Classical analogues of quantum paradoxes

Henryk Gzyl

Abstract

In this note we discuss a few simple classical (as opposed to quantum) prediction problems. The thrust of this work is to examine the predictive role of probability theory and we shall see that some situations are not really paradoxical. In particular we want to separate the role of probability as a predictive tool from one of the basic sources of strangeness in quantum mechanics, namely the principle of superposition of states. For that we shall re-examine some of the quantum paradoxes, recast as a problem about making a prediction about the result of a random experiment in classical physics when some information is known.

1 Preliminaries

There is a vast, academic and non-academic, which does not mean non-serious, popular literature, about the strangeness of quantum phenomena and how this strangeness forces us to re-examine our image or our paradigms about the material world. A very small sample is [A],[C],[d'E], [F],[K],[KN],[M],[T] and the fabulous collection [WZ]. But no item in this collection separates the two issues mentioned in the abstract, namely, that the strangeness in quantum theory does not come from its being a curious mathematical model about phenomena that seem to be intrinsically random, but from the inner structure of the mathematical model itself. In quantum mechanics the strangeness does not come from its being a probabilistic theory, but from a combination of two facts: on one hand the superposition principle (implicit in the fact that states are modeled by vectors over the field of complex numbers), and on the other hand, from the fact that probabilities are obtained from the absolute values of the components of the vector describing the states.

What we are going to do in this note is to examine some problems in which a prediction has to be made about a classical experimental setup that imitates and parallels the setup associated to some of the quantum paradoxes. Thus in spirit we are close to Mermin's [Me]. Our emphasis is on examining in **what does the process of prediction consist of**, and **what is the influence of the available information** on it.

In a lengthy (but necessary for making this note self contained) appendix we recall the very basic ingredients of mathematical probability theory. The readers familiar with that material, may skip it, but those that have never heard of it, must read it. It is basically a collection of standard definitions accompanied by an explanation for their need. The important part of that section is the description of the predictive tool in probability theory, namely the concept of conditional expectation, and the formal use of the concept of σ-algebras as carriers of information. The content of each section is described below, but the message is that a predictor is described by a conditional expectation, which is a random variable whose value depends on information provided by a measurement, and once the measurement is made, the predictor provides us with a prediction.

In section 2 we consider an alternative, simple version of the two slit experiment. Instead of two slits, we have two point light sources. Here the randomness will come from the fact that we do not know which source is on, and we have to predict the observed signal.

In section 3 we shall consider the classical version of Schroedinger's cat and in section 4 the classical version of the EPR paradox (after Einstein, Rosen and Podolsky). Both in the first and last example the role of prior information role will be important for completing the prediction of the outcome. In section 5 we present a proof of the proof of Bell's inequalities, just to emphasize that there is nothing in them having to do with quantum phenomena. See [K] for a glimpse at the vast literature in this area.

We close this section with a few words about **random variables**. The formal definition is given in section 6, but an understanding of the concept is essential for following section (2)-(5). A random variable is the mathematical object (a function, to be specific) designed to model the result of a measurement in an experiment, and an experiment can be thought of as a sequence (finite, infinite, discrete or continuous, of measurements). The values taken by the random variable, are identified with the results of possible measurements of the variable, but the actual mathematical framework chosen does not have anything to do with the actual measurement process. A couple of example will clarify the issue. For example, to describe the results of the toss of a coin we need functions taking values in the set $\{H, T\}$, or to describe life-times, we shall use variables taking values in $[0, \infty)$, regardless of how toss the coin or how we measure the life-times.

But more important, a random variable is a function which stands for a collection of values. This issue is related to the question: In what state is the coin? Or, has the particle decayed or not?. The state is known after the coin or the particle are observed. This is the exact counterpart of the fact that a function is a collection of values taken at points and should not be confused with the values of the function at specific points. This is essential when interpreting conditional expectations, which are random variables, whose values are the best predictors of some variable when another variable is observed.

2 The two sources experiment

To avoid dealing with waves behind diffracting holes, we shall consider an experimental setup in which we have two sources emitting continuously monochromatic light of frequency ν, located at $\mathbf{x}_{\pm} = (0, 0, \pm h)$. An observer at $\mathbf{x}$ would see a signal $u_{\pm}(\mathbf{x}) = \frac{e^{i\nu\|\mathbf{x}-\mathbf{x}_{\pm}\|}}{2\pi\|\mathbf{x}-\mathbf{x}_{\pm}\|}$ coming from each of them. But regretfully Mr. Prankster is in charge of the lights and he tosses one or two coins to decide which source he will let shine or not. Thus the observer does not know in advance what Mr. Prankster will do but he has to predict what will be observed on the basis of any observation that he may gather.

Assume first that the decision will be made on the basis of a coin toss. In order to model this set up he chooses $\Omega = \{H, T\}$ and then the signal at any arbitrary (but fixed $\mathbf{x}$, to avoid dealing with function valued random variables) all he knows is that the signal has to be modeled by a random variable $U = u_+(\mathbf{x})I_H + u_-(\mathbf{x})I_T$, and if does not have any other information, then the best prediction he can make is that on the average he will observe $E[U] = p_H u_+(\mathbf{x}) + p_T u_-(\mathbf{x})$.
This is a good point to insist on the difference between a prediction like $E[U]$ and an observed value: When one observes a random variable, U in this case, each time one of its possible values is seen, even if the initial set up is the same. Since these differ from observation to observation, one performs a statistical analysis on the data, and an expected (value to be seen upon average) is provided. But assume that he is given information consisting of the specification of which bulb is on. This is modeled by a σ-algebra $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \Omega\}$, i.e., the full information about the setup. He knows that his best predictor must be computed as $E[U \mid \mathcal{F}] = U$, and when he is told that the result of the toss is an H, he then predicts that he will observe $u_+(\mathbf{x})$.
Assume now that the observer knows that Mr. Prankster is using two coins. Then to model the outcome of a single toss he considers $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$, and all the information about the experiment is contained in the σ-algebra $\mathcal{P}(\Omega)$. Assume that he knows the probabilities $\{p_1, p_2, p_3, p_4\}$ as listed above. The signal to be observed is modeled by the random variable (random field)

$$U = (u_+(\mathbf{x}) + u_-(\mathbf{x}))I_{\{(H,H)\}} + u_+(\mathbf{x}))I_{\{(H,T)\}} + u_-(\mathbf{x}))I_{\{(T,H)\}}.$$

