APPPHYS225 - Thursday 25 September 2008

Classical probability review, part 2

1. Joint systems
2. Conditioning, Bayes' Rule
3. Paradoxes
4. Conditional expectation
5. State discrimination and measures of distinguishability

Joint systems

Suppose now we have two six-sided dice. We can at first consider them to be independent systems living in separate sample spaces:

$$\Omega^A = \{\omega_1^A, \omega_2^A, \omega_3^A, \omega_4^A, \omega_5^A, \omega_6^A\}, \quad \Omega^B = \{\omega_1^B, \omega_2^B, \omega_3^B, \omega_4^B, \omega_5^B, \omega_6^B\}.$$

Clearly we can define random variables on each space, such as

$$X^A(\omega_i^A) = i, \quad X^B(\omega_j^B) = j.$$

Note that at this level of description, ω_1^B is not in the domain of $X^A(\cdot)$ and therefore $X^A(\omega_i^B)$ is undefined. Likewise, we have two probability distribution functions $m^A(\cdot)$ and $m^B(\cdot)$, which we might as well take to be uniform.

We can clearly construct a joint sample space by taking Cartesian products:

$$\Omega^{AB} = \{\omega_1^A \times \omega_1^B, \omega_1^A \times \omega_2^B, \omega_1^A \times \omega_3^B, \omega_1^A \times \omega_4^B, \omega_1^A \times \omega_5^B, \omega_1^A \times \omega_6^B, \omega_2^A \times \omega_1^B, \ldots, \omega_6^A \times \omega_6^B\}.$$

Now Ω^{AB} has 36 elements, corresponding to all possible outcomes of the rolling of a pair of six-sided dice. What about the random variables and probability distribution functions? Consider the following definition:

$$R^{AB}(\cdot) = R^A(\cdot) \times R^B(\cdot),$$
$$R^{AB}(\omega_i^A \times \omega_j^B) = R^A(\omega_i^A) R^B(\omega_j^B),$$

where the final expression indicates simple scalar multiplication of the numerical values of $R^A(\omega_i^A)$ and $R^B(\omega_j^B)$. Making use of the identity functions

$$1^A(\cdot) = \chi_{\Omega^A}(\cdot), \quad 1^B(\cdot) = \chi_{\Omega^B}(\cdot),$$

we can thus define *ampliations* of the random variables we initially define on the factor spaces Ω^A and Ω^B to the joint space Ω^{AB}. For example,

$$X^A(\cdot) \mapsto \Upsilon^B[X^A](\cdot) \equiv X^A(\cdot) \times 1^B(\cdot), \quad \Upsilon^B[X^A](\omega_i^A \times \omega_j^B) = X^A(\omega_i^A) = i,$$
$$X^B(\cdot) \mapsto \Upsilon^A[X^B](\cdot) \equiv 1^A(\cdot) \times X^B(\cdot), \quad \Upsilon^A[X^B](\omega_i^A \times \omega_j^B) = X^B(\omega_j^B) = j.$$

Often we will simply write

$$X^A(\omega_i^A \times \omega_j^B) = i, \quad X^B(\omega_i^A \times \omega_j^B) = j,$$

with all the ampliation stuff implied. Note that we can now also consider things like

$$X^{A\times B}(\cdot) \equiv X^A(\cdot) \times X^B(\cdot), \quad X^{A\times B}(\omega_i^A \times \omega_j^B) = ij,$$
$$X^{A+B}(\cdot) \equiv X^A(\cdot) \times 1^B(\cdot) + 1^A(\cdot) \times X^B(\cdot), \quad X^{A+B}(\omega_i^A \times \omega_j^B) = i + j.$$

Normally in games we consider $X^{A+B}(\cdot)$ to correspond to the numerical value of the roll. Incidentally, note that $X^{A+B}(\cdot)$ is a random variable on Ω^{AB} that does not simply factor into the product of a random variable on Ω^A with another random variable on Ω^B.

