula (20), as Gauss does. Although, in this case, the result is the same, the interpretation of the estimate is different:  $\hat{x}$  is interpreted now as the *most likely* value for  $x$ , that is, the value that maximizes the function

<sup>6</sup> We discuss in some detail the Gauss–Laplace rule in Section 1.4.

<sup>7</sup> The reason for assuming a uniform distribution is that prior to making the measurement, all possible values of  $x$  may be considered equally likely.

$$l(x|\mathbf{y}) = p(\mathbf{y}|x), \quad (24)$$

now called the *likelihood function*. Note that the likelihood function is *not* a probability distribution (because, in general, it does not integrate to unity).

There is a major difference between Gauss's and Fisher's approaches because in the former a probabilistic statement is made for  $\hat{x}$ : it is the value having maximum probability, once that  $\mathbf{y}$  has been observed, whilst in the latter no such probabilistic statement is possible. In this sense we may say that maximum-likelihood estimation is not a probabilistic estimation. Orthodox statistics consider other estimation methods, that we can not review here, but to which similar arguments apply [23, 31, 36].

### Epistemological aspects of statistical tests

If Fisher's position on estimation is, in our opinion, not fully convincing, much more interesting is his view of statistical testing. To introduce this subject, let us consider an example of *significance testing*, simple but of high metrological import.

Suppose that we assume, for some measurement process, that model (5) holds. As we have already noted, the potentially critical assumption with this model is the absence of any (noticeable) systematic effect. This hypothesis implies that, for each observation  $y_i$ ,

$$E(y_i) = x, \quad (25)$$

where  $E$  is the expectation operator. A straightforward way of checking the validity of this hypothesis is to apply the measurement system to a standard object, whose value is known,  $x = x_0$ , with negligible uncertainty. In these conditions, the measurement process is described by

$$\mathbf{y} = x_0 + \mathbf{v}, \quad (26)$$

where  $\mathbf{v}$  is a vector of  $N$  independent, zero-mean, normal variables, with variances all equal to an unknown value  $\sigma^2$ . So, if we take the arithmetic mean of the observations  $\bar{y}$ , we expect that it is almost equal to  $x_0$ . The question is *how much may we allow  $\bar{y}$  to differ from  $x_0$ , while still maintaining our model?*

To answer this question we observe that it is possible to prove that the *scaled distance*

$$d = \sqrt{N-1} \frac{\bar{y} - x_0}{\hat{\sigma}}, \quad (27)$$

where  $\hat{\sigma}^2 = (N-1)^{-1} \sum (y_i - \bar{y})^2$  is an estimate of the variance, has a known distribution,

$$p(d) = p_{t,N-1}(d), \quad (28)$$

where  $p_{t,\nu}(\cdot)$  is a  $t$ -Student distribution, with  $\nu$  degrees of freedom. Then the acceptance region, that is, the region where the difference  $d$  is not critical for

our model is a *wide enough* interval around the origin. How wide should it be?

We must fix a small probability  $\alpha$  (typical values are 0.05, 0.01, or 0.001) and then we identify the points  $t_{\alpha/2}$ ,  $t_{1-\alpha/2}$ , such that

$$\int_{-\infty}^{t_{\alpha/2}} p_{t,N-1}(d)dd = \frac{\alpha}{2}, \quad \int_{t_{1-\alpha/2}}^{+\infty} p_{t,N-1}(d)dd = \frac{\alpha}{2}. \quad (29)$$

Then the acceptance region is  $A = [t_{\alpha/2}, t_{1-\alpha/2}]$ : the *a priori* probability that  $\bar{y}$  falls into the acceptance region is high, and is equal to  $1 - \alpha$ , whilst the probability that it falls outside is small. So, once we have made our experiment and have calculated the actual value of  $\bar{y}$ , if it falls into the acceptance region we maintain our model; we say that the test *has corroborated it*; otherwise we reject it, because what we have observed is *highly unlikely* under that model, and we consider the possibility of some systematic effect.

Let us then summarise the logic underlying significance testing: in general,

- We assume a probabilistic model, relying on an  $H_0$  hypothesis and calculate the probabilistic distribution of the observations (or of some function of them, such as the arithmetic mean  $\bar{y}$  just considered),
- We partition the space of the observations into two regions, an acceptance region, where the observations are likely to occur, and a rejection region, where the observations are unlikely to occur; this partitioning is based on the assumption of a conventional value  $\alpha$ , called the level of significance of the test,
- Then we conduct the experiment and if the observations fall into the acceptance region, we maintain the model, otherwise we abandon it.

This logical process may be synthetically called a *hypothetic–deductive inference* [23, 36]: hypothetic, because it starts from assuming a probabilistic model, and deductive because from the model it deduces the probability distribution for the observations, on which the test is based.

Significance testing plays a fundamental role in statistics, as much as *it is the only way of checking statistical models*. Although this way of testing statistical hypotheses, in an embryonic form, may be traced from the very dawn of probability and statistics [23], it was undoubtedly developed by orthodox statisticians. Yet some of them considered another class of statistical tests, *hypothesis tests*. Although in many textbooks they are put together with significance tests, their epistemological status is very different and we think it is wise to keep them distinct. We cannot discuss hypothesis testing thoroughly here, so we just mention it as it differs from significance testing.

In significance testing we test a statistical hypothesis *with respect to its negation: the result may be interpreted as a validation of a statistical model*.

In hypothesis testing we compare two alternative hypotheses,  $H_0$  and  $H_1$ , and although the way we treat both of them is not symmetric, at the end

of such a test we cannot reach absolute conclusions about  $H_0$ , because such conclusions depend on the alternative hypothesis we have chosen.

So the domains of application of these tests are quite different: the former are more suited for *scientific investigation*, the latter for addressing *decision making*. This is, very briefly, the core of the criticism that Fisher directed towards hypothesis testing, that was instead supported by two other orthodox statisticians, Neyman and Pearson.

On this point we agree with Fisher's position.

### **Final remarks on the contribution of orthodox statistics to metrology**

Orthodox statistics has been and still is very influential to metrology. Its contribution is manifold, as we have seen. In our opinion we may elicit two main contributions, namely

- *Addressing the design and evaluation of experiments*, by providing valuable tools for the design (via the design-of-experiments approach) and the evaluation of the influence of the various factors (through the analysis-of-variance method)
- Providing an invaluable tool for *checking statistical models*, by significance testing

On the other hand, the orthodox approach to estimation is, in our opinion, not fully satisfactory and its limit is even more apparent in the following section.

In the second half of the 20th century, when orthodox methods had reached their systematisation and were very popular among many experimenters, including the metrologists, the metrology community felt the need for a critical revision of its entire approach to uncertainty.

## **1.3 The Guide to the Expression of Uncertainty in Measurement**

In the late 1970s, the metrological community recognised the need of reaching an internationally agreed way of expressing uncertainty in measurement. It also recognised the need to accompany the reporting of the result of any measurement by some quantitative indication of its quality, not only in primary metrology, but also in everyday measurements. So, in 1978, the Bureau International des Poids et Mesures (BIPM) carried out an investigation on a large number of laboratories and prepared Recommendation INC-1 (1980), whose guidelines were adopted by the Conference International des Poids et Mesures CIPM. Then an international working group was instituted (ISO/TAG 4/WG 3) for the development of a technical guide. One of the major scientific problems to be faced was the composition of random and systematic effects causing uncertainty. This also required an evolution in the concept of uncertainty

itself. The work of the group was paralleled by intensive scientific debate on such themes. In 1993 an important result was attained with the publication of the *Guide to the Expression of Uncertainty in Measurement* (GUM) [26]. The document had a great impact both on the technical and the scientific side and further stimulated international debate on measurement uncertainty and related topics. Good introductions to the GUM are already available [34] and here we only want to highlight some points that are particularly relevant to our subject and to introduce some trends of the contemporary debate on uncertainty that are the object of the next section.

As we have mentioned, the main problem to be faced was the composition of systematic and random effects in the evaluation of uncertainty. To do this the GUM chose to adopt the paradigm of *indirect measurements*, in which ‘the value of the measurand is obtained by measurement of other quantities functionally related to the measurand. This may be expressed as

$$x = g(\mathbf{z}), \quad (30)$$

where  $x$  is the measurand,  $\mathbf{z}$  a vector of input quantities, and  $g$  is a function. We call this expression the (*GUM*) *evaluation model* or *formula*. Basically it allows us to propagate the uncertainties on the quantities  $\mathbf{z}$  to the measurand  $x$ . Such uncertainties, in turn, may be evaluated on the basis of different pieces of information, which the GUM classifies in two main categories: those coming from a series of observations (type A) and those coming from other sources, such as information provided by the instrument manufacturers, by calibration, by experience, and so on (type B). So the focus moved from the type of the uncertainty sources (systematic versus random) to the type of information on them (type A versus type B). Consequently, it was possible to support, on a pragmatic basis, a common treatment for both of them.

Let us now see how can we deal with direct measurement, that is, measurement which is obtained from the output of a measuring instrument or, more generally, from a measuring system (MS). We may interpret one of the  $z_i$ , for example, the first one, as the indication  $y$  of the MS, that is,  $z_1 = y$ , and the remaining  $z_i$  as ‘corrections’, that should be ideally applied to correct the effect of the various error sources. The (possible) spread of the indications is accounted for by considering the variability of the random variable  $y$ . The evaluation procedure for the standard uncertainty then proceeds as follows.

