Probabilistic inferences related to the measurement process

Giovanni Battista Rossi
Università degli Studi di Genova - DIMEC, Genova, Italy
E-mail: gb.rossi@dimec.unige.it

Abstract. In measurement indications from a measuring system are acquired and, on the basis of them, some inference about the measurand is made. The final result may be the assignment of a probability distribution for the possible values of the measurand. We discuss the logical structure of such an inference and some of its epistemological consequences. In particular, we propose a new solution to the problem of systematic effects in measurement.

1. Introduction
We consider two main questions:

- What statistical-probabilistic inferences are related to the measurement process?
- How to deal with systematic effects?

The first is motivated by some increasing interest in inferences in measurement, within the current debate on measurement uncertainty [9, 11, 16] (note that references are listed in a chronological order, to give a feeling of the historical development of the subject).

The second is a classical problem, traceable to Gauss [1, 18], for which solutions have been proposed only quite recently [6, 8, 11, 13].

For addressing such questions, we start by considering, in section 2, statistical-probabilistic inferences in general and, following authoritative references [4, 14], we group them in two main classes, deductive and inductive, and we further parse the latter in two subclasses, depending upon the assumptions made.

Then, in section 3, we briefly review the problem of systematic deviations in measurement, by recalling some issues of the historical debate about them.

In section 4 we start a systematic examination of the measurement process by briefly recalling a general probabilistic model for it [13, 16].

In section 5 we dwell in a thorough investigation of the logic steps of the measurement process and in particular we show that, if systematic deviations are described in probabilistic terms, as recommended by current good practices [8], measurement is not reducible to a purely inductive process. Here a difficulty arises, from an epistemological standpoint, since measurement is considered, in science, the key way for learning from experience [12]. This is another, more modern, way for formulating the problem of systematic deviations in measurement.

Lastly, in section 6, we propose a solution to this problem, which is, in our opinion, more satisfactory than the others presently known, and thus provides a sounder foundation to the practice of treating systematic effects in probabilistic terms.
2. Probabilistic models and inferences

In our experience, there are two main ways of working with probability in a research environment: developing models and making inferences. Although often related, these two processes are logically different and their difference is important.

A model may be understood as an abstract system aiming at representing, from some standpoint and up to some limitations, a real system, or a class of real systems. In particular, a probabilistic model is one that includes random variables, or random relations, or both of them.

A probabilistic inference, instead, concerns a real system: assumptions are made, data are acquired and, on the basis of both of them, probabilistic statements about the system under investigation are derived.

Let us illustrate these ideas with simple examples.

As a simple example of a probabilistic model, consider the case of coin tossing. Such an experiment is based on repeated independent executions of an elementary experiment, which has just two possible outcomes, heads, \( A \), or tails, \( \bar{A} \).

The model for the elementary experiment is simply

\[
P(A) = p, \quad P(\bar{A}) = q = 1 - p.
\]

Then, if \( y \) is the \( N \)-dimensional vector of the outcomes of \( N \) trials - the elements of \( y \) equal either \( A \) or \( \bar{A} \) -, the model for the overall experiment is

\[
P(y) = p^{n_A} (1 - p)^{N - n_A}, \tag{1}
\]

where \( n_A \) is the number of \( A \)-outcomes in \( y \). From a different standpoint, if we regard \( n_A \) as a random variable, its probability distribution is given by

\[
P(n_A) = \binom{N}{n_A} p^{n_A} (1 - p)^{N - n_A}. \tag{2}
\]

As we can see from this simple example, a model is just a mathematical description of a class of phenomena and does not strictly imply to perform any specific experiment.

A statistical-probabilistic inference, instead, concerns the way of deducing consequences, to be stated in terms of probabilities, from a set of observations and a set of assumptions. Here the experiment must be performed and the goal is to learn from data.

According to authoritative references \[4, 14\], such inferences may be grouped in two main classes, deductive and inductive.

Hypothetic deductive inferences include significance tests \[2, 4, 14, 18\]. In these tests a probabilistic model - usually consisting in a probability distribution - is assumed for describing the behavior of a real system: this is the hypothesis. Thanks to the model, the behavior of the system in some given situation may be predicted, in probabilistic terms: this is the deduction. Then an experiment is performed reproducing such a situation and the real outcome of the system is observed and compared with the prediction. If there is agreement between prediction and observation the model (the hypothesis) is maintained, otherwise it is rejected.