Note again that the vales of U, that is of the observed field, depend on the result of the throw of the two coins. Now in the absence of any information, the best prediction that the observer can make is given by $E[U] = p_1(u_+(\mathbf{x}) + u_-(\mathbf{x})) +$

$p_2u_+(\mathbf{x}) + p_3u_-(\mathbf{x}) = (p_1+p_2)u_+(\mathbf{x}) + (p_1+p_3)u_-(\mathbf{x})$. Clearly p_1+p_2 and p_1+p_3 are, respectively the probabilities that the sources at $\mathbf{x}_\pm$ are on. These were denoted by p_H and p_T a few lines above.

Assume now that all the observer is able to find out is that either the two sources may be in the same state. That is, assume that both of the sources are on or off, or that either of them may be on and the other off. The given information corresponds to the partition of $\{\{(H,H),(T,T)\},\{(H,T),(T,H)\}\}$ of Ω. Denote by $\mathcal{G}$ the σ-algebra that it generates. The best predictor given this information is the random variable $E[U\,|\,\mathcal{G}]$ taking values

$$E[U\,|\,\{(H,H),(T,T)\}] = \frac{p_1(u_+(\mathbf{x})+u_-(\mathbf{x}))}{p_1+p_4} = \frac{p_1}{p_1+p_4}(u_+(\mathbf{x})+u_-(\mathbf{x}))$$

$$E[U\,|\,\{(H,T),(T,H)\}] = \frac{p_2u_+(\mathbf{x})+p_3u_-(\mathbf{x})}{p_2+p_3} = \frac{p_2}{p_2+p_3}u_+(\mathbf{x}) + \frac{p_3}{p_2+p_3}u_-(\mathbf{x}).$$

respectively whenever the event $\{(H,H),(T,T)\}$ or $\{(H,T),(T,H)\}$ occurs. The probabilities of occurrence of each of the events are, respectively p_1+p_4 and p_2+p_3. To make the role of the prior information more apparent, assume that the observer knows that one of the events in the following partition of Ω occurs: $\{\{(H,H),(T,T)\},\{(H,T)\},\{(T,H)\}\}$. The difference with the previous case is that when the observer finds out that one of the sources is uncovered, he knows that the other is covered. If we denote by $\mathcal{G}_1$ the σ algebra generated by that partition, then the best predictor of the random field U is the random variable (random field)

$$E[U\,|\,\mathcal{G}_1] = u_+(\mathbf{x})I_{\{(H,T)\}} + u_-(\mathbf{x})I_{\{(T,H)\}} + \frac{p_1}{p_1+p_4}(u_+(\mathbf{x})+u_-(\mathbf{x}))I_{\{(H,H),(T,T)\}}.$$

Observe for example, that if the toss of the coins were (T,T), then we would see no signal, but if it were (H,T), we would see $u_+(\mathbf{x})$. And equally important, that the three possible results would occur with probabilities p_2, p_3 and p_1+p_4. The difference between this and the first case treated, is that now both sources may be on or off, whereas in the first case one was on while the other was off.

3 Is the bird in the cage or not?

Let us now examine the classical variant of the Schroedinger cat paradox. Our setup consists of a light bulb with an exponential lifetime, and a bird in a cage. The setup is such that if the light bulb burns out, the latch on the bird's cage is released and the bird flies away. The issue is whether, upon observation, will we find the bird in the cage or not? To describe the result of one measurement of a lifetime, the following sample space is adequate: $\Omega = [0,\infty)$, and all questions we can ask about the measurement are represented by $\mathcal{F} = \mathcal{B}([0,\infty))$. As

probability measure we choose $P(dt) = \frac{1}{\tau}e^{-t/\tau}$. The random variable of interest is the lifetime of the bulb, described by $T : \Omega \to [0,\infty)$ such that $T(u) = u$, that is we shall find convenient to think of the points $u \in \Omega$ as "life-histories" of the light bulb. The distribution of this variable is $P(\{T > t\}) = e^{-t/\tau}$, which is to be read as: the fraction of life-histories longer that t is $e^{-t/\tau}$.
Let the position of the latch be denoted by $\top$ if it is open and by $\perp$ if it is closed. The random state of the latch at time t is described by a random variable $X_t : \Omega \to \{\top, \perp\}$ given by

$$X_t(u) = \top \text{ whenever } \mathrm{T(u) \leq t;\ \ or\ \ X_t(u)} = \perp \text{ whenever } \mathrm{T(u) > t.}$$

What happens here? Well, the observer decides when he is going to look at the bird, that is the observer chooses t, whereas nature chooses the life history u of the bulb. If $u = T(u) \leq t$ then the bulb is still on at the time of observation, whereas if $u = T(u) > t$, the bulb has burnt out and the bird has flown away. Is there anything paradoxical here? It does not seem so.

If the caretaker of the bird is to receive a payment d if nothing has happened, and he has to pay a penalty δ if the bird escapes, what is our best prediction of his fortune at time t? Clearly if no information is available, his fortune at time t is $d(1 - e^{-t/\tau}) - \delta e^{-t/\tau}$. We leave it up to the reader to see why, and direct her/him to Peres' [P] for a discussion from the point of view of a physicist.