The variables that appear in the evaluation formula (30) are regarded as random. So, if  $\hat{\mathbf{z}}$  is a ‘best estimate’ of  $\mathbf{z}$  (which usually means that it is its expected value,  $\hat{\mathbf{z}} = E(\mathbf{z})$ ),  $\Sigma_{\mathbf{z}}$  the covariance of  $\mathbf{z}$  and  $\mathbf{b}$  the vector of the sensitivities of  $x$  with respect to  $\mathbf{z}$ , calculated for  $\mathbf{z} = \hat{\mathbf{z}}$ , that is,

$$b_i = \left. \frac{\partial g}{\partial z_i} \right|_{\mathbf{z}=\hat{\mathbf{z}}}, \quad (31)$$

then an estimate of  $x$  may be obtained as

$$\hat{x} = g(\hat{\mathbf{z}}), \quad (32)$$

and the standard uncertainty,  $u$ , to be associated to  $\hat{x}$ , is

$$u = \sqrt{\mathbf{b}^T \Sigma_{\mathbf{z}} \mathbf{b}}. \quad (33)$$

The generalisation in the case of a vector measurand  $\mathbf{x}$  is not given explicitly, but is simple to obtain.

The GUM allows substantial discretion for choosing a formal statistical inference framework and concentrates mainly on practical aspects.

The debate on uncertainty, stimulated by the GUM, has also involved theoretical aspects. Bayesian inference was rapidly recognised as a sensible approach to the problems considered by the GUM, in particular when dealing with a combination of different sources of information. Prior to entering into the debate, we briefly review, in the next section, some of the main ideas of Bayesian inference.

## 1.4 Issues in the contemporary debate on measurement uncertainty

### Bayesian estimation

Consider an experiment in which we perform repeated trials, in each of which an event  $E$  may occur or not. Let  $p$  be the probability of its occurrence in a single trial. Then the probability that  $E$  occurs  $m$  times in  $N$  repeated trials is

$$P(n_N = m | p) = \binom{N}{m} p^m (1 - p)^{N-m}, \quad (34)$$

where  $n_N$  is the number of occurrences of event  $E$  in  $N$  trials. This result was obtained by Jacob Bernoulli and was one of the earliest findings in the theory of probability. Reverend Thomas Bayes (1702–1761) in his *Essay* [1], published posthumously in 1763, considered the problem which is inverse to the above: suppose that we do not know  $p$  and that we perform  $N$  trials of the experiment and find that event  $E$  occurs  $m$  times: how can we estimate the probability  $p$ ?

To *estimate the parameter  $p$*  Bayes intends to *assign a probability distribution to it*, that is, to find a rule for calculating the probability that the value of  $p$  falls in any assigned interval  $[a, b]$ , with  $0 \leq a < b \leq 1$ . He obtains the following formula.

$$P(a \leq p \leq b \mid n_N = m) = \frac{\int_a^b p^m (1 - p)^{N-m} dp}{\int_0^1 p^m (1 - p)^{N-m} dp}. \quad (35)$$

This result comes from assuming a uniform prior distribution for  $p$ , over its range  $[0, 1]$ . Bayes justifies this assumption by observing that it is the proper one when ‘concerning such an event I have no reason to think that, in a certain number of trials, it should rather happen any one possible number of

times than another'. This is a formulation of the principle of *indifference* or of *insufficient reason*: when there is not sufficient reason for treating different possible cases in a different way, they should be treated in the same way [23].

The key idea underlying Bayes' solution was further investigated and generalised, perhaps independently, by Laplace in his *Essay on the probability of causes* [2] and then in his *Analytic theory of probability* [4]. He formulated what is now known as the Bayes–Laplace rule as follows,

if an event can be produced by a number  $n$  of different causes, then the probabilities of these causes given the event are to each other as the probabilities of the event given the causes, and the probability of the existence of each of these is equal to the probability of the event given the cause, divided by the sum of all the probabilities of the event given each of the causes.

In symbols, if we denote the  $i$ th cause by  $A_i$ , we have

$$P(A_i|E) = \frac{P(E|A_i)}{\sum_i P(E|A_i)}. \quad (36)$$

As we have seen, Gauss used a similar, Bayesian, argument in his *Theoria motus* and Laplace's treatise on probability was highly influential throughout the 19th century. On the other hand, orthodox statisticians disagreed on the possibility of calculating the probability of causes and preferred a different approach to estimation, as we have seen. Bayesian estimation became popular again due to the works of de Finetti, Ramsey, Jeffreys, and others [36].

In modern terms, a Bayesian estimation problem may be formulated and solved as follows [22]. Consider a series of observations  $\mathbf{y}$ , depending upon a unobservable parameter  $x$ : then the parameter  $x$  may be estimated by assigning a probability distribution to it, conditioned by the observations:

$$p(x|\mathbf{y}) = \frac{p(\mathbf{y}|x)p(x)}{\int p(\mathbf{y}|x)p(x) \mathrm{d}x} \propto p(\mathbf{y}|x)p(x). \quad (37)$$

The probability distribution  $p(x)$  is called the *prior distribution* for the parameter  $x$  and may, in general, incorporate prior knowledge about it. A special, but very important case for us, is that in which the indifference principle is used for assigning the  $p(x)$ , which is also called, in this case, a *vague prior*. If we adopt a vague prior, the essence of Bayesian estimation may be summarised as follows. *It is a probabilistic inference aimed at assigning a probability distribution to some quantity  $x$  on the basis of a set of observations  $\mathbf{y}$  and of an hypothesis on a probabilistic relation,  $p(\mathbf{y}|x)$ , linking the quantity and the observations.* As such it may be called a *hypothetic inductive inference* [23, 36].

Bayesian inference has been applied to measurement problems in different ways so far, corresponding to different ways of formulating the core hypothesis. We review them in the following sections.

Furthermore, it is useful to compare the definition above with that which we provided for *significance tests*: we regarded them as *hypothetic deductive inferences* instead. This distinction is essential for understanding the logic of the measurement process and we return to it at the end of this section.

## A Bayesian approach to the evaluation of measurement uncertainty

A first Bayesian approach to the evaluation of uncertainty in measurement is due to Weise and Wöger [24], and other authors, and is documented by several papers illustrating its application to measurement problems (see [34] for a bibliography). We now try to summarise it, considering the presentation by Lira [34] in particular, with the important case of the direct measurement of a quantity by a MS, subject to both random variations and an additive systematic effect, by repeated observations.

If we assume, for maximum simplicity, that the indications of the MS are already properly scaled so that its response function is unitary, we may model this process as

$$\mathbf{y} = x + \theta + \mathbf{v}, \quad (38)$$

where  $x$  is the measurand,  $\mathbf{y}$  is a vector of  $N$  indications of the MS,  $\theta$  is a unknown constant systematic effect, and  $\mathbf{v}$  is a vector of random variations, that are independent realizations of a normal random variable  $v$ , with zero mean and known<sup>8</sup> variance  $\sigma^2$ . Note that the structure is similar to a Gauss problem, but it differs from it due to the presence of the systematic effect  $\theta$ . Note also that the structure is similar to that of formula (16) but with an important difference: here we have a single instrument and so we have no way of directly observing the effects of  $\theta$ .

In this approach we need to begin with an evaluation equation which we may state in the following way,

$$x = g(\tilde{y}, z) = \tilde{y} + z, \quad (39)$$

where  $\tilde{y}$  is the mean value of  $y$  and  $z$  is the (ideal) ‘correction’ of the systematic effect  $\theta$ ; that is, very simply,

$$z = -\theta. \quad (40)$$

Note that formula (39) is a special case of the GUM evaluation formula (30). Then we may consider the vector of parameters  $[x, \tilde{y}, z]$ <sup>9</sup> and apply the Bayes–Laplace rule to it,

<sup>8</sup> The case of an unknown variance may also be treated, but our aim here is to keep the example as simple as possible.

<sup>9</sup> We may be surprised by the apparent dishomogeneity between  $\tilde{y}$  on one side and  $x$  and  $z$  on the other, because  $\tilde{y}$  is a mean value. The reason is that  $y$  varies during the repeated observations, whilst  $x$  and  $\theta$  do not. So, to combine them in the same expression, we have to consider the mean value of  $y$  instead of its individual realizations. Actually here there is a criticality because the motivation of formula (39) in this approach is essentially heuristic.



$$p(x, \tilde{y}, z | \mathbf{y}) \propto p(\mathbf{y} | x, \tilde{y}, z) p(x, \tilde{y}, z). \quad (41)$$

Because the indications  $\mathbf{y}$  depend on  $x$  and  $z$  only through the expected value  $\tilde{y}$ , the formula simplifies as

$$p(x, \tilde{y}, z | \mathbf{y}) \propto p(\mathbf{y} | \tilde{y}) p(x, \tilde{y}, z).$$

The joint distribution  $p(x, \tilde{y}, z)$  may be factorised as

$$p(x, \tilde{y}, z) = p(x | \tilde{y}, z) p(\tilde{y}, z) = p(x | \tilde{y}, z) p(\tilde{y}) p(z),$$

having further assumed the independence of  $\tilde{y}$  from  $z$ . With the evaluation equation above in mind, we obtain

$$p(x | \tilde{y}, z) = \delta(x - g(\tilde{y}, z)) = \delta(x - \tilde{y} - z),$$

where  $\delta$  is the Dirac-delta operator. If we also assume a uniform prior for  $\tilde{y}$ , we have

$$p(x, \tilde{y}, z | \mathbf{y}) \propto p(\mathbf{y} | \tilde{y}) \delta(x - \tilde{y} - z) p(z).$$

To reach the final distribution, we integrate out  $\tilde{y}$  and  $z$  and we obtain the marginal distribution

$$p(x | \mathbf{y}) \propto \int \int p(\mathbf{y} | \tilde{y}) \delta(x - \tilde{y} - z) p(z) d\tilde{y} dz. \quad (42)$$

In order to proceed with the analytical calculations, let us now assume, as anticipated, that  $v$  is normal with known variance  $\sigma^2$ . Then

$$p(\mathbf{y} | \tilde{y}) \propto \exp\left(-\frac{1}{2} \frac{(\tilde{y} - \bar{y})^2}{\sigma^2/N}\right)$$

and, finally,

$$p(x | \mathbf{y}) \propto \int \exp\left(-\frac{1}{2} \frac{(x - \bar{y} - z)^2}{\sigma^2/N}\right) p(z) dz. \quad (43)$$

A distribution is thus assigned to the measurand on the basis of the observations  $\mathbf{y}$  and of the following hypotheses.