Let us demonstrate this in the coin tossing experiment. Suppose that we have one specific coin - this one! - and that, after carefully observing it, we assume the additional hypothesis of symmetry. Then we can fix the needed probability \( P(A) = p = p_0 = 1/2 \) and the model, for the coin under investigation, is now totally specified and may be expressed by a probability distribution for \( n_A \), given by formula (2), with \( p = 1/2 \). For a number of trials \( N = 10 \) the possible values for \( n_A \) are the integer numbers from 0 to 10, that constitute the space of the outcomes. Then we divide the space of outcomes in two subspace: \( R \), where the deviations from the expected value are not significant, ad \( \bar{R} \), where they are significant. For example, we may choose \( R = [2, 8] \). Lastly, we have to perform the experiment, with the real coin, and observe the result. If we obtain, e.g., \( n_A = 9 \), we reject our hypothesis, that is we conclude that

\[
P(A) \neq 1/2. \tag{3}
\]
Note that this is a probabilistic statement, although expressed by an inequality.
To sum up, let us outline the logical structure of a hypothetic deductive probabilistic inference. We may summarize such a procedure in the following steps.

a1) We hypothesise a probabilistic model,

a2) on the basis of which we deduce the probability distribution for the observations in a given experiment, which allows us

a3) to define an acceptance region $R$ for the observation, that is a region in which the observation complies with the model;

a4) then we perform the experiment and acquire $y$:

a5) if $y$, or some function of it, as $n_A$ in our example, falls into the acceptance region, the model is corroborated, otherwise it is falsified by the observation and we may consider abandoning it.

In an inductive inference, instead, the probability distribution is not previously assumed, but is inferred - induced - from data. This inference, in turn, may be done in two main ways, either by assuming some probabilistic data-producing mechanism, or by considering the way we learn from experience. The former strategy is typical of Bayesian inference [7]: since it hypothesises some kind of data-producing mechanism we will call it hypothetic inductive; the latter instead is typical of a predictive approach [5] and so we will call it a predictive inference [14].

Continuing with our example, suppose that we want to estimate the parameter $p$. We may do this by using the celebrated Bayes-Laplace rule, which allows assigning a probability the the possible causes of some observation. It reads

$$ P(C|E) = \frac{P(E|C)P(C)}{P(E)} \propto P(E|C)P(C), \quad (4) $$

that is to say that the probability of one possible cause, $C$, given an observed effect, $E$, is proportional to the product of the probability of the effect given that cause, times the probability of the cause.

The key point in this inference is to assume some model that describes the relation between the parameter $p$, the cause, and the observable effects, here the possible values of $n_A$. One such a model is again provided by the previously established Bernoullian frame (formula 2). We rewrite it as

$$ P(n_A|p) = \binom{N}{n_A} p^{n_A}(1-p)^{N-n_A}, \quad (5) $$

outlining the fact that such a model actually provides the sought link between $p$ and $n_A$. Note that the right side of formula 5 is to be interpreted as a function of both $n_A$ and $p$. By applying the Bayes-Laplace rule, we obtain

$$ P(p|n_A) \propto p^{n_A}(1-p)^{N-n_A}, \quad (6) $$

where now the right side of the equation is a function of $p$, with $n_A$ fixed and corresponding to the result of our experiment. This yields a Beta distribution. If we find, as above, $n_A = 9$, the corresponding expected value of $p$ is

$$ \hat{p} = \frac{n_A + 1}{N + 2} = 0.83. \quad (7) $$

Note that the same model, the Bernoullian one, has been used in these two inferences, but in two different ways. So, the difference between a model and an inference should now be clear: a model is a general description of a class of phenomena, which may, or not, be used to support
inferences concerning actual manifestations of such phenomena; an inference instead is a process in which we learn from data, using (or not, as we will show in the following) a model.