4 Split coins and recoiling particles

In this section we shall consider non quantum variants of the EPR paradox. Again the aim is to emphasize the role of prior knowledge. Suppose you have a coin-like object with two faces which can be either R(ed) or B(lue), but these properties for all you know, have been assigned at random. The assignment is by means of a random variable taking values in the set $\{(R,R),(R,B),(B,B)\}$. When performing an experiment consisting of one measurement, it suffices to take $\Omega = \{\{R,R\},\{R,B\},\{B,B\}\}$. Clearly we think of the color assignment as a multi set: sets in which an element may appear repeated a number of times. This description takes care of the intuitive symmetry of a coin respect to the labeling of the faces. An observation of a coloring of a coin is actually a pair of observations, which we list sequentially. Formally speaking $X : \Omega \to \{R,B\} \times \{R,B\}$ and an equivalence relation has been defined on the range, which identifies (R,B) and (B,R). The values of X are obviously $X(\{B,B\}) = (X_1(\{B,B\}), X_2(\{B,B\})) = (B,B)$, etc. This may seem like overdoing it, but is really necessary if we want our model to reflect the symmetry of the coin with respect to "flips" (We may also think of colorings as mappings $\{1,2\} \to \{B,R\}$ which are invariant under the "flip" $F : \{1,2\} \to \{1,2\}$, $F(1) = 2$, and $F(2) = 1$).

If we assign probabilities to the elementary events by Ω by $P(\{B,B\}) = p_1$, $P(\{R,B\}) = p_2$ and $P(\{R,R\}) = p_3$, then clearly $P(\{X_1 = B\}) = p_1 + \frac{p_2}{2}$ and $P(\{X_1 = R\}) = p_3 + \frac{p_2}{2}$.

Suppose now that Mr. Prankster splits the coin in two (each half comprising a face), and a half given to some team mate. You have to guess the color of that face. You will receive either of the following pieces of information about the half in Mr. Prankster's hand: (i) the color of that half, (ii) the color of that half and the coloring of the coin, or (iii) only the coloring of the coin.

Case 1 Assume you are told that $X_2 = R$. In this case all you can predict is that either $X_1 = B$ or that $X_1 = R$ with the two probabilities given above.

Case 2 Assume that you are told that $X_2 = B$ and that the coloring is $\{B,R\}$. Clearly in this case you predict that $P(X_1 = R \,|\, X = \{B,R\},\ X_2 = B) = 1$. You do not have to invoke super luminal propagation of information or anything. Just apply basic probability (which in this case could not be more trivial).

Case 3 Here there are several possibilities depending on the explicit form of the information. For example, if told that both faces are equal but nothing else, you will compute basic conditional probabilities and conclude that the coloring is B with probability $\frac{p_1}{p_1+p_3}$ or R with the probability $\frac{p_3}{p_1+p_3}$. If you are told that the coloring is mixed, you will assert that one face is R and the other is B with probability 1.

A variation on this theme, corresponding to a simple version of the problem of the recoiling particles, is the following: Assume that the particles are marked (distinguishable) and an integer $1 \leq X_i \leq N, i = 1,2$ is assigned to each. Assume that the numbers are chosen at random but not necessarily independently of each other. For example, the numbers may measure "quanta" of energy transferred to each particle during the splitting process. Assume that you may observe X_1 and you want to predict X_2.
The underlying sample space for this situation is $\Omega = \{1,2,...,N\} \times \{1,2,...,N\}$, and the information about the measurement is described by a probability law $\mathcal{P}$ on Ω given by $P(X_1 = i, X_2 = j) = p_{ij}$, and each observable is modeled by a coordinate map, for example $X_1 : \Omega \to \{1,...,N\}$, such that $X_1(a,b) = a$.. When the X_i are independent $P(X_1 = i, X_2 = j) = P(X_1 = i)P(X_2 = j) = p_i p_j$, but this is not necessary for what comes below.

Any prediction about X_2 will depend about the information we are given about X_1 and about the pair. Let us consider three cases (i) we only know X_1, (ii) we are given X_1 and we now that $a \leq X_1 + X_2 \leq b$ is constant, but we do not know which constant, and (iii) is as the second case, but the value of the constant $X_1 + X_2 = k$ is specified.

Case 1 We know from Example 1 in the appendix that $E[X_2 = j \,|\, X_1] = \frac{p_{X_1 j}}{p_{X_1}}$. That is the best predictor of the probability of the event $\{X_2 = j\}$ depends on the observation of X_1. When we learn that $X_1 = i$, at that moment we know that we predict $X_2 = j$ with probability $\frac{p_{ij}}{p_i}$. But before the observation of X_1, the best predictor is a random variable.

Case 2 This time the generic information is the trace of $\sigma(X_1)$ on the event $\{a \leq X_1 + X_2 \leq b\}$, that is the σ-algebra generated by the collection $\{a - i \leq X_2 \leq b - i;\; 1 \leq i \leq N\}$. Let us denote this by the obvious $\sigma(X_1) \cap \{a \leq X_1 + X_2 \leq b\}$. In this case the best predictor of X_2 (not of any particular value of the variable) is

$$E[X_2 \,|\, \sigma(X_1) \cap \{a \leq X_1 + X_2 \leq b\}]$$

and we leave it up to the interested reader to compute the possible values of this random variable according to Example 1 in the appendix.