- A probabilistic model for the observations,  $p(\mathbf{y} | \tilde{y})$
- The evaluation equation  $x = g(\tilde{y}, z)$
- A probability distribution for  $z$ ,  $p(z)$

This approach requires the previous assumption of an evaluation equation, which relies on an essentially heuristic basis (see Footnote 9). In the next section we consider a different approach: we first present a general probabilistic model of the measurement process and then we consider an alternative Bayesian approach, based on that model.

## A probabilistic model of the measurement process

Recently a general probabilistic model of the measurement process (MP) has been proposed [35, 42]. It starts from the basic consideration that measurement is performed through a measuring system (MS) [25] and envisages a general functional description of it. The MS interacts with the measurand and produces an observable output, the indication, which is related to or, in other words, is caused by, the value of the measurand. So it is quite natural to describe the behaviour of the MS by an input–output model, whose input is the value of the measurand and whose output is the instrument indication. Such an input–output relationship may be experimentally determined by calibration. When we perform a measurement, we get an indication on the basis of which we are able to identify, within the uncertainty limitations, the value of the measurand, because we know in advance, thanks to the calibration, the cause–effect relation linking the two. The measurement process may thus be broken down into two subprocesses, namely:

- *Observation*, the process of producing an observable output that is caused by the measurand and depends on its value
- *Restitution*, the process of identifying the value of the measurand from the indication(s) of the MS

Consequently, *measurement* may be viewed as the process resulting from the chaining of observation and restitution and allowing a value (the measurement value) to be assigned to the measurand. Observation is always performed by the MS, whilst restitution may either be embedded in the MS or performed off-line, depending upon the technology. In any case it seems conceptually correct to distinguish between the two, because the former is a chain of physical transformations, whilst the latter is a kind of information processing. We also show how this distinction is practical, as much as it permits the development of a systematic approach to the modelling of measurement processes, which allows the final result of measurement to be expressed as a probability distribution over the set of the possible values of the measurand. Let us now see how can we describe all this in deterministic terms first. This may be seen as the description of the ideal MP and paves the way to the presentation of the probabilistic model. In a deterministic framework, *observation* may be described by a function that expresses the cause–effect relationship holding between the value of the measurand  $x$  and the indication  $y$ ; that is,

$$y = f(x). \quad (44)$$

The function  $f$  may be called the *response characteristic* of the MS, because it expresses the input–output behaviour of the MS, or also *calibration function*, because it may be experimentally determined by calibration [25]. For example, if  $x$  is the temperature,  $t$ , of an object, and the MS is a measuring chain made of a thermocouple, an amplifier, and a voltmeter, then  $y$  is the voltage reading,  $V$ , from the voltmeter and  $f$  includes the (direct) thermo-electric function for the thermocouple,  $f'$ , and the gain of the amplifier,  $A$ . So we have

$$V = Af'(t).$$

This deterministic description is ‘ideal’ in that we assume that the measuring system behaves exactly according to its response function  $f$  and that no other quantity influences the measurement.

*Restitution*, on the other hand, may be viewed as the inversion of observation, because for any indication  $y$  we provide the measurement value  $\hat{x}$  by

$$\hat{x} = f^{-1}(y). \quad (45)$$

In our example, we have

$$\hat{t} = A^{-1}f'^{-1}(V/A).$$

Note that for standard thermocouples both  $f'$  and  $f'^{-1}$  are standardised functions (polynomials).

In a traditional environment, restitution may be performed manually, whilst in a computerised measuring process it is performed automatically. Anyway, irrespective of the technology, the concept is the same.

Finally, *measurement* is the concatenation of the two transformations,

$$\hat{x} = f^{-1}[f(x)] = x, \quad (46)$$

and results in a unitary transformation, due to the fact that the deterministic model provides a description of an ideal MP. The meaning of this last equation is the following. If the MS behaved exactly according to its characteristic function  $f$ , and no other uncertainty cause applied, then we would obtain the exact value of the measurand. Of course this is not the case in real measurements, but this ideal scheme allows us to introduce the probabilistic framework, which instead properly represents an uncertain environment.

The results thus far obtained are summarised in Table 1, under the ‘deterministic model’ heading. We provide additional arguments in support of this model later on, in Section 2.2.

The probabilistic model may be obtained by translating what we have thus far exposed in probabilistic terms. The natural probabilistic description of the observation process, that is, the natural counterpart of the calibration function, is provided by the conditional probability distribution

$$p(y|x),$$

**Table 1.** Comparison between the deterministic model and the probabilistic one.

Process/ Subprocess	Deterministic Model	Probabilistic Model
Observation	$y = f(x)$	$p(\mathbf{y} x, \boldsymbol{\theta})$
Restitution	$\hat{x} = f^{-1}(y)$	$p(x \mathbf{y}) = \int_{\boldsymbol{\Theta}} p(\mathbf{y} x, \boldsymbol{\theta}) [\int_X p(\mathbf{y} x, \boldsymbol{\theta}) dx]^{-1} p(\boldsymbol{\theta}) d\boldsymbol{\theta}$
Measurement	$\hat{x} = f^{-1}[f(x)] = x$	$p(\hat{x} x) = \int_Y \delta[\hat{x} - E(x \mathbf{y})] [\int_{\boldsymbol{\Theta}} p(\mathbf{y} x, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}] d\mathbf{y}$

where

$x \in X$  : value of the measurand

$\mathbf{y} \in \mathbf{Y}$  :  $N$ -dimensional observation vector

$\boldsymbol{\theta} \in \boldsymbol{\Theta}$  :  $K$ -dimensional parameter vector

$\hat{x} \in X$  : measurement value

that is, the probability distribution of the indication  $y$ , for any given measurement value  $x$ . In other words, whilst in the deterministic model for each value of the measurand we get one and only one indication, in the probabilistic case we may obtain a plurality of indications, ruled by a probability distribution. Consequently, *restitution may be described as the probabilistic inversion of the transformation defining observation*. Such an inversion may be performed according to the Bayes–Laplace rule. If we assume a uniform prior for  $x$ , we obtain

$$p(x|y) = \frac{p(y|x)}{\int p(y|x) dx} \propto p(y|x).$$

In this way we account for random variations, but how can we deal with systematic effects? The systematic effect (of any type, additive, multiplicative, etc.,) of an influence quantity  $\theta$  may be expressed *by allowing the distribution  $p(y|x)$  to also be conditioned by  $\theta$* , thus becoming

$$p(y|x, \theta).$$

If we now apply the Bayes–Laplace rule, the result will be still conditioned by  $\theta$ ; that is,

$$p(x|y, \theta) \propto p(y|x, \theta).$$

To attain the final distribution  $p(x|y)$  it is sufficient to ‘decondition’ with respect to  $\theta$ , by applying the principle of total probability; that is,

$$p(x|y) \propto \int p(y|x, \theta) p(\theta) d\theta.$$

The generalisation to a vector of observations  $\mathbf{y}$  and to a vector of influence parameters is immediate and yields, for observation

$$p(\mathbf{y}|x, \boldsymbol{\theta}) \quad (47)$$

and for restitution

$$p(x|\mathbf{y}) \propto \int p(\mathbf{y}|x, \boldsymbol{\theta})p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}. \quad (48)$$

Restitution yields a probabilistic distribution rather than a single value, as happened in the deterministic case. Yet even now it is possible to define a single measurement value as

$$\hat{x} = E(x|\mathbf{y}). \quad (49)$$

This formula *provides the most general definition of  $\hat{x}$*  as a function of the indications  $\mathbf{y}$ . When it is possible to make it explicit as a function of  $\boldsymbol{\theta}$  also, we obtain

$$\hat{x} = h(\mathbf{y}, \boldsymbol{\theta}), \quad (50)$$

which is another, essentially equivalent, way of expressing the evaluation equation envisaged by the GUM, but with an important difference: such an equation *is now derived by the model*.