Turning now to the probability distribution functions, we note that

$$m^{A\times B}(\cdot) \equiv m^A(\cdot) \times m^B(\cdot), \quad m^{A\times B}(\omega_i^A \times \omega_j^B) = m^A(\omega_i^A) m^B(\omega_j^B) = 1/36$$

provides the proper joint probability distribution function on Ω^{AB} (it is clearly normalized). While we are free to define

$$m^{A+B}(\cdot) \equiv m^A(\cdot) \times 1^B(\cdot) + 1^A(\cdot) \times m^B(\cdot) = 1/3,$$

this function on Ω^{AB} is not a valid probability measure. Note that the action of $m^{A\times B}(\cdot)$ on subsets of Ω^{AB} follows in an obvious way.

While the matrix representations of our new joint random variables are rather cumbersome to write down, we note that they have dimension 36 which is the product of the matrix representation dimensions on the factor spaces. In fact, using notation familiar from quantum mechanics we can write

$$(R^{AB}) = (R^A) \otimes (R^B),$$

where $\otimes$ denotes tensor (Kronecker) product. If you didn't see this in your prevous quantum class don't worry - we'll review this later. If you do know how to take tensor products of matrices, perhaps you could verify the above relation for $(X^{A\times B})$ and (X^{A+B}).

Finally we mention the issue of *marginalization*. Suppose we retain the joint sample space Ω^{AB} but I now tell you that the dice are weighted and that I am going to roll them in some sneaky way that could correlate their outcomes. I summarize the information numerically by giving you a new joint probability distribution function $n^{AB}(\cdot)$. If we forget about the B die, what is the marginal probability distribution function $n^A(\cdot)$ for the A die only? In the functional notation we can write

$$n^A(\omega_i^A) = \sum_{j=1}^{6} n^{AB}(\omega_i^A \times \omega_j^B).$$

In matrix notation we would like to have a procedure for going from $\mathrm{diag}(n^{AB}(\omega_1^A \times \omega_1^B), \ldots, n^{AB}(\omega_6^A \times \omega_6^B))$ to $\mathrm{diag}(n^A(\omega_1^A), \ldots, n^A(\omega_6^A))$ via linear algebra-type operations. Again, from previous quantum classes it may not surprise you to hear that this is a partial trace operation; this also will be reviewed a bit later on in the course.

Conditioning

Suppose I roll the dice without showing you the exact outcome, but I tell you that $X^{A+B} = 7$. There are obviously several joint configurations consistent with this but we can rule out others, such as $\omega_1^A \times \omega_1^B$; how should you update your original $m^{A\times B}(\cdot)$ to obtain a *conditional* probability distribution $m^{A\times B}(\cdot | X^{A+B} = 7)$?

Most of you will have seen *Bayes' Rule* on some previous occasion:

$$\Pr(E|F) = \frac{\Pr(F|E)\Pr(E)}{\Pr(F)},$$

which can be thought of as a summary of the equations

$$\begin{aligned}\Pr(E,F) &= \Pr(E|F)\Pr(F),\\ \Pr(F,E) &= \Pr(F|E)\Pr(E),\\ \Pr(E,F) &= \Pr(F,E).\end{aligned}$$

For the present discussion it will be most useful to use the slightly modified form

$$\Pr(E|F) = \frac{\Pr(E,F)}{\Pr(F)}.$$

Here $\Pr(E,F)$ is the joint probability of E and F, while $\Pr(E|F)$ is the conditional probability of E given F. The probabilitiles $\Pr(E)$ and $\Pr(F)$ are understood to be prior probabilities, that is, the probabilites we would have assigned to the events E and F before gaining any updated information. Note here the use of the term *events*, which should immediately alert you to how we are going to proceed. Inovking Bayes' Rule for our dice scenario, we define

$$\begin{aligned}E &= \{\omega_i^A \times \omega_j^B\},\\ F &= \{\omega \in \Omega^{AB} : X^{A+B}(\omega) = 7\},\end{aligned}$$

(F is a level set of $X^{A+B}(\cdot)$) and find

$$m^{A\times B}(\omega_i^A \times \omega_j^B | X^{A+B} = 7) = \Pr(E|F) = \frac{\Pr(E,F)}{\Pr(F)} = \frac{m^{A\times B}(E \cap F)}{m^{A\times B}(F)}.$$