Case 3 Let us examine the most standard example corresponding to the classical version of the EPR paradox. Assume that it is known that $X_1 + X_2 = k$. If nothing is known about X_1, the best predictor of X_2 is

$$E[X_2 \,|\, X_1 + X_2 = k] = \sum_{i=1}^{N} \frac{(k-i)^+ p_i}{p_i} I_{\{X_1 = i\}}$$

where the explanation of the term $(k-i)^+$ is clear: the value $k-i$ is an allowed value for X_2 while it is bigger or equal to 0.
Notice that when we are given the event $\{X_1 = i\} \cup \{X_1 + X_2 = k\}$ (which is not empty, that is it describes a possible event only when $k \geq i$, then

$$E[X_2 \,|\{X_1 = i\} \cup \{X_1 + X_2 = k\}] = k - i!$$

This is clear because $\{X_1 = i\} \cup \{X_1 + X_2 = k\} = \{X_2 = k - i\}$, that is specifying the conserved quantity and one of the variables and the value of one of them, automatically specifies the other, regardless whether we measure it or not. In this case knowledge does not have to do with transmission of information, it has to do with logic.

Let us now consider these three cases under the assumption that X_1 and X_2 are continuous, say, real valued random variables, having a joint distribution function $\rho(x_1, x_2)$. Let us now re-examine some possibilities.

Case 4 Assume that all knowledge that may be available to us consists of the result of a measurement of X_1. What is our best predictor of X_2?. We already know that it is $E[X_2 \,|\, X_1]$, and in example 2 of the appendix, we show that this is the random variable $\int x_2 \frac{\rho(x_2,X_1)}{\rho(X)} dx_2$, whose value becomes known once X_1 is measured. When $X_1 = x_1$ is observed, our best predictor of X_2 becomes

$$E[X_2 \,|\, X_1 = x_1] = \int x_2 \frac{\rho(x_2, x_1)}{\rho(x_1)} dx_2.$$

Case 5 Let us consider now the case where only $Z \equiv X_1 + X_2$ is available for observation. What is our best predictor of X_2 in this case? This is similar to the case considered above, except that now we must first find the joint distribution of (Z, X_2). It is a simple computation to verify that it is $\hat{\rho}(z, x_2) = \rho(z - x_2, x_2)$ and that $\hat{\rho}(z) = \int \rho(z - x_2, x_2) dx_2$. Now the computation of $E[X_2 \,|\, Z]$ is a particular case of *Case 4*.

Case 6 Assume to finish that we may measure both Z and X_1. In this case note that for bounded continuous functions h, f, g, a simple change of variables should convince anyone that

$$\begin{aligned} E[h(X_2)f(X_1 + X_2)g(X_1)] &= E[h(Z - X_1)f(Z)g(X_1)] \\ &= E[E[h(X_2) \,|\, Z, X_1]f(Z)g(X_1)] \end{aligned}$$

from which it is clear (see definition (6.4) that $E[h(X_2) \,|\, Z, X_1] = h(Z - X_1)$. This is clearly what is expected, and the predictive formalism reflects it.

5 Bell inequalities

There are some inequalities, called Bell inequalities, satisfied by any set of three random variables taking values in $[-1, 1]$. And it said that their violation is a manifestation of the fact that the system under study is quantum and not classical, or that it "obeys a logic" different that the logic of classical probability. Actually, the following is an exercise of chapter 1 of [B]:

Let X, Y, Z, be three random variables taking values in $[-1, 1]$. Show that $1 - E[XY] \geq |E[XZ] - E[YZ]|$.

One possible proof of the inequality runs as follows: Consider

$$(1 + X)(1 - Y)(1 + Z) = 1 + X - Y - Z - XY + XZ - YZ$$

$$(1 - X)(1 + Y)(1 - Z) = 1 - X + Y + Z - XY + XZ - YZ$$

adding these two and noticing that the left hand side is always positive, rearranging and taking expected values, we obtain

$$E[YZ] - E[XZ] \leq 1 - E[XY].$$

To obtain the other half of the inequality, consider

$$(1-X)(1+Y)(Z-1) = -1 + X - Y + Z + XY - XZ + YZ$$
$$(1+X)(Y-1)(1+Z) = -1 - X + Y - Z + XY - XZ + YZ$$

adding and noticing that the left hand side is always negative, we obtain after rearranging and taking expected values, that

$$1 - E[XY] \leq E[XZ] - E[YZ]$$

which finishes the proof. And for fun consider the particular case: Let ξ, η, ζ be random variables taking (any) two values each, such that $P(\xi = a) = 1/2$, $P(\eta = b) = 1/2$ and $P(\zeta = c) = 1/2$, and consider that random variables $X = I\{\xi = a\}$; $Y = I_{\{\eta=b\}}$ and $Z = I_{\{\zeta=c\}}$. A direct application of the inequalities plus multiplication by 2, lead us to the inequality

$$2 - P(\xi = a \,|\, \eta = b) \geq |P(\eta = b \,|\, \zeta = c) - P(\xi = a \,|\, \zeta = c)|.$$

Comment 5.1 *Note that in the assumption about X, Y and Z, nothing is said about whether these variables represent measurements at near or far away locations, nor about the times the measurements are to be taken, or about possible physical or other kind of interpretation of the random variables. The only assumption made is about their possible values, nothing is even assumed about their joint distribution. If the inequalities are violated, it must be that the underlying mathematical structure of the probabilistic model under examination does not coincide with the classical probabilistic model, not due to some non-explicit assumption about causality or whatever.*

Here is an example of a quantum system that does seem to satisfy the Bell inequalities: Consider three identical, distinguishable and spinless particles moving on the real line, under the action of a three particle force given by the potential $V(x_1, x_2, x_3) = 0$ or $= +\infty$ depending on whether $(x_1^2 + x_2^2 + x_3^2$ is respectively ≤ 1 or ≥ 1. Mathematically speaking, this system is equivalent to a particle confined to move freely inside a sphere of radius 1. Can there exist a state $\Psi(x_1, x_2, x_3)$ such that the probability density $|\Psi(x_1, x_2, x_3)|^2$ violates Bell inequalities? Offhand, it seems that the answer is: “No, there is not.” But life is full of surprises.