Finally, the overall measurement process may be described by combining (chaining) observation and restitution in order to obtain the distribution of the measurement value for each possible value of the measurand, that is,  $p(\hat{x}|x)$ . This may be done by observing that the measurement value  $\hat{x}$  is a function of the indications  $\mathbf{y}$ , which, in turn, in observation are regarded as a vector random variable, conditioned by  $x$ . So, applying a formula for the propagation of distributions, we obtain

$$p(\hat{x}|x) = \int_{\mathbf{Y}} \delta[\hat{x} - E(x|\mathbf{y})] \int_{\boldsymbol{\Theta}} [p(\mathbf{y}|x, \boldsymbol{\theta})p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}] \, \mathrm{d}\mathbf{y} \quad (51)$$

which describes the overall *measurement* process. This completes the set of formulas of the probabilistic model, collected in Table 1, where they are compared with the corresponding deterministic ones. Further generalisations are possible (e.g., considering a vector measurand [35]), but are not treated here, for the sake of simplicity.

Note the different meaning of the distributions  $p(x|\mathbf{y})$  and  $p(\hat{x}|x)$ . The former,  $p(x|\mathbf{y})$ , is the distribution that describes restitution: whenever we observe the (vector, in general) indication  $\mathbf{y}$  we may assign to the measurand the distribution  $p(x|\mathbf{y})$ . This distribution is thus *the basis for providing the measurement value,  $\hat{x} = E(x|\mathbf{y})$ , and its uncertainty*. For example, the standard uncertainty may be defined as

$$u = \sqrt{\text{Var}(x|\mathbf{y})}, \quad (52)$$

and the expanded uncertainty, at a coverage level  $p_0$ , as the value  $U > 0$  such that

$$\int_{\hat{x}-U}^{\hat{x}+U} p(x|\mathbf{y}) \, dx = p_0. \quad (53)$$

Instead,  $p(\hat{x}|x)$  is the distribution that describes the overall measurement process and relates the measurement value  $\hat{x}$  to the value of the measurand  $x$ . Then it may be used *for declaring the performance of the measuring system*.

In summary, we may say that the two distributions considered thus far,  $p(x|\mathbf{y})$  and  $p(\hat{x}|x)$ , are *complementary in meaning and purpose*: the former is the basis for expressing the uncertainty in a specific measurement; the latter is useful for expressing the performance of the measuring system in its measuring range (spanned by  $x$ ) [39,42]. We may say that the former is of primary interest for instrument users and the latter for instrument producers.

Let us now look at the application of this model to the example discussed in the previous section, that is, the measurement of a constant quantity  $x$ , by repeated observations from an instrument affected both by random variations and systematic additive deviations, according to the model (38)

$$\mathbf{y} = x + \theta + \mathbf{v}.$$

For observation, we obtain

$$p(\mathbf{y}|x, \theta) = \prod_i p_v(y_i - \theta - x) \quad (54)$$

and for restitution

$$p(x|\mathbf{y}) \propto \int \prod_i p_v(y_i - \theta - x) p(\theta) \, d\theta. \quad (55)$$

When the distribution of  $v$  is normal with known variance  $\sigma^2$ , we have

$$p(x|\mathbf{y}) \propto \int \exp\left(-\frac{1}{2} \frac{(x - \bar{y} + \theta)^2}{\sigma^2/N}\right) p(\theta) \, d\theta, \quad (56)$$

which is equivalent to formula (43), because  $z = -\theta$ .

The derivation is now more straightforward and there is no need to assume an evaluation equation, rather the results derive directly from the model (38).

Lastly, the distribution that characterises the overall measurement process is

$$p(\hat{x}|x) \propto \int \exp\left(-\frac{1}{2} \frac{(\hat{x} - x - \theta)^2}{\sigma^2/N}\right) p(\theta) \, d\theta. \quad (57)$$

In the accompanying DVD numerical examples of the application of this model to uncertainty evaluation and to risk analysis are provided, with the related software codes. Additional examples may be found in References [34, 39, 40].

### Some notes on inference in measurement

We call *probabilistic* an *inference* that yields the assignment of a probability distribution to some parameter under investigation.

We have encountered two kinds of such inferences so far, namely

- (a) *Hypothetic deductive* inferences, in significance testing
- (b) *Hypothetic inductive* inferences, in Bayesian estimation

Let us briefly recall their logical structure. In a hypothetic deductive inference

- (a1) We hypothesise a probabilistic model,
- (a2) On the basis of which we deduce the probability distribution of the observation  $\mathbf{y}$  in a given experiment, which allows us
- (a3) To define an acceptance region  $A$  for the observation, which is a region in which the observation complies with the model;
- (a4) Then we perform the experiment and acquire  $\mathbf{y}$ :
  - If  $\mathbf{y}$  falls into the acceptance region, the model is corroborated,
  - Otherwise it is ‘falsified’ by the observation and we may consider abandoning it.

In a hypothetic inductive inference, instead, if we consider the most important case for us, that of assuming a noninformative prior,

- (b1) We hypothesise a probabilistic relation, in a given experiment, linking the observation  $\mathbf{y}$  to a parameter  $x$ , expressed as a conditional distribution  $p(\mathbf{y}|x)$ ;
- (b2) We perform the experiment and acquire the observation  $\mathbf{y}$ ;
- (b3) On the basis of the observation and of the hypothesised probabilistic relation, we assign a probability distribution to  $x$ , induced through the observation.

Let us now considered the logical structure of the measurement process, as outlined in the previous section. It includes the following steps.

- (c1) Assume a probabilistic relation between the value of the measurand and the indications of the MS, parametrical in respect to some influence parameters: this relation is a model of the observation process;
- (c2) Assume a probability measure over the space of the influence parameters;
- (c3) Perform observation and acquire the indications of the MS;
- (c4) Apply, in the restitution phase, the Bayes–Laplace rule and obtain a probability distribution for the measurand, still conditioned upon the influence parameters;
- (c5) Decondition the probability distribution with respect to the influence parameters, which concludes the restitution phase and the overall measurement process.

If we analyse this procedure in the light of what we have so far exposed, we recognise in steps c1, c3, and c4 a Bayesian inference, so that we may say that the measurement process *embeds* a Bayesian inference.

On the other hand, we also note that steps c2 and c5 are not typical of a Bayesian inference. They include the assumption of a probability distribution for some parameters (step c2) and their use according to the rules of the calculus of probability (step c5). We say that these two steps form a *hypothetic-deductive process*<sup>10</sup>: so we conclude that in general *in a measurement process we have the combination of a hypothetic-inductive inference and of a hypothetic-deductive process*.

This conclusion does not apply only to the approach based on the probabilistic model of the MS, but also to the previous approach, based on formula (39). Even in that case no inference is made on the influence parameters giving rise to systematic effects: actually no inference is possible because their effects are not observable via the indications.

We thus now have a new way of posing the problem of systematic effects. Because influence parameters giving rise to systematic effects must be treated via a hypothetic deductive process, what guarantees the validity of the final measurement result?

This question is a special case of the general requirement for scientific statements: they must be *controllable*, as it must be possible to design and perform experiments whose results may falsify such theories. This principle, the *falsifiability* of scientific theories, is central to Popper's epistemology and widely accepted [36].

So what can we do in the case of measurement?

The answer, from what we have seen so far, is simple and straightforward: the validity of the measurement process, which includes a hypothetic-deductive treatment of systematic effects, may be controlled by a significance test, that is, by a hypothetic-deductive inference.

Let us briefly see how this inference can be stated. Consider a measurement process described by  $p(\hat{x}|x)$ . Remember that this distribution accounts for systematic effects too. Suppose that we dispose of a standard whose value,  $x_0$ , is known with uncertainty negligible for our purpose. Then we can measure the standard through the measurement process under consideration and perform a significance test on the difference  $\hat{x}_0 - x_0$ , where  $\hat{x}_0$  is the measurement value obtained after measuring the standard. For a significance level  $\alpha$ , the acceptance region will be  $A = [-a, +a]$ , such that

$$\int_{-a}^{+a} p(\hat{x}_0 - x_0|x_0) d\hat{x} = \alpha. \quad (58)$$

---

<sup>10</sup> We distinguish between a hypothetic-deductive process and a hypothetic-deductive inference: in the latter we learn from experience, whilst in the former we do not. We show how to apply a hypothetic deductive inference to measurement in a moment.



This procedure formalises what is done in the practice of metrology, for example, in the verification of the calibration of a MS or in the control of a measurement process by check standards.

## 2 Towards a probabilistic theory of measurement

### 2.1 Origin and early development of the formal theory of measurement

#### Helmholtz

So far we have considered a mainstream of studies, centered on the problem of measurement uncertainty, from Gauss up to contemporary practice. In reality, there is another very important area of studies in the science of measurement that arose towards the end of the 19th century and concerns the problem of the foundations of measurement and the development of a formal theory for it. These two mainstreams, although conceptually related, have developed essentially in parallel and with few connections. The reason for this lack of connection is historical and we think that, at present, a merger of these two approaches is much needed and is a major challenge for metrology. We thus briefly review some of the main steps in the historical development of measurement theory and then we show why this theory also requires a probabilistic approach. We then overview what has been done so far and discuss what we may expect in the near future [41].

The beginning of the modern theory of measurement is usually traced to a genial work by Helmholtz, '*Counting and Measuring from the Viewpoint of the Theory of Knowledge*,' published in 1887 [6]. In this essay he poses the problem of the foundation of measurement, because he investigates 'the objective meaning of the fact that we express as quantities, through *concrete numbers*, situations of real objects' and he discusses 'under what circumstances we are allowed to do so.' 'Concrete numbers', in his language, are those arising from the counting of real objects.

He finds a brilliant solution to the problem by establishing an analogy between measurement and counting. The key idea is that in many cases what we want to measure is literally a 'quantity,' in the sense that it is the amount of something, and thus it may be considered to be composed of the sum of a number of elementary parts, or units, of that something. In these cases measurement is equivalent to the counting of such units.