Again it is important to elicit the logical structure of this inference, which may be expressed in the following steps.

b1) We hypothesise a probabilistic relation, in a given experiment, linking the observation $y$ to a parameter $x$, expressed by a conditional distribution $P(y|x)$;
b2) we perform the experiment and acquire the observation $y$;
b3) on the basis of the observation and of the hypothesised probabilistic relation, we assign (induce) a probability distribution to $x$.

Lastly, let us briefly touch on predictive inferences. Here the goal is to assign a probability to the possible future observations, on the basis of the past observations. In our example we look for $P(A|y)$ and $P(\bar{A}|y)$. In contrast to the hypothetic inductive approach, here we do not assume any model for the process, rather we make assumptions on the way we learn from observations. The basic idea of this approach may be traced to Carnap [3]; a very nice introduction to the philosophy of predictivism may be found in [5].

With reference to our simple example, by assuming some exchangeability (the order of the observations is not relevant) and invariance (the roles of $A$ and $\bar{A}$ are exchangeable) requirements it is possible to obtain a so-called lambda-representation, which reads

$$P(A|y) = \frac{\lambda}{\lambda + N} p_0 + \frac{N}{\lambda + N} \frac{n_A}{N},$$

that is to say that the predictive probability is a weighted mean between the initial probability, $p_0$, and the relative frequency, $n_A/N$, where $\lambda$ is the weight of the initial probability. For example, considering our previous numerical example, if we obtain $n_A = 9$ and we assume $\lambda = 10$, we obtain

$$P(A|y) = 0.7.$$ (9)

The logical sequence of steps for this kind of inference is as follows.

c1) We assume some properties that characterise the way we learn from experience, such as exchangeability and invariance conditions,
c2) we perform the experiment and acquire the observation $y$,
c3) on the basis of the observation we assign a probability distribution to the possible future outcomes.

3. Systematic effects in measurement
We are now almost ready to discuss the logic of the measurement process. But prior of doing that, let us review, very briefly, some issues in the problem of systematic effects in measurement.

The fundamental contribution by Gauss needs first mentioning [1]. "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." 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 characterizes the classical theory of errors and may be its main limitation [18].
In the first half of the nineteenth century the "orthodox" school of statistics (a school that refer's mainly to Fisher’s work - the term "orthodox" is due to Jaynes [14]) contributed with the analysis-of-variance (ANOVA) method, by showing how to turn some systematic effects into random, by properly extending the experiment and how to separate the different contributions to uncertainty [2, 4, 18].

Yet, in the second half of the nineteenth century the international community of metrologists recognised that, in general, it is not possible to assume that all systematic effects have been eliminated or turned into random, and some residual systematic effect needs to be included in the overall uncertainty budget. Finding a common way for dealing with uncertainty sources of a different nature was one of the main challenges to be faced in the preparation of the Guide to the expression of uncertainty in measurement [8], which recommended to treat all uncertainty sources as random variables. But how to justify such a - though very reasonable - approach from the theoretical side? We just mention here the contribution by Wöger [6], who discussed the rationale of acting according to the principle of maximum entropy for assigning probability distributions to systematic deviations.

Yet, from an epistemological standpoint, a problem still exists, as we will show in the following, where we will discuss in depth the logical and epistemological implications of systematic effects and we will propose a new solution, hopefully more satisfactory than those currently available. But prior of doing that, we have to introduce a model of the measurement process.

4. A probabilistic model of the measurement process

Following the pattern proposed in section 2, we need a general model of the measurement process, for which we will use the one originally published in [13]. A gentler, although somewhat less general introduction to this model may also be found in [16]. For the reader’s convenience, we briefly recall the essential features and the main formulae to be used later on; the reader already familiar with this model may obviously skip this section without loss of continuity.

In this model, measurement is viewed as the concatenation of two phases,

- observation, in which the object to be measured is inputted to the measuring system that produces indications, and
- restitution, where the measurement value is obtained on the basis of instrument indications, thanks to a previously established calibration function.