To finish, we direct the reader to the exposé [B] for a nice presentation of the “geometric” ideas behind the inequalities as well as further references.

6 Appendix: The very basics of probabilistic modeling

6.1 Ensembles a.k.a Sample Spaces

The basic construct in probability theory is that of a sample space. This corresponds to what is called (but never explicitly defined) an ensemble in most

books in statistical thermodynamics. The construct is specified by a triplet $(\Omega, \mathcal{F}, \mathbb{P})$ where the set Ω should be thought of as the "experiments" (i.e., sequences of measurements), or sometimes it may be convenient to *think* of it as "the set of sates of nature". The set $\mathcal{F}$ is the mathematical construct describing the information about our system (i.e., it contains the objects that describe all questions **we assume valid** to ask about our system). Its elements are called events and formally it is a σ-algebra of subsets of Ω, that is a collection of sets satisfying:
i) $\emptyset \in \Omega$
ii) $A \in \mathcal{F} \Rightarrow A^c \equiv \Omega - A \in \mathcal{F}$
iii) If A_k for $k = 1, 2, ...$ is a countable collection of elements of Ω, then $\bigcup_k A_k \subset \Omega$.

Observe that the arithmetical structure of the set operations parallel that of the logical connectives in ordinary speech. That is why $\mathcal{F}$ is to be thought of as the class of allowed questions about the collection of experiments described by Ω. And to finish with the list,we have

Definition 6.1 *We say that $\mathbb{P}$ is a probability measure or law on the sample space $(\Omega, \mathcal{F})$, that is a function $\mathbb{P} : \Omega \to [0, 1]$ satisfying the conditions*
1)$\mathbb{P}(\Omega) = 1$
2)If A_k for $k = 1, 2, ...$ is a countable collection of ***disjoint*** *elements of Ω, then $\mathbb{P}(\bigcup_k A_k) \equiv \sum_k \mathbb{P}(A_k)$.*

Comment 6.1 *A pair $(S, \mathcal{S})$ where S is any set and $\mathcal{S}$ is a σ-algebra of subsets of S is called a* ***measurable space****, because an additive function (called measure or volume) can be consistently defined on it. The difference between probability and arbitrary measures is that the later do not have to be normalized to* 1.

6.2 Observables a.k.a Random Variables

And important notion in both classical and quantum mechanics is the concept of random variable, it is central to mathematical probability theory, just because it is the mathematical object embodying the notion of measurement, **not the actual process or act of measurement**, but the object describing the result of the measurement. The technical definition is the following:

Definition 6.2 *We assume we have already chosen a sample space $(\Omega, \mathcal{F})$ and we also have a measurable space $(S, \mathcal{S})$. An S-valued random variable is a function $X : \Omega \to S$ such that for any $A \in \mathcal{S}$ we have $\{X \in A\} \equiv X^{-1}(A) \in \mathcal{F}$.*

Comment 6.2 *We can interpret $X(\omega)$ as the value of the observable X in the experiment ω. It is here where the mathematical modeling of the randomness is* *located. Notice that different values of the observable are achieved in different experiments. Thus, the interpretation of the ω and of and of $X(\omega)$ are tailored to complement each other. Also, if the sets in $\mathcal{S}$ describe questions about the results of measurements, then the measurability condition just transfers questions about measurements onto questions about experiments taking given values. When $S = \mathbb{R}$ and $\mathcal{S} = \mathcal{B}(\mathbb{R})$, we just say that X is a random variable. For any topological space S, the class $\mathcal{B}(S)$ denotes the smallest σ-algebra containing the open sets of S and is called the Borel σ-algebra of S.*

Also, now and then we shall make use if the notation I_A to denote the indicator function of the set A, defined by $I_A(x) = 1$ or 0 depending on whether $x \in A$ or not.

The important example for us corresponds to the case in which $\mathcal{F}$ is generated by a partition $\pi = \{\Lambda_1, ..., \Lambda_K\}$. In this case any random variable is described by

$$X = \sum_{j=1}^{K} x_j I_{\Lambda_j}$$

that is, X takes value x_j whenever the event Λ_j takes place.

Three simple but important **examples** of σ-algebras are worth recording.
(i)First, consider a partition $\Pi = \{A_i, ..., A_n, , ...\}$, that is, a finite or infinite disjoint collection of subsets of Ω whose union is Ω. The σ-algebra $\mathcal{F} = \sigma(\Pi)$ that the partition generates is the collection of all possible union of sets of Π. Any $\mathcal{F}$-measurable function (random variable) is of the form $X = \sum_n x_n I_{A_n}$.
(ii) Conversely, any random variable X taking only a countable collection of values $\{x_1, x_2, ...\}$ generates a σ-algebra, obviously denoted by $\sigma(X)$ generated by the partition $\{\{X = x_n\} \,|\, n = 1, 2, ...\}$.
(iii)To finish the list, there is the trivial σ-algebra $\mathcal{F}_0 \equiv \{\emptyset, \Omega\}$. The reader should verify that all $\mathcal{F}_0$-measurable random variables are constant.