Counting is possible thanks to the properties of natural numbers which undergo an order based on the relation 'greater than or equal to,' denoted by  $\geq$ , and may be added to each other by an addition operation, denoted by  $+$ .

Similarly, measurement is possible and well founded whenever it is possible to identify the empirical counterparts of the order relation and of the addition operation for the objects carrying the characteristic of interest.

The main idea of Helmholtz, that measurement represents properties of objects by assigning numbers to them in such a way as to reproduce empirical relations in the numerical domain, has been the basis for the development of a theory of measurement.

## Campbell

The first organic presentation of a theory of measurement was by Norman Campbell, in the second part of his book, *Physics: The Elements* [7], published in 1920. Like Helmholtz, he considers the problem of ‘Why can and do we measure some properties of bodies while we do not measure others’ and goes further in this investigation by asking, ‘What is the difference between the properties which determine the possibility or impossibility of measuring them.’ In order to answer this question, he distinguishes two main kinds of quantities, *fundamental*, such as mass or length, and *derived*, such as density, for example. Both of them require an empirical property of *order*, which is – as for Helmholtz – the basic requirement for measurement. But fundamental quantities allow for a *physical-addition* operation also. *Why is this operation so important?*

Because it is the key to permitting *the general procedure for fundamental measurement* to be applied. Such a procedure consists in constructing a *measurement scale*, that is, a *series of standards* with properly assigned numerical values, and then in comparing any unknown object,  $r$ , to it, in order to select the element in the series which is equivalent to it. Then it will be possible to assign to  $r$  the same number (measure) as the selected element.

The physical addition operation must satisfy – as Helmholtz had already pointed out – the logical properties of addition, that is, the associative and the commutative properties, and there must be experimental evidence of this.

On the other hand, derived quantities do not require a specific scale to be devised, because they may be measured thanks to a physical law relating them to other measurable quantities. In the case of density  $\rho$ , for example, we may define it as the ratio of mass to volume, that is,  $\rho = m/V$ , and thus reduce its measurement to one of mass and volume.

Campbell’s investigation contributed to a deeper understanding of the nature of measurement and his position has been, and still is, highly influential and he was deeply involved in a controversy that arose in the 1930s in the scientific community and that would strongly influence the development of measurement science.

## The Committee of the British Association for the Advancement of Science

In the beginning of the 20th century measurement was popular not only in physics and engineering, but also in experimental psychology and in behavioural sciences. So it was quite natural for the scientific community

to consider whether the epistemological value of measurement in this new domain was well founded. With this aim, in 1932, the British Association for the Advancement of Science appointed a committee composed of physicists and psychologists, to consider and report upon the possibility of quantitative estimates of sensory events. The report of the committee, published only in 1939, after years of discussions, admitted that it had been impossible for the two sides to reach a common understanding of measurement [9]. The physicists, in particular, took a strong stance against the possibility of actually making measurements in the behavioural sciences.

Without entering into detail, the committee considered typical psychophysical experiments on ‘just perceptible differences’ and on ‘equal appearing intervals.’ The psychologists claimed that from those experiments and by assuming some feasible psychophysical law, such as Fechner’s law, it was possible to arrive at quantifying sensations. The physicists, instead, denied that, mainly because, in their opinion, direct estimation of sensations was not possible and additivity was inconceivable for them.

The report of the committee had an enormous influence in the following years and we may say that it led to an essentially parallel development of measurement science in physical science on one side and in behavioural sciences on the other, with consequences up to the present day. But let us see now some reactions from an outstanding psychologist, Stanley Stevens, who did not himself attend the committee, although his work on loudness was thoroughly discussed by them.

## Stevens

Stevens, who was at Harvard and was dealing with problems similar to those considered by the committee in the same period, felt the need for a more general theory of measurement. This generalisation was aimed at enlarging the number of feasible measurement scales [10]. In Campbell’s view there was only one type of measurement scale, the one holding for quantities for which an empirical operation of addition was possible. Stevens instead proposed his famous fourfold classification of measurement scales, which is still in use and is summarised in Table 2. The classification is based on the notion of admissible transformations, that is, transformations that leave the scale form invariant. In the table we may see the scale types (column 3) and the groups of admissible transformations (column 5).<sup>11</sup> In doing so he shifted the focus from empirical relations, such as order, additivity..., to the *invariance properties* of the scales.

*Nominal* scales are involved in classification operations and numbers serve only to distinguish one class of objects from another. Any *biunivocal* transformation is permissible, because identification is still possible. Examples are colour measurements and pattern recognition techniques.

---

<sup>11</sup> The content of the other columns is presented later on.

**Table 2.** Summary of the main scales for fundamental measurement, as considered in the representational theory, based on the original classification by Stevens.

Empirical Structure	Empirical Relations	Scale Type	Representation	Admissible Transformations
Nominal	Equivalence among elements in each class	Nominal	$a \sim b \iff m(a) = m(b)$	Biunivocal
Order	Weak order among the objects	Ordinal	$a \succ b \iff m(a) \geq m(b)$	Monotone increasing
Difference	As above plus weak order among intervals	Interval	$\Delta_{ab} \succ \Delta_{cd} \iff m(a) - m(b) \geq m(c) - m(d)$	Linear positive $m' = \alpha m + \beta$ $\alpha > 0$
Extensive	As above plus a concatenation operation	Ratio	$a \sim b \circ c \iff m(a) = m(b) + m(c)$	Similarity $m' = \alpha m$ $\alpha > 0$

*Ordinal* scales permit a rank ordering of objects and remain invariant under *monotonic increasing* transformations. They include hardness of minerals and earthquake or wind intensity.

*Interval* scales entail a constant unit of measurement; that is, they introduce a metric, and so permit the calculation of differences between any two values. They remain invariant under *linear positive* transformations. Fahrenheit or Celsius temperatures are good examples, as well as position or time, intended as calendar.

*Ratio* scales also feature constant units of measurement, but, in addition, they allow the ratio of two values to be evaluated, because a true zero exists. They are invariant under any simply multiplicative transformation, or *similarity*. They include ‘extensive’ quantities, such as mass or length, but also, in Stevens’ view, perceptual quantities, such as loudness or brightness

So in order to overcome the position of the physicists, Stevens generalises the notion of measurement scale, already introduced by Campbell for fundamental measurements. Moreover, he argues that direct estimation of sensations is possible, as happens in *magnitude estimation*. Such a test may be performed, for example, by presenting a line of a given length and telling the observer to call it some number, say, 10. Then a line of some other length is presented and the subject is asked to assign it a number, considering that the first line was 10 and so forth. The important point is that, thanks to such tests, it is possible for Stevens to consider *equality between ratios, as the empirical relation for ratio scales, instead of addition*.

Summarising, Stevens proposes to overcome the severe limitation in measurability posed by the report of the British Association, by increasing the number of allowable measurement scales and by considering equality of ratio as an empirical relation. Yet his arguments did not convince the physicists.

With Stevens we have reached the second half of the 20th century. At that time a considerable body of results had been obtained in measurement theory and there was a need for a systematization, which was achieved with the representational theory of measurement.

## 2.2 The representational theory of measurement

### The representational framework

A remarkable systematisation of the formal theory of measurement was achieved in the second half of the 20th century and referred to as representational theory. A comprehensive presentation is offered in the gigantic treatise, *Foundations of Measurement*, by Krantz, Luce, Suppes, and Tversky [13], as well as in other parallel works, such as those by Roberts [16] and Narens [20]. These studies share a common framework, which essentially may be seen as a combination of the viewpoints of Campbell and Stevens, that are seen as complementary rather than opposing. The main idea, traceable, as we have seen, to Helmholtz, is that the numbers we obtain through measurement represent empirical relations. This framework also applies to fundamental physical measurements as intended by Campbell, here called extensive. But now *extensive* is regarded as *a special*, though very important, kind of measurement, not as *the only* one worthy of this name. Consequently, the classification of scales proposed by Stevens may be retained and each scale is now characterized by

1. A *representation theorem*, showing how empirical relations are mapped into corresponding numerical relations
2. A *uniqueness theorem*, specifying which class of transformations maintain the properties of the scale

The uniqueness theorem allows the meaningfulness of statements concerning measurement to be addressed. In fact, we may say that a statement concerning the results of measurement on a given scale, is meaningful if its truth is unaffected by admissible transformations on that scale.

A summary of the representation framework has been presented in Table 2 above. We have already discussed the invariance properties of the scales, here called uniqueness conditions, when presenting Stevens' contribution. Let us now briefly comment on empirical structures (column 1), the associated empirical relations (column 2), and representation theorem (column 4).

In *nominal structures* we only have the *equivalence* of elements belonging to the same class, denoted by  $\sim$  and the result of operating on this scale is a classification.

*Order structures* are characterised by an empirical relation of *weak order* that we denote by the symbol  $\succsim$ . The relation of weak order plays a fundamental role in measurement and is satisfied also in the other scales to follow.

In the case of *difference structures* we are mainly concerned with *intervals* of objects and with a weak order relation among them. For example, if  $a, b$  are two elements of  $A$ , their interval will be denoted by  $\Delta_{ab}$  (being positive if  $a \succ b$ ) and  $\Delta_{ab} \succsim \Delta_{cd}$  means that the first interval is, empirically, greater than or equal to the second one.