If we denote the measurand (the quantity to be measured) by \( x \), the vector of instrument indications by \( y \) and the vector of influence quantities or parameters by \( \theta \), observation may be described by the conditional distribution

\[
P(y|x, \theta)
\]

and restitution by the probabilistic inversion of observation, obtained through the Bayes-Laplace rule, followed by a probabilistic de-conditionning from influence quantities/parameters (see [16], in particular, for a careful discussion of this point). This yields

\[
P(x|y) = \sum_{\theta} P(y|x, \theta) \left( \sum_{x} P(y|x, \theta) \right)^{-1} P(\theta).
\]

A probabilistic description of the overall measurement process may be obtained by combining the two transformations above. After introducing a proper position parameter for the final distribution, \( \mu(x|y) \), which in many cases may be identified with the expectation, we obtain

\[
P(\hat{x}|x) = \sum_{y} \delta[\hat{x} - \mu(x|y)] \left( \sum_{\theta} P(y|x, \theta) P(\theta) \right).
\]
Apart from details, it is very important to note that this model is intended to describe the measurement process as it is performed when good practices are followed. It is thus intended to have a "descriptive" rather than "normative" character.

The reader is invited to check this model with respect to her/his experience in measurement: any feedback will be welcome.

5. Logic of the measurement process

Let us now discuss the logic of the measurement process and in particular of the restitution phase. After examination of formula 11, it quite easy to recognize that restitution includes a hypothetic-inductive inference. Indeed, thanks to the model of the observation phase (the hypothesis), on the basis of the instrument indications, we assign a probability distribution to the measurand (induction). We may observe that, in doing so, we learn from data in a few ways: we learn were to locate the final distribution, we may obtain a more concentrated distribution by averaging over repeated observations and we may also obtain the variance of the distribution from the data [16].

We may thus say that restitution includes a hypothetic-inductive inference.

But it would not be correct to say that restitution is reducible to a hypothetic-inductive inference.

In fact, suppose that $\theta$ represents a systematic deviation: then in order to implement formula 11 we must assign a probability distribution to it, $P(\theta)$, and then to use it to obtain the final distribution. Yet our final knowledge about $\theta$ will be equal to the initial one, i.e., $P(\theta|y) = P(\theta)$, which implies that we do not learn from data about $\theta$. So $\theta$ is not subject to a hypothetic inductive inference, rather it is treated, necessarily, in a purely deductive way: a probability distribution for it is assumed and consequences are derived, without any interaction with the data.

To probe this further, let us outline the logical structure of the measurement process. It includes the following steps:

- d1) to assume a probabilistic relation between the value of the measurand and the indications of the measuring system, parametrical in respect to some influence parameters: this relation is a model of the observation process;
- d2) to assume a probability measure over the space of the influence parameters;
- d3) to perform observation and acquire the indications of the measuring system;
- d4) to apply, in the restitution phase, the Bayes-Laplace rule and obtain a probability distribution for the measurand, still conditioned upon the influence parameters;
- d5) to de-condition 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 d1, d3 and d4 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 d2 and d5 are not typical of a Bayesian inference. They include the assumption of a probability distribution for some parameters (step d2) and their use according to the rules of the calculus of probability (step d5). We say that these two steps form a hypothetic deductive process: 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.

The presence of a deductive process associated to measurement is undoubtedly a problem, since measurement is normally intended as the way for learning from experience par excellence [12]. So, how can we accept that measurement includes a deductive process? How can we deal with this problem?
We will provide an answer to these questions after reviewing some general epistemological principles.

6. Epistemology of the measurement process
6.1. Scientificity and falsifiability
Modern science, according to Antiseri [17], may be characterised as a kind of knowledge which is public, progressive and controllable. Controllability is most important for us and is the core of the scientific method, traceable up to Galileo Galilei [10].

The control of scientific theories, in principle, may be done by performing an experiment whose result is predicted by the theory and by checking whether the actually observed result is in agreement with the predicted one. If this is the case, we say that the theory has been corroborated by the experiment, otherwise we say that it has been falsified, and it should be rejected.

The possibility for some theory of being controlled by experiments whose results may falsify it, that is its falsifiability, is, according to Popper [17], an essential requirement for asserting the scientificity of the theory.