Clearly the numerator vanishes for any configuration not in the level set, and is equal to $1/36$ for any configuration that is in the level set. The denominator is actually independent of $\omega_i^A \times \omega_j^B$, and in fact can be seen to be equal to the sum over all $\omega \in \Omega^{AB}$ of $m^{A\times B}(\omega \cap F)$. Hence it is simple a normalization factor for the conditional probability distribution. So in the end,

$$\begin{aligned}m^{A\times B}(\omega_i^A \times \omega_j^B | X^{A+B} = 7) &= \frac{1}{6}, \quad \omega_i^A \times \omega_j^B \in F,\\ &= 0, \quad \omega_i^A \times \omega_j^B \notin F,\\ F &= \{\omega_1^A \times \omega_6^B, \omega_2^A \times \omega_5^B, \omega_3^A \times \omega_4^B, \omega_4^A \times \omega_3^B, \omega_5^A \times \omega_2^B, \omega_6^A \times \omega_1^B\}.\end{aligned}$$

The basic structure of Bayes' Rule, which we have just highlighted, is that you condition your probability distribution function by 'eliminating' all configurations that are inconsistent with the information gained and then renormalizing whatever is left over.

Paradoxes

Grinstead & Snell, Example 4.6 (Monty Hall).

Conditional expectation

Suppose an experiment is performed (the die is rolled) and we do not know the precise outcome, but we learn the value of a certain random variable $F = f$. How does this knowledge change our predictions regarding the values of other random variables?

Before gaining knowledge of f we can use our probability distribution function $m(\cdot)$ to compute expectation values of arbitrary random variables:

$$\boxed{\langle E \rangle = \sum_{\omega_i \in \Omega} E(\omega_i) m(\omega_i).}$$

Following the discussion above, it should naturally follow that after learning f we adopt the conditional probability distribution $m(\cdot | F = f)$ and compute *conditional expectation values*:

$$\boxed{\langle E \rangle_{F=f} = \sum_{\omega_i \in \Omega} E(\omega_i) m(\omega_i | F = f) = \sum_{\omega_i \in \Omega} E(\omega_i) \frac{m(\omega_i \cap \Omega_f)}{m(\Omega_f)},}$$

where Ω_f here denotes the level set of F corresponding to value f.

At this point you already get the main idea, which is quite simple, but before moving on we would like to use some of the tools we have developed illustrate an elegant notion from modern probability theory: conditioning of one random variable by another is actually a form of orthogonal projection. For this purpose we actually want to think of "the conditional expectation of E given F" as a new random variable, denoted $\langle E \rangle_F(\cdot)$ and defined by

$$\boxed{\langle E \rangle_F(\omega_i) = \langle E \rangle_{F=F(\omega_i)}.}$$

In words, $\langle E \rangle_F(\omega_i)$ is the conditional expectation value that we would assign to $E(\cdot)$ if we were to learn that the value of $F(\cdot)$ was equal to $F(\omega_i)$. Looking at the above expression for $\langle E \rangle_{F=f}$ and utilizing the basis expansion

$$\boxed{F(\cdot) = \sum_j f_j \chi_{f_j}(\cdot),}$$

it follows from inspection that we can write

$$\boxed{\langle E \rangle_F(\cdot) = \sum_j \sum_{\omega_i \in \Omega} E(\omega_i) \frac{m(\omega_i) \chi_{f_j}(\omega_i)}{\langle \chi_{f_j} \rangle} \chi_{f_j}(\cdot) = \sum_j \frac{\langle E \chi_{f_j} \rangle}{\langle \chi_{f_j} \rangle} \chi_{f_j}(\cdot).}$$

On the other hand, if we define an inner product and norm on random variables via

$$\boxed{(A, B) \equiv \langle AB \rangle, \quad |A|^2 = \langle A^2 \rangle,}$$

the orthogonal projection of $E(\cdot)$ onto the basis functions of $F(\cdot)$ should be given by

$$\boxed{E(\cdot) \mapsto \sum_j \left(E, \frac{\chi_{f_j}}{|\chi_{f_j}|} \right) \frac{\chi_{f_j}}{|\chi_{f_j}|},}$$

hence

$$\boxed{\langle E \rangle_F(\cdot) = \sum_i \left\langle E \frac{\chi_{f_i}}{\sqrt{\langle \chi_{f_i} \rangle}} \right\rangle \frac{\chi_{f_i}(\cdot)}{\sqrt{\langle \chi_{f_i} \rangle}} = \sum_i \frac{\langle E \chi_{f_i} \rangle}{\langle \chi_{f_i} \rangle} \chi_{f_i}(\cdot).}$$

We thus confirm that $\langle E \rangle_F(\cdot)$ corresponds to the orthogonal projection of $E(\cdot)$ onto the basis functions of $F(\cdot)$ (technically, onto the σ-algebra generated by the level sets of

$F(\cdot)$).