Actually, usually the construction of sample spaces and of the basic random variables is carried out simultaneously. A typical example of sample space is the following: Consider a finite set $R = \{r_1, ..., r_k\}$ and form $\Omega \equiv R^N$ where N is either finite or ∞. In any case think as follows $\Omega = \{\omega : \{1, 2, ..., N\} \to R\}$ is the collection of sequences of length N (finite or infinite). Each sequence describes an experiment and to describe the measurements it is convenient to introduce the following R-valued random variables: $X_i : \Omega \to R$ defined by $X_i(\omega) \equiv \omega(i)$. Mathematically speaking this is just a coordinate map, but it is the mathematical object that models the result of the $i - th$ measurement in the experiment Ω. The information about this setup is provided but the results of all possible experiments. This corresponds to the σ-algebra generated by the collection (of "cylinders") $\{X_{i_1} \in B_1, ..., X_{i_m} \in B_m : 1 \leq i_1 \leq i_2 \leq ... \leq i_m; B_1, ..., B_m \subseteq R\}$

To construct a probability on $(\Omega, \mathcal{F})$ there are many ways to proceed. If we want the X_i to be **statistically independent** we start from any assignment $\mathbb{Q}(B) = \sum_{q_i \in B} q_i$ where $\sum_{q_i \in R} q_i = 1$ and define $\mathbb{P}(\{X_{i_1} \in B_1, ..., X_{i_m} \in B_m\}) = \prod \mathbb{Q}(B_i)$. That is we start from a probability $\mathbb{Q}$ on R which we may have to find by doing statistical analysis, and we end up with a probability law on the class of all possible questions about the experiments.
But more interesting probability laws are conceivable. For example, when $N < \infty$, we may define a measure on Ω by putting

$$\mathbb{P}(\omega) \equiv \frac{e^{\beta \sum E(i,j) X_i(\omega), X_j(\omega)}}{Z(\beta)}$$

where $Z(\beta)$ is a normalization factor, which should remind us of the partition function. Anyway, this is as simple as you can get and say something interesting and familiar. But the list examples in the applied literature, ranging from signal processing to mathematical finance is enormous.

6.3 A simple ensemble

A standard construction in probability theory, which can obviously identified with the physicist's notion of ensemble (which by the way is never made explicit in any book on statistical physics), is useful for describing sequences of independent measurements. Denote by S the set in which the measurements take value. Let us assume that it is either a discrete set or a subset of some R^d. In the first case we consider $\mathcal{S} = \mathcal{P}(S)$ to be the σ-algebra on S. In the second case $\mathcal{S} = \mathcal{B}(S)$. We set $\Omega \equiv \prod_{j=1}^{N} S$, and we describe all the questions about the experiments by $\mathcal{F} \otimes_{j=1}^{N} \mathcal{S}$. We understand that N is either finite or $N = \infty$ depending on whether we need to describe finite or infinite sequences of measurements. The results of the measurements are described by the coordinate maps $X_n : \Omega \to S$ defined by $X_n(\omega) = s_n$ if $\omega = \{s_1, s_2, ...\}$

Thus, note that in this set up, $\omega = \{s_1, s_2, ...\}$ clearly denotes an experiment, that is, a sequence of measurements yielding values s_1, s_2, ...

A probability assignment which makes the X_n statistically independent is the following: Let $\mu : \mathcal{S} \to [0,1]$ be a given probability on $(S, \mathcal{S})$, which we want to describe the distributions of particular measurements, and we put $P(\{X_{n_1} \in A_1, ..., X_{n_k} \in A_k\}) \equiv \mu(A_1)...\mu(A_k)$, for any $n_1 < ... < nk$ and any $A_1, ...A_k$ in $\mathcal{S}$. In particular $P(\{X_n \in A\}) = \mu(A)$, and thus the X_n are **by construction** independent and identically distributed.

For example in order to describe an experiment consisting on the measurement of the life-time of a light bulb, or the life-time of a decaying particle, or the penetration of a particle in an absorbing medium, we take $N = 1$, $S = [0, \infty)$, $\mathcal{S} = \mathcal{B}(S)$ and $\mu(A) = \frac{1}{\tau} \int_A e^{-t/\tau} dt$.

6.4 Basic predictions

In order to do predictions we need to introduce two notions, that of mathematical expectation and that of conditional expectation. Logically speaking, the former is a particular case of the later, but it is easier to begin with the simpler notion. The definition or construction of the mathematical expectation is done stepwise. We shall define a **simple function** or **simple random variable** by $X \equiv \sum x_i I_{A_i}$ where $\{A_1, A_2, ..., A_m\}$ denotes a finite partition of Ω.

Now we **define** the expected value (or the mathematical expectation) of a simple random variable by

$$E[X] = \sum x_i \mathbb{P}(A_i)$$

This is the intuitive thing to do, and so it is to extend this to any variable: one can prove that any **positive** random variable X is an increasing limit of simple functions X_n, then we define $E[X] \equiv \lim_n E[X_n]$, and to finish one notices that any random variable can be written as $X = X^+ - X^-$, and when the expected of each of these is finite we say that X is integrable and write $E[X] = E[X^+] - E[X^-]$. Actually to express that summarily we write $E[|X|] < \infty$.

We are finally ready to introduce the most basic probabilistic notion: that of conditional expectation.

Definition 6.3 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability space, and $\mathcal{G}$ be s sub-σ-algebra of $\mathcal{F}$. Let Y be an integrable, $\mathcal{F}$-measurable, real valued random variable. The* ***conditional expectation of** Y* ***given** $\mathcal{G}$ is a $\mathcal{G}$-measurable random variable denoted by $E[Y \mid \mathcal{G}]$ satisfying the following condition*

$$E[YH] = E[E[Y \mid \mathcal{G}]H] \text{ for any } \mathcal{G} - \text{measurable, bounded H.} \tag{1}$$

Let us list some of its properties and then show how it is computed starting from (1).