Finally, in *extensive structures* an empirical *concatenation* operation, or physical addition as Campbell named it, is present. In connection with this operation we also define a ternary relation  $a \sim b \circ c$ , meaning that the element  $a$  is equivalent to the empirical sum of  $b$  plus  $c$ .

The main representation theorem for the three structures we are dealing with thus reads as follows.

- For order structures:

$$a \succsim b \iff m(a) \geq m(b), \quad (59)$$

- For difference structures:

$$\Delta_{ab} \succsim \Delta_{cd} \iff m(a) - m(b) \geq m(c) - m(d), \quad (60)$$

- For extensive structures:

$$a \sim b \circ c \iff m(a) = m(b) + m(c). \quad (61)$$

Each structure includes the properties of the previous ones, so, for example, difference structures also satisfy the representation theorem that holds for order structures and so on.

The representational theory has been developed mainly in the field of behavioural sciences but has been brought to the attention of physicists and engineers since the 1970s, mainly by Finkelstein [19], and has received, afterwards, contributions also from that community. Such a theory was initially stated in an essentially algebraic fashion and, until very recently, it has only partially been treated in probabilistic terms. Moreover, due to its growth, especially in the field of behavioural science, little or no attention has been paid to the role of the measuring instrument (or system). We thus now briefly review some of the probabilistic developments proposed in the representational approach, leaving for the next section the presentation of a complete probabilistic theory of measurement that also accounts for the role of the measuring system.

## Probabilistic issues

Because uncertainty is a constitutive aspect of measurement, several attempts have been made to include probabilistic issues in the representational

framework. A basic bibliography may be found in Roberts [16], Chapter 6], Krantz *et al.* [13, Vol. 2, Chapters 16–17], and in Luce and Suppes [33]. Although we cannot survey these references in detail, we may say that they mainly deal with a probabilistic treatment of comparison tests aimed at constructing order scales. The general problem that they consider may be formulated in the following terms. Given a probabilistic description of the empirical relations, under which conditions is it possible to arrive at a meaningful representation? This issue is also called *probabilistic consistency*.

To illustrate the problem, suppose we want to measure some perceived quantity, for example, the intensity of a class of sounds, and we want to construct an order scale for them [37]. Empirical relations are in this case defined by the responses of a class of subjects to the sounds under investigation. If we consider any two sounds,  $a$ ,  $b$ , we cannot expect in general that a definite relation, for example,  $a \succ b$ , where  $\succ$  here means ‘louder than’, definitely holds for them, due to inter- and intrasubjects variability. We may rather expect that a probability may be attached to such a relation; that is,  $\mathbb{P}(a \succ b)$ , where  $\mathbb{P}$  denotes the probability of a relation.

What property should we require for this probability in order for representation to be possible? We have seen that for deterministic order scales the key property is transitivity; that is, for  $a$ ,  $b$ ,  $c$  belonging to  $A$ , if both  $a \succ b$  and  $b \succ c$  hold,  $a \succ c$  should also hold. A suitable probabilistic replacement for this is the so-called *weak probabilistic transitivity*; that is,

$$\mathbb{P}(a \succ b) \geq \mathbb{P}(a \prec b) \quad \text{and} \quad \mathbb{P}(b \succ c) \geq \mathbb{P}(b \prec c) \quad \implies \mathbb{P}(a \succ c) \geq \mathbb{P}(a \prec c). \quad (62)$$

If this property holds, the following representation theorem may be proved [27].

$$m(a) \geq m(b) \iff \mathbb{P}(a \succ b) \geq \mathbb{P}(a \prec b). \quad (63)$$

Yet results such as this one, although important in some regards, are not completely satisfactory from a general standpoint. In fact, this means that in order to assign measures to the objects we must know the probability of the empirical relations. Empirical relations are no longer regarded as deterministic, but the measures may still be assigned exactly, once the required probabilities are known. Measurement is still, in some sense, a deterministic process, although it is no longer based *directly* on empirical relations but rather, *indirectly*, on their probabilities.

This is not yet what is needed for a completely satisfactory probabilistic theory of measurement. The need for a complete probabilistic formulation of the representational framework was clearly expressed by Leaning and Finkelstein [18], where a probabilistic framework was also proposed. Unfortunately that approach was not developed in detail in the following years. The most famous paper on the probabilistic approach is perhaps one from Falmagne [17], in which the author presented a probabilistic representation for extensive measurements, in the case that some special constitutive relations hold between

relational probabilities and measure values. The paper was a novelty at that time and had the merit of dealing with the quite complex structure of extensive measurement, but it dealt with special cases only.

Actually, there is an inherent difficulty in achieving a complete probabilistic formulation, partially documented in the studies concerning the so-called *random-utility model* (see Roberts [16], Section 6.2). Another difficulty is related to the notion of the *probability of a relation* (and of a relational system), which has been scarcely investigated in the past [36]. Only relatively recently has a major contribution been provided on this subject in two papers, by Regenwetter [28] and Regenwetter and Marley [32], which provide useful results. We account for them, later on, in proposing a full probabilistic theory, but, before that, we have to discuss the role of the measuring system in a formal theory of measurement.

### The role of the measuring system in a formal theory of measurement

Although measuring instruments have been key players in the development of modern science, their role in the theory of measurement seems to have been underestimated [41]. Campbell, for example, as a physicist, was aware of their importance, yet he concentrated mainly on the problem of scale construction and considered the issues related to the measuring system as technical rather than theoretical ones. Lately, the formal theory of measurement has been developed mainly in the area of behavioural sciences, where the role of the measuring system is not felt as central. The need of explicitly accounting for the role of the measuring system in a theory of measurement has been pointed out by Gonella [21] and, more recently and with additional arguments, by Mari [29], who claims that measurement is essentially an evaluation performed by a calibrated MS. We essentially agree with this position, *provided that a proper (broad) definition of MS is adopted*. Let us then discuss this point in more detail.

In principle, measurement may be performed by selecting one object,  $a$ , and comparing it with a previously established measurement scale in order to identify a standard  $s$  *to which the element  $a$  is equivalent*. After that, we assign  $m(a) = m(s)$ . But how can we actually do that?

Measurement is in general performed *through the mediation of a measuring system*. So what is the role and the behaviour of the MS precisely?

In Section 1.4 we have observed that, because the value of the measurand is not directly observable, the function of the MS is to interact with the object and to produce, as a result of the interaction, an observable output, which is ‘caused’ by the measurand. From the observable output it is possible to go back to the cause and to properly assign a value to the measurand. We then propose to define the measuring system as an *empirical system* able to *interact* with any object carrying the quantity under investigation and to produce, as a result of the interaction, an *observable output*, on the basis of which it



is possible to assign a value to the measurand. Note that this definition is general enough to accomplish also measurements by a panel or a jury with respect to a previously defined reference scale, as often occurs in the case of perception [37]. Provided that this definition is accepted, how can we formally state this? In particular how can we characterise the interaction of the MS with the measurand object?

The solution is straightforward, because the property that we need is simply the following. *The output of the measuring system should depend only on the state of the quantity* (and thus should not depend on the specific object manifesting that state). The behaviour of the MS may thus be described and characterised by a function  $\varphi : A \mapsto \mathbb{R}$  such that, for each  $a, b \in A$ ,

$$a \sim b \Leftrightarrow \varphi(a) = \varphi(b). \quad (64)$$

The output of the MS does not depend on the specific object but only on the value of the measurand, thus another useful description is provided by the *calibration function*  $f$ , that we have informally introduced in Section 1.4. We may now provide a formal definition: let  $x$  be the value of the measurand and  $f : X \mapsto \mathbb{R}$  the calibration function; then

$$\varphi(a) = f[m(a)]. \quad (65)$$

The deterministic description of the measurement process is then still provided by the formulae (44–46). All this holds in a deterministic framework and thus describes the ideal measurement. Later on we present the corresponding probabilistic model, which is also in agreement with what we have anticipated in Section 1.4.

## 2.3 A probabilistic theory of measurement

### Probabilistic relations

We now briefly sketch a probabilistic theory of measurement that we have recently proposed and published [38].

The key point for attaining such a theory is the introduction of the notion *probability of a relation* [28, 32]. It is important to note first that the term ‘relation’ may be understood both in a general and in a specific meaning. For example, when we write  $a \succsim b$ , we mean that the relation  $\succsim$  holds for the couple  $(a, b)$  in  $A$  (this is the specific meaning). On the other hand, when we speak of the relation  $\succsim$  on  $A$ , we refer to the set of all the pairs of elements of  $A$  which satisfy it (this is the general meaning). Consider now this second, general, meaning. Then, a *probabilistic relation* of some kind, for example, a weak order, on a finite set  $A$ , may be defined by *assigning a probability distribution over the class of all possible relations of that kind on  $A$* . This is illustrated by the simple example in Table 3, concerning a set with only three elements,  $A = \{a, b, c\}$ .