Having gone to the trouble of defining conditional expectation as a map on observables, we can easily derive the following fact:

$$\langle\langle E\rangle_F\rangle = \sum_j \sum_i \frac{\langle E\chi_{f_i}\rangle}{\langle\chi_{f_i}\rangle}\chi_{f_i}(\omega_j)m(\omega_j) = \sum_i \frac{\langle E\chi_{f_i}\rangle}{\langle\chi_{f_i}\rangle}\langle\chi_{f_i}\rangle = \sum_i \langle E\chi_{f_i}\rangle,$$

$$\sum_i \langle E\chi_{f_i}\rangle = \sum_i \sum_j E(\omega_j)\chi_{f_i}(\omega_j)m(\omega_j) = \sum_i \chi_{f_i}(\omega_j)\sum_j E(\omega_j)m(\omega_j) = \langle E\rangle.$$

Hence, the expectation of the conditional expectation of E with respect to any F is just the expectation of E. Similarly, the average of a conditional probability distribution over the possible values of the conditioning observable is equal to the original function (prove this to yourself as an exercise):

$$\sum_i m(\cdot|F = f_i)m(F = f_i) = m(\cdot).$$

With a bit of thought you can convince yourself that this can be interpreted to say that in classical probability, if we make a measurement but then ignore the result, there is no net effect on our probability model. In quantum probability this is not true because of the feature we generally call measurement backaction—this can be seen formally in the structure of the theory, and in any experimental scenario one can also find mechanistic explanations for it, with the essential ingredient in both cases being the existence of non-commuting observables.

Note that the expectation values that appear in these definitions are taken with respect to the original probability distribution function $m(\cdot)$. Thus, they are specified by the original *state* (before conditioning) on the algebra of random variables. Suppose for some reason we have access to the state—as it is formally defined—but not to the probability distribution function $m(\cdot)$ itself. As long we have a way of specifying how all the original expectation values for our algebra of random variables should be updated to conditional expectation values when we gain information about the value of a measured observable, we can update the state to a conditional state without ever making reference to $m(\cdot)$. In the classical probability setting this is pretty much a purely formal observation, but it will help us later in the term when we consider a sort of derivation of the quantum-mechanical projection postulate using classical conditional expectation. In quantum probability the density operator ρ is really a representation of state in the sense that it can be used to compute the expectation value of every observable in a non-commutative algebra, and we will be able to figure out how to update it after learning the result of a measurement without ever needing to make reference to a probability distribution function on the sample space, which is good since in the quantum case there is no sample space!

State discrimination and measures of distinguishability

Suppose I give you a coin and tell you that its weighting is such that it either follows the probability distribution function $m(\cdot)$ or $n(\cdot)$. We'll specify the sample space as

$$\Omega = \{\omega_1, \omega_2\}.$$

How well can you distinguish between $m(\cdot)$ versus $n(\cdot)$ if I demand that you make a guess based on the outcome of a single trial (coin flip)?

The first thing we should recognize is that this discrimination problem can be hard or easy, depending on the exact nature of $m(\cdot)$ and $n(\cdot)$. Consider the possibility that

$$m(\cdot)n(\cdot) = 0,$$

by which we mean that one of the following two cases holds:

$$\begin{array}{|l|}\hline m(\omega_1) = 1, \quad m(\omega_2) = 0, \quad n(\omega_1) = 0, \quad n(\omega_2) = 1, \\ \hline m(\omega_1) = 0, \quad m(\omega_2) = 1, \quad n(\omega_1) = 1, \quad n(\omega_2) = 0. \\ \hline \end{array}$$

If the two probability distributions are *orthogonal* in this sense, we can actually distinguish between them with a single trial. Note that it is possible for two distributions to be 'nearly' orthgonal, for example

$$m(\omega_1) = 0.99, \quad m(\omega_2) = 0.01, \quad n(\omega_1) = 0.01, \quad n(\omega_2) = 0.99,$$

in which case we could still make a distinction with high (but not absolute) confidence on the basis of a single trial. On the other extreme, a pair of distributions such as

$$m(\omega_1) = 0.501, \quad m(\omega_2) = 0.499, \quad n(\omega_1) = 0.499, \quad n(\omega_2) = 0.501,$$

will be relatively difficult to distinguish.