Theorem 6.1 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G}$ be a subσ-algebra of $\mathcal{F}$. Then the following hold:*
(i) $E[1 \mid \mathcal{G}] = 1$.
(ii) If $X_1 \geq X_2$ are random variables, then $E[X_1 \mid \mathcal{G}] = E[X_2 \mid \mathcal{G}]$.
(iii) For $a, b \in \mathbb{R}$ $E[aX_1 + bX_2 \mid \mathcal{G}] = aE[X_1 \mid \mathcal{G}] + E[X_2 \mid \mathcal{G}]$.
(iv) (Filtering or tower property) If $\mathcal{H} \subset \mathcal{G}$ is another subσ-algebra, then for any integrable X; $E[E[X \mid \mathcal{G}] \mid \mathcal{H}] = E[X \mid \mathcal{H}]$.
(v) If Z is any bounded, $\mathcal{G}$-measurable random variable, then $E[ZX \mid \mathcal{G}] = ZE[X \mid \mathcal{G}]$.
(vi) $E[X \mid \mathcal{F}_0] = E[X]$

Even though we do not state it explicitly, the mapping $X \rightarrow E[X \mid \mathcal{G}]$ behaves in all respect as an expected value, except that it is a random variable. When

$\mathcal{G} = \sigma(Y)$, where Y is a given random variable, we shall write $E[X \,|\, Y]$ instead of $E[X \,|\, \sigma(Y)]$. The result that explains why conditional expectations are (best) predictors is contained in the following

Theorem 6.2 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let X be an integrable random variable and let Y be another random variable. Then the function $g(Y)$ that minimizes $E[(X - g(Y))^2]$ is $g(Y) = E[X \,|\, Y]$*

Comment 6.3 *Implicit in the statement of the theorem is an intuitive result that we need to complete the proof.*

Let us state a simple version of that result:

Theorem 6.3 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $Y_1, ..., Y_N$ be a collection of random variables. If Z is a random variable, measurable with respect to $\sigma(Y_1, ..., Y_N)$, then there exists a Borel measurable function $h : \mathbb{R}^N \to \mathbb{R}$ such that $Z = h(Y_1, ..., Y_N)$.*

Proof of theorem (6.2). It is basically a computation. Consider

$$E[(X - g(Y))^2] = E[((X - E[X \,|\, Y]) + (E[X \,|\, Y] - g(Y)))^2] =$$

$$E[(X - E[X \,|\, Y])^2] + E[(E[X \,|\, Y] - g(Y))^2] + 2E[(X - E[X \,|\, Y])(E[X \,|\, Y] - g(Y))]$$

We shall now verify that the last term vanishes. Therefore the assertion is clear, since we can choose $G(Y) = E[X \,|\, Y]$. Observe that from the definition it follows that by (1) (i.e., the defining property of the conditional expectation)

$$E[(X - E[X \,|\, Y])(E[X \,|\, Y] - g(Y))] = E[E[(X - E[X \,|\, Y]) \,|\, Y](E[X \,|\, Y] - g(Y))]$$

and now by properties (i) and (v) in theorem (6.1), with X there replaced by 1 and Z by $E[X \,|\, Y]$, it follows that the conditional expectation under the integral sign vanishes.

To illustrate how the definition is applied, let us now work out two standard examples:

Example 1 Consider the a σ-algebra $\sigma(\pi)$, generated by the partition $\pi = \{A_1, ..., A_n, ..\}$ of Ω. Assume that $\mathbb{P}(A_j) > 0 \ \forall j \geq 1$. Then since $E[X \,|\, \sigma(\pi)]$ is $\sigma(\pi)$-measurable, it can be written as $E[X \,|\, \sigma(\pi)] = \sum_j c_j I_{A_j}$. Clearly, to compute the c_j it suffices to apply the defining identity (1) to the binary random variables I_{A_k}. Thus

$$E[E[X \,|\, \sigma(\pi)] I_{A_k}] = E[X I_{A_k}] = E[(\sum_j c_j I_{A_j}) I_{A_k}] = c_k \mathbb{P}(A_k),$$

and therefore

$$E[X \mid \sigma(\pi)] = \sum_j \frac{E[XI_{A_j}]}{\mathbb{P}(A_j)} I_{A_j}. \tag{2}$$

Example 2 Consider now two random variables X and Y, about which you know the joint distribution $\rho(x,y)$, i.e., for any bounded function $F(X,Y)$, $E[F(X,Y)] = \int F(x,y)\rho(x,y)dxdy$. Assume that X is integrable, which presently means that $E[|X|] = \int |x|\rho(x,y)dxdy < \infty$. To compute $E[H(X) \mid Y]$ for any bounded $H(X)$, note that for any bounded $g(Y)$,

$$E[H(X)g(Y)] = \int H(x)g(y)\rho(x,y)dxdy = \int g(y)(\int h(x)\frac{\rho(x,y)}{\hat{\rho}(y)}dx)\hat{\rho}(y)dy$$

where we set $\hat{\rho}(y) = \int \rho(x,y)dx$. Also, temporarily put $E[H(X) \mid Y] = h(y)$, then

$$E[E[H(X) \mid Y]g(Y)] = E[h(Y)g(Y)] = \int h(y)g(y)\hat{\rho}(y)dy.$$

Since this identity is true for any $g(y)$, then it must be true that

$$h(Y) = E[H(X) \mid Y] = \int h(x)\frac{\rho(x,Y)}{\hat{\rho}(Y)}dx).$$

Comment 6.4 *what is important is that we are displaying the conditional expectation as a random variable. That is, the best predictor of $H(X)$ given that Y is measured is a random variable, whose value is known once Y is observed, and not before.*

6.5 The notion of independence

The proper definition of independence is at the level of σ-algebras. We have, within the usual model $(\Omega, \mathcal{F}, P)$

Definition 6.4 *(a) A collection $\{\mathcal{G}_i \mid i \in \mathcal{I}\}$ (where $\mathcal{I}$ is any set of indices) of sub-σ-algebras of $\mathcal{F}$, is said to be independent whenever for any finite subset $J \subset \mathcal{I}$, and foe any collection $\{A_i \mid i \in J\}$ we have*

$$P(\cap_{\{i \in J\}} A_i) = \prod_{\{i \in J\}} P(A_i).$$

(b) A family $\{X_i \mid i \in \mathcal{I}\}$ is independent whenever $\sigma(X_i)$ are independent.