This requirement becomes critical when uncertainty comes into play. In fact a probabilistic model in general cannot be falsified by a single observation, or even by a single set of observations, since a probabilistic model is unable to predict exactly what will happen. This point has been raised to deny the scientificity of probabilistic theories and models [14]. Yet probability plays nowadays such a fundamental role in science that it would quite hard to confine it outside science. As an extreme opposite stance, Jaynes even claims the probability is the logic of science. So how can we deal with the falsifiability issue? A very agreeable position, in our opinion, is the one supported by Costantini [14], that probabilistic theories and model may be falsified by hypothetic-deductive inferences. Actually, even from our brief presentation in section 2, it should be apparent that the logical structure of such inferences strictly follows the classical pattern of the scientific method, that we have just briefly recalled. Although there is some conventionality in the choice of the acceptance region, hypothetic-deductive inferences really act as controllers of probabilistic models.

6.2. Epistemological status of statements concerning systematic effects in measurement
Let us now turn back to the problem of systematic effects in measurement. As we have already mentioned, systematic effects may be treated in probabilistic terms, but this implies that the restitution phase of the measurement process includes a deductive process, which is not controlled by the data acquired during the measurement process. How can we then ensure the scientificity of the overall measurement process? We suggest that the general principles that we have just reviewed may be applied not only to control general theories and models in science, but they may be also applied to ensure good scientific and technical practice.

So we propose to do that for measurement. Then the answer 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 the conditional distribution $P(\hat{x}|x)$, defined in formula 12. 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 the probabilistic model to be checked is the distribution $P(\hat{x}|x_0)$, which now is fully specified and is a function of $\hat{x}$ only. 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 $R = [a, b]$, such that

$$\sum_{\hat{x}=a}^{b} P(\hat{x} - x_0 | x_0) = \alpha$$

and we have to check whether

$$\hat{x}_0 \in R$$

or not. This procedure may be implemented in different ways, that are part of a good practice in measurement. They include, e.g., the verification of the calibration of a measuring system or the control of a measurement process by check standards. Even inter-comparisons, including key comparisons, may be considered in this perspective, although they require a more complicated mathematical model that can not be examined here.

To sum up, we suggest that the probabilistic treatment of systematic deviations may be considered a sound scientific approach, since the hypotheses implied in it may be, in general, at least in principle, checked by an auxiliary experiment that we call measurement verification and that consists in a hypothetic-deductive inference.

Acknowledgments
The author is grateful to Professor Domenico Costantini, formerly chairman of the Institute of Statistics of the University of Genova, for many exciting discussions on the philosophy of probability.

This work has been partially funded through a grant by the Italian Ministry of Education and research.

References
[1] Gauss C F 1823 *Theoria combinationis observationum erroribus minimis obnoxiae* (Gottingen) English translation by Stewart G W, 1995 (Philadelphia: SIAM)
[2] Fisher R A 1956 *Statistical methods and scientific inference*(Edinburg: Oliver and Boyd)
[3] Kemeny J G 1963 *The philosophy of Rudolf Carnap* ed Schilpp (London: Cambridge University Press) p. 711
[4] Hacking I 1965 *Logic of statistical inference* Digitally reprinted in 2009 (Cambridge: Cambridge University Press)
[5] Geisser S 1980 *Bayesian analysis in econometrics and statistics* ed A Zellner (Amsterdam: North Holland)
[6] Wöger W 1987 *IEEE Trans. Instrumentation and Measurement* 36 655
[7] Press S J 1989 *Bayesian statistics* (New York: Wiley)
[8] BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML 1993 *Guide to the expression of uncertainty in measurement* (Geneva: ISO)
[9] Estler W T 1999 *CIRP Annals* 48 1
[10] Rossi P 2002 *La nascita della scienza moderna in Europa* (Torino: Laterza)
[11] Lira I 2002 *Evaluating the measurement uncertainty* (Bristol: IOP)
[12] Mari L 2003 *Measurement* 34 17
[13] Rossi G B 2003 *Measurement* 34 85
[14] Costantini D 2004 *I fondamenti storico-filosofici delle discipline statistico-probablistiche* (Torino: Bollati Boringhieri)
[15] Rossi G B 2006 *Measurement* 39 34
[16] Cox M G, Rossi G B, Harris P M and Forbes A 2008 *Metrologia* 45 493
[17] Reale G and Antiseri D 2008 *Storia della filosofia* (Milano: Bompiani)
[18] Rossi G B 2009 *Data modeling for metrology and testing in measurement science* ed F Pavese A Forbes (Boston: Birkhauser - Springer) p 31