**Table 3.** An illustrative example of a probabilistic order structure on  $A = \{a, b, c\}$ .

Relational System $\mathcal{A}_i$	Weak Order Relations $\succsim_i$	$x_a$	$x_b$	$x_c$	$\mathbb{P}(\mathcal{A}_i)$ (Example)
$\mathcal{A}_1$	$a \succ b \succ c$	3	2	1	0.2
$\mathcal{A}_2$	$a \succ c \succ b$	3	1	2	0.2
$\mathcal{A}_3$	$b \succ a \succ c$	2	3	1	0.0
$\mathcal{A}_4$	$b \succ c \succ a$	1	3	2	0.0
$\mathcal{A}_5$	$c \succ a \succ b$	2	1	3	0.0
$\mathcal{A}_6$	$c \succ b \succ a$	1	2	3	0.0
$\mathcal{A}_7$	$a \sim b \succ c$	2	2	1	0.1
$\mathcal{A}_8$	$a \sim c \succ b$	2	1	2	0.1
$\mathcal{A}_9$	$b \sim c \succ a$	1	2	2	0.0
$\mathcal{A}_{10}$	$a \succ b \sim c$	2	1	1	0.3
$\mathcal{A}_{11}$	$b \succ a \sim c$	1	2	1	0.0
$\mathcal{A}_{12}$	$c \succ a \sim b$	1	1	2	0.0
$\mathcal{A}_{13}$	$a \sim b \sim c$	1	1	1	0.1

In the table all the possible weak orders on  $A$  are listed (column 2) and a probability is assigned to each of them (column 6). It may be that the probability of some of them is null, but it is necessary that no relation which is *not* a weak order has a nonnull probability. Note that when we assign a probability to a weak order, say  $\succsim_i$ , we also formally assign it to the *order structure*  $\mathcal{A}_i = (A, \succsim_i)$ .

If we now consider, in this example, a specific relation holding for a specific pair of elements, for example,  $a \succ b$ , we may note that it is verified in  $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_5, \mathcal{A}_8$ , and  $\mathcal{A}_{10}$  and consequently its probability is

$$\mathbb{P}(a \succ b) = \mathbb{P}(\mathcal{A}_1) + \mathbb{P}(\mathcal{A}_2) + \mathbb{P}(\mathcal{A}_5) + \mathbb{P}(\mathcal{A}_8) + \mathbb{P}(\mathcal{A}_{10}) = 0.8.$$

In general, for each couple of elements  $a, b \in A$ , we may calculate the probability of the empirical relations  $a \succsim b$  as

$$\mathbb{P}(a \succsim b) = \sum_{a \succsim b \in \mathcal{A}_i} \mathbb{P}(\mathcal{A}_i). \quad (66)$$

What we have so far presented is a *probabilistic order structure*. In a similar way it is possible to define a probabilistic counterpart also for interval and extensive structures.

### The measurement scale

Thanks to the notion of probabilistic relation it is possible to propose a *probabilistic counterpart of the representation theorem*. Consider a finite set of objects  $A$  and either

1. A probabilistic order structure
2. A probabilistic interval structure
3. A probabilistic extensive structure

Then it is possible to assign a discrete random variable  $x_a$  to each element  $a \in A$  in such a way that (for each  $a, b, c, d \in A$ ):

1. For a probabilistic order structure:

$$\mathbb{P}(a \succ b) = P(x_a \geq x_b); \quad (67)$$

2. For a probabilistic difference structure:

$$\mathbb{P}(\Delta_{ab} \succ \Delta_{cd}) = P(x_a - x_b \geq x_c - x_d); \quad (68)$$

3. For a probabilistic extensive structure:

$$\mathbb{P}(a \sim b \circ c) = P(x_a = x_b + x_c). \quad (69)$$

Proof of this probabilistic representation theorem is provided in Reference [38]. Let us now illustrate formula (67) with the example in Table 3. In the table, for each  $\mathcal{A}_i$  (column 1), a proper assignment of values to the random variables is presented (columns 3–5). For example, when  $\mathcal{A}_1$  holds,  $x_a = 3$ ,  $x_b = 2$ , and  $x_c = 1$ . So it is possible to calculate the probability distribution for each random variable: for example, because  $x_a = 3$  in  $\mathcal{A}_1$  and  $\mathcal{A}_2$ ,

$$P(x_a = 3) = \mathbb{P}(\mathcal{A}_1) + \mathbb{P}(\mathcal{A}_2) = 0.4.$$

It is now easy to check that the  $\mathcal{A}_i$ s for which, say,  $a \succ b$  holds, namely  $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_5, \mathcal{A}_8$ , and  $\mathcal{A}_{10}$ , are the same for which also  $x_a > x_b$  holds and consequently

$$\mathbb{P}(a \succ b) = P(x_a > x_b) = 0.8.$$

In the accompanying DVD this example is studied in more detail and the related software is addressed.

## The measurement process

In the deterministic description of the MS (Section 2.2) we have assumed that the output of the MS does not depend upon the specific object selected, but only on its state, and from that we have deduced that the observation transformation can be described by a function  $y = f(x)$ , defining a unique relation between each value of the measurand  $x$  and the corresponding indication  $y$  of the MS. Here we maintain the hypothesis that the output of the MS does not depend upon the specific object selected, but only on its state, that is, on the specific value that it manifests when we make the measurement, but we assume that, for each such value, a plurality of indications is possible, governed by a probabilistic distribution. Consequently, as we have seen in Section 1.4, a probabilistic description of the MS may be obtained by considering a

conditional probability distribution which describes the observation phase, that is,

$$P(y|x). \quad (70)$$

Then the restitution phase follows, described by

$$P(x|y) = P(y|x) \left[ \sum_{x \in X} P(y|x) \right]^{-1}. \quad (71)$$

Lastly, the overall measurement process is characterised by

$$P(\hat{x}|x) = \sum_{y \in Y} \delta[\hat{x} - \mu(x|y)] P(y|x), \quad (72)$$

where  $\mu$  is a position parameter appropriate for the scale that we are considering (i.e.,  $\mu$  is the median, if the scale is ordinal, or the expected value, if the scale is interval or ratio) and  $\delta$  is the unitary discrete impulse. Note that the integrals appearing in Table 1 are now replaced by sums, because now we are dealing with discrete random variables. This hypothesis, anyway, is not critical, because it is essentially equivalent to requiring that the measuring system has a finite resolution (see [29, 38] for a discussion of this point). We omit, for the sake of brevity, the treatment of influence quantities that may be explicated as presented in Section 1.4. Lastly, the *calibration* of the MS may be now intended as the operation aiming at obtaining the conditional distribution  $P(y|x)$ . A summary of the main results of this probabilistic theory of measurement is provided in Table 4.

**Table 4.** Synopsis of the proposed theory: deterministic versus probabilistic approach.

The Measurement Scale		
Scale Type	Deterministic Approach	Probabilistic Approach
Order	$a \succ b \Leftrightarrow m(a) \geq m(b)$	$\mathbb{P}(a \succ b) = P(x_a \geq x_b)$
Interval	$\Delta_{ab} \succ \Delta_{cd} \Leftrightarrow$ $m(a) - m(b) \geq m(c) - m(d)$	$\mathbb{P}(\Delta_{ab} \succ \Delta_{cd}) =$ $P(x_a - x_b \geq x_c - x_d)$
Ratio	$a \sim b \circ c \Leftrightarrow m(a) = m(b) + m(c)$	$\mathbb{P}(a \sim b \circ c) = P(x_a = x_b + x_c)$
The Measuring Process		
Process	Deterministic Approach	Probabilistic Approach
Observation	$y = f(x)$	$P(y x)$
Restitution	$\hat{x} = f^{-1}(y)$	$P(x y) = P(y x) \left[ \sum_{x \in X} P(y x) \right]^{-1};$ $\hat{x} = \mu(x y)$
Measurement	$\hat{x} = f^{-1}(f(x)) = x$	$P(\hat{x} x) = \sum_{y \in Y} \delta[\hat{x} - \mu(x y)] P(y x)$

### 3 Final remarks

Probability and metrology are two closely linked disciplines.

The science of measurement has taken advantage of the development of probability and statistics and has assumed methods and tools. But at the same time it has also greatly contributed to the development of these disciplines: the early theory of errors is an outstanding example of this.

We think that both in the present and in the future measurement problems may be best faced not simply by looking for existing statistical methods to adopt, but rather by considering probability as the natural tool for dealing with matters in which determinism is not appropriate for providing a satisfactory description and explanation of facts. In this way the relationship within the two disciplines may be rich and fruitful and the dialogue between the related scientific communities intense and enriching for both parts.

Moreover we have seen how measurement science has developed according to two distinct mainstreams, because of the division which arose between scientists in physics and engineering on the one hand and in psychology and behavioural science, in the first half of the last century. Such a division has been, in our opinion, detrimental in many respects, because the two approaches, one based on the study of the measurement process, the other on the problem of measurability and of the construction of the measurement scale, naturally complement each other and are both necessary to attain a satisfactory overall theory of measurement. We have also seen how a new and unconventional way of using probability, the probability of relations, has been recently proposed and is extremely promising as it paves the way to a better foundation for measurement. So we believe in the possibility of a new foundation for a unique science of measurement, valuable for all domains of knowledge in which measurement is seen as a necessary tool for reinforcing knowledge and for gathering information, and we consider probability as the natural logic for such a science.

### Appendix: Symbols and Notation

Some of the main symbols are listed in Table 5. As a general criterion, we have tried to keep notation as lean as possible. Due to the broadness of the subject, some symbols are polysemantic: we have preferred to establish an easy connection between similar ideas, rather than resorting to a wide mass of difficult-to-relate symbols. We adopt the usual convention of denoting vectors and matrices by bold characters. We do not use special notation (such as capitals or bold) for random variables. So the same symbol may be used for denoting a random variable as well as its specific value. For example, the probability distribution of  $v$  may be denoted either as  $p_v(\cdot)$  or, in a shorthand notation, as  $p(v)$ .