Up to now we have been talking about discriminating between probability distribution functions, but we could just as well talk about discriminating between states. For example, we could ask about a state ρ_m in which

$$\langle \chi_{\omega_1} \rangle = 1, \quad \langle \chi_{\omega_2} \rangle = 0,$$

versus a state ρ_n in which

$$\langle \chi_{\omega_1} \rangle = 0, \quad \langle \chi_{\omega_2} \rangle = 1.$$

In classical probability theory this is a subtle distinction, but remember than in quantum probability we will have states but not probability distribution functions.

By now it should seem natural that we would like to define some measures of distinguishability for probability distribution functions (or states). We'll start by discussing a measure called the *probability of error*. Coming back to the scenario in which I have given you a coin that is weighted such that it follows either the probability distribution function $m(\cdot)$ or $n(\cdot)$, let's say that we require you to make a guess on the basis of a single measurement of the random variable

$$X(\omega_1) = 1, \quad X(\omega_2) = 2, \quad X(\cdot) = \chi_{\omega_1}(\cdot) + 2\chi_{\omega_2}(\cdot).$$

We can use conditional probabilities to do this in an optimal way. The basic idea is to define an expanded probability model that includes not only the result of the coin flip but the probability distribution function from which it was drawn:

$$\Omega = \{\omega_m \times \omega_1, \omega_m \times \omega_2, \omega_n \times \omega_1, \omega_n \times \omega_2\}.$$

If I specify that the the choice of $m(\cdot)$ versus $n(\cdot)$ was originally made by flipping a fair coin, then the initial probability distribution function on Ω is

$$\begin{array}{|l|}\hline d(\omega_m \times \omega_1) = \frac{1}{2} m(\omega_1), \quad d(\omega_m \times \omega_2) = \frac{1}{2} m(\omega_2), \\ \hline d(\omega_n \times \omega_1) = \frac{1}{2} n(\omega_1), \quad d(\omega_n \times \omega_2) = \frac{1}{2} n(\omega_2). \\ \hline \end{array}$$

We should probably ampliate $X(\cdot)$ to the full Ω in order to keep things tidy,

$$\Upsilon[X](\omega_m \times \omega_1) = 1, \quad \Upsilon[X](\omega_m \times \omega_2) = 2, \quad \Upsilon[X](\omega_n \times \omega_1) = 1, \quad \Upsilon[X](\omega_n \times \omega_2) = 2,$$

and we can also define a random variable on Ω that corresponds to the identity of the coin probability distribution, which we are trying to discern:

$$Y(\omega_m \times \omega_1) = m, \quad Y(\omega_m \times \omega_2) = m, \quad Y(\omega_n \times \omega_1) = n, \quad Y(\omega_n \times \omega_2) = n.$$

Before learning the value of $\Upsilon[X](\cdot)$, our probability distribution on $Y(\cdot)$ is

$$\Pr(Y = m) = d(\{\omega_m \times \omega_1, \omega_m \times \omega_2\}) = \frac{1}{2}m(\omega_1) + \frac{1}{2}m(\omega_2) = \frac{1}{2},$$

$$\Pr(Y = n) = d(\{\omega_n \times \omega_1, \omega_n \times \omega_2\}) = \frac{1}{2}n(\omega_1) + \frac{1}{2}n(\omega_2) = \frac{1}{2}.$$

Suppose we now learn that $\Upsilon[X] = 1$. Then in terms of the conditional probability distribution (and being a bit intuitive with the notation),

$$d(\omega_m \times \omega_1 | \Upsilon[X] = 1) = \frac{d(\omega_m \times \omega_1