Independence is a family property, and is a tricky property to verify. And it is sometimes confused with functional independence. Let us examine this source of confusion in a simple case. Consider $\Omega = \mathbb{R}^2$ and $\mathcal{F} = \mathcal{B}(\mathbb{R}^2)$. Consider the two random variables $X, Y : \Omega \to \mathcal{R}$ $X(x,y) = x$, $Y(x,y) = y$. These are functionally independent at least in two senses. First they are linearly independent as vectors in the class of functions defined over Ω, and second they are functionally independent, that is $Y \notin \sigma(X)$ (nor viceversa), that is, Y cannot be written as function of X. But they may not be independent as random variables, as the following two examples show:

Example 3 Consider the uniform distribution on the unit disk on Ω, that is $P(dx, dy) = \frac{1}{\pi}$ whenever $(x,y) \in \{(\xi,\eta) \,|\, (\xi)^2 + (\eta)^2 \leq 1\}$, and equal to zero otherwise. We leave it up to reader to find two sets A, B in the disk and verify that $P(\{X \in A, Y \in B\}) \neq P(\{X \in A\})P(\{Y \in B\})$.

Example 4 Within the same setup of the previous example, Assume that X and Y have a joint Gaussian distribution with correlation $\rho \neq 0$. They again, they are not independent, no matter how functionally independent they are.

Let us now examine another source of confusion. Consider the following model for the coin tossing experiment. $\Omega = \{0,1\}^{\mathbb{N}} = \{\omega : \mathbb{N} \to \{0,1\}\}$. Now experiments are specified by the "observation" of the random variables $X_n : \Omega \to \{0,1\}$ by $X_n(\omega) = \omega(n)$. Observe that specifying an experiment $\omega \in \Omega$ amounts to prescribing the result of the measurement of all X_n. The information available to the observer is modeled by the σ-algebra generated by the cylinder sets $\{\cap_{\{i \in J\}}\{X_i = \alpha_i\};\, |\, \alpha_i \in \{0,1\};\; J \text{ finite}\}$.One extends P from its values on cylinders, which to make life easy we take as $P(\{\cap_{\{i \in J\}}\{X_i = \alpha_i\}) = (\frac{1}{2})^{|J|}$, where we denote the cardinality if J by $|J|$. Let $\Theta : \Omega \to \Omega$ be the shift operator defined by $X_n \circ \Theta = X_{n+1}$, that is, the shift of a sequence by one unit to the left, and let Θ^n describe the $n-th$ iterate of Θ. It is rather easy to see that the mappings Θ^n are not independent, whereas the X_n are independent by construction. Consider for example the events $\Lambda_1 = \{X_2 = 0\}$ and $\Lambda_2 = \{X_1 = 1, X_2 = 1\}$. Consider for example: $\{\Theta \in \Lambda_1\} \cap \{\Theta^2 \in \Lambda_2\} = \emptyset$, but $P(\{\Theta \in \Lambda_1\}) = 1/2$ and $P(\{\Theta^2 \in \Lambda_2\}) = 1/4$.

Notice that the flow Θ^n is deterministic, that is, given the initial point ω of the orbit, the values $\Theta^n(\omega)$ are completely determined. But to give the initial point an infinite sequence of independent random variables has to be observed.

References

[A] Aczel, A. *"Entanglement"* Plume-Penguin, New York, 2003.

[B] Borkar, V.A. *"Probability Theory"* Springer-Verlag, Berlin, 1995.

[C] Cromer, A. *"Uncommon Sense"* Oxford Univ. Press, Oxford, 1993.

[D] Deutsch, D. *"The fabric of reality"* Penguin Books, New York, 1997.

[d'E] d'Espagnat, B. *"Reality and the Physicist"* Cambridge Univ. Press, Cambridge, 1990.

[F] Fine, A. *"The Shaky Game: Einstein Realism and the Quantum theory"* Chicago Univ. Press, Chicago, 1990.

[K] Kavafos, M. *"Bell's Theorem, Quantum Theory and Conceptions of the Universe"* Kluwer Acad. Pubs., Dordrecht, 1989 .

[KN] Kavafos, M. and Nadeau, R. *"The Conscious Universe"* Springer-Verlag, Berlin, 1990.

[L] Lindley, D. *"Where did the weirdness go?"* Basic Books, New York, 1996.

[M] Malament, D.B. *"Notes on the geometric interpretation of Bell type inequalities"* Downloaded from the authors web page.

[Ma] Mayants, L. *"The Enigma of Probability in Physics"* Kluwer Acad. Pubs., Dordrecht, 1984.

[Me] Mermin, D. *"Bringing home the atomic world: quantum mysteries for anyone"* Am. J.Phys. **49** (1981), pp. 940-943.

[P] Peres, A. *"The classic paradoxes of quantum theory"* The Foundations of Physics, **14** (1984),pp. 1131-1145.

[S] Silverman, M.P. *"More than a mystery"* Springer-Verlag, New York, 1995.

[TM] Tarozzi, G. and v.d Merwe, A. *"The Nature of Quantum Paradoxes'*, Kluwer Acad. Press, Dordrecht, 1988.

[WZ] Wheeler, J. and Zurek, H. *"Quantum Theory and Measurement"* Springer-Verlag, Berlin, 1990.

Henryk Gzyl
Depto. de Estadística; Univ. Carlos III de Madrid.
España
hgzyl@est-econ.uc3m.es; hgzyl@reacciun.ve