**Table 5.** List of the main symbols of Sections 1 and 2, respectively.

$x, \mathbf{x}$	parameter(s) to be estimated, measurand (either scalar or vector)
$y, \mathbf{y}$	observation(s), instrument indication(s)
$f, \mathbf{f}$	scalar function, vector function
$v, \mathbf{v}$	random errors affecting the indications of a measuring system
$\max$	operator that calculates the maximum of a function
$\mathbf{A}$	matrix
$E, Var$	expectation operator, variance operator
$\sigma, \sigma^2$	standard deviation, variance
$N, n, m$	integers
$\theta, \boldsymbol{\theta}$	influence parameter(s) producing systematic effects
$p$	probability density function, also called probability distribution
$P$	probability function, discrete probability distribution
$p_{t,\nu}$	$t$ -Student probability density function, with $\nu$ degrees of freedom
$\hat{x}$	the “hat” symbol indicates an estimator or an estimated value; if applied to the measurand it denotes the measurement value
$\bar{y}, \tilde{y}$	arithmetic mean of $y$ , mean value of $y$
$u, U$	standard uncertainty, expanded uncertainty
$g$	function appearing in the GUM evaluation formula
$z, \mathbf{z}$	corrections of influence quantities
$A$	set of objects manifesting the characteristic $x$
$m$	measure function, $m : A \mapsto \mathbb{R}$
$\succsim$	empirical weak-order relation on $A$ , empirical weak order relation between intervals
$\succsim_i$	$i$ th empirical weak-order relation definable on $A$
$\Delta_{ab}$	interval between elements $a$ and $b$ of $A$
$\circ$	binary empirical operation of concatenation (i.e., empirical sum) of elements of $A$
$\sim$	empirical equivalence relation on $A$ defined, for $a, b \in A$ by $a \sim b \Leftrightarrow (a \succsim b) \text{ and } (b \succsim a)$
$\mathcal{A} = (A, \succsim)$	empirical order system on $A$
$\mathcal{A}_i = (A, \succsim_i)$	$i$ th empirical order system on $A$
$\mathbb{P}$	probability function whose argument is a relation or a relational system (such as an order system)
$X$	set of the possible values of the measurand, image of $A$ in $\mathbb{R}$ , through the measure function $m$ ; that is, $X = m(A)$
$Y$	set of the output values (indications) of the measuring system
$\mu$	position parameter of a probability distribution (e.g., expected value or median)
$\varphi, f$	characteristic functions of the measuring system: $\varphi : A \mapsto Y, f : X \mapsto Y$ ; <ul style="list-style-type: none"> <li>the property characterising <math>\varphi</math> is: <math>\forall a, b \in A, a \sim b \Leftrightarrow \varphi(a) = \varphi(b)</math></li> <li>the function <math>f</math> is defined by: <math>\forall a \in A, y = \varphi(a) = f[m(a)]</math></li> </ul>
$g$	measurement function, $g : X \mapsto X$ , defined, for $a \in A, x = m(a)$ , by: $\hat{x} = g(x) = f^{-1}[f(x)] = x$ , and $m(a) = g[m(a)] = g(x) = x$

## References

1. Bayes, T.: An essay towards solving a problem in the doctrine of chances. *Phil. Trans. R. Soc. London*, **53**, 370-418 (1763)
2. Laplace, P. S.: Mémoire sur la probabilité des causes par les événements. *Mem. Acad. R. Sci.*, **6**, 621-656 (1774)
3. Gauss, C. F.: *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Hamburg (1809) English translation by Davis, C. H., Dover (1963) reprinted 2004
4. Laplace, P.S.: *Theorie analytique des probabilités*, Courcier, Paris, (1812) In: *Oeuvres Complètes de Laplace*. Gauthier-Villars, Paris, vol. VII
5. Gauss, C. F.: *Theoria combinationis observationum erroribus minimis obnoxiae*. Gottingen, (1823) English translation by Stewart, G.W., SIAM, Philadelphia (1995)
6. von Helmholtz, H.: Zählen und Messen Erkenntnis – theoretisch betrachtet, Philosophische Aufsätze Eduard Zeller gewidmet. Fuess, Leipzig (1887)
7. Campbell, N. R.: *Physics - The elements*. (1920) *Representations as Foundations of science*. Dover, New York (1957)
8. Fisher, R. A.: *Statistical methods for research workers*. Oliver and Boyd, Edinburgh (1925)
9. Ferguson, A., Myers, C. S., Bartlett, R. J.: Qualitative estimates of sensory events. Final Report – British Association for the Advancement of Science, **2**, 331-349 (1940)
10. Stevens, S. S.: On the theory of scales and measurement. *Science*, **103**, 667-680 (1946)
11. Fisher, R. A.: *Statistical methods and scientific inference*. Oliver and Boyd, Edinburgh (1956)
12. Mandel, J.: *The statistical analysis of experimental data*. Wiley (1964) Repr. Dover (1984)
13. Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A.: *Foundations of Measurement*. Vol 1–3, Academic Press, New York (1971–1990)
14. Sheynin, O. B.: Laplace's theory of errors. *Archive for history of exact sciences*, **17**, 1–61 (1977)
15. Sheynin, O. B.: C. F. Gauss and the theory of errors. *Archive for history of exact sciences*, **20**, 21–72 (1979)
16. Roberts, F.S.: *Measurement theory*. Addison-Wesley, Reading, MA (1979)
17. Falmagne, J.C.: A probabilistic theory of extensive measurement. *Philosophy of Science*, **47**, 277–296 (1980)
18. Leaning M.S., Finkelstein, L.: A probabilistic treatment of measurement uncertainty in the formal theory of measurement. In: Streker, G. (ed) *ACTA IMEKO* 1979. Elsevier, Amsterdam (1980)
19. Finkelstein, L., Leaning, M.S.: A review of the fundamental concepts of measurement. *Measurement*, **2**, 25–34 (1984)
20. Narens, L.: *Abstract measurement theory*. MIT Press, Cambridge (1985)
21. Gonella, L.: Measuring instruments and theory of measurement. In: *Proc. XI IMEKO World Congress*, Houston, (1988)
22. Press, S.J.: *Bayesian statistics*. Wiley, New York (1989)
23. Costantini, D., Garibaldi, U., Penco, M. A.: *Introduzione alla statistica- I fondamenti dell'argomentazione incerta*. Muzzio, Padova (1992)

24. Weise, K., Wöger, W.: A Bayesian theory of measurement uncertainty. *Measurement Science and Technology*, **4**, 1–11 (1993)
25. BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML: International Vocabulary of Basic and general terms in Metrology. Second Edition (1994)
26. BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML: Guide to the Expression of Uncertainty in Measurement. ISO, Geneva (1995)
27. Michellini, R.C., Rossi, G.B.: Measurement uncertainty: a probabilistic theory for intensive entities. *Measurement*, **15**, 143–157 (1995)
28. Regenwetter, M.: Random utility representations of finite m-ary relations. *Journal of Mathematical Psychology*, **40**, 219–234 (1996)
29. Mari, L.: Beyond the representational viewpoint: a new formalization of measurement. *Measurement*, **27**, 71–84 (2000)
30. Monti, C. M., Pierobon, G.: *Teoria della probabilità*. Zanichelli, Bologna (2000)
31. Hacking, I.: An introduction to probability and inductive logic. Cambridge Press, Cambridge (2001) Italian edition: *Il Saggiatore*, Milano (2005)
32. Regenwetter, M., Marley, A.A.J.: Random relations, random utilities, and random functions. *Journal of Mathematical Psychology*, **45**, 864–912 (2001)
33. Luce, R.D., Suppes, P.: Representational measurement theory. In: Stevens' *Handbook of Experimental Psychophysics*. Vol. 4, Wiley, (2002)
34. Lira, I.: *Evaluating the measurement uncertainty*. IOP, Bristol (2002)
35. Rossi, G.B.: A probabilistic model for measurement processes. *Measurement*, **34**, 85–99 (2003)
36. Costantini, D.: *I fondamenti storico-filosofici delle discipline statistico probabilistiche*. Bollati Boringhieri, Torino (2004)
37. Rossi, G.B., Crenna, F., Panero, M.: Panel or jury testing methods in a metrological perspective. *Metrologia*, **42**, 97–109 (2005)
38. Rossi, G.B.: A probabilistic theory of measurement, *Measurement*, **39**, 34–50 (2006)
39. Rossi G.B., Crenna F., Cox M.G., Harris P.M.: Combining direct calculation and the Monte Carlo Method for the probabilistic expression of measurement results. In: Ciarlini, P., Filipe, E., Forbes, A.B., Pavese, F., Richter, D. (eds) *Advanced Mathematical and Computational Tools in Metrology VII*, World Scientific, Singapore (2006)
40. Rossi, G.B., Crenna F.: A probabilistic approach to measurement-based decisions. *Measurement*, **39**, 101–119 (2006)
41. Rossi, G.B.: Measurability. *Measurement*, **40**, 545–562 (2007)
42. Cox, M.G., Rossi, G.B., Harris, P.M., Forbes, A.: A probabilistic approach to the analysis of measurement processes, *Metrologia*, **45**, 493–502 (2008)



Data Modeling for Metrology and Testing in  
Measurement Science

Pavese, F.; Forbes, A.B. (Eds.)

2009, XVIII, 486 p. 111 illus., Hardcover

ISBN: 978-0-8176-4592-2

A product of Birkhäuser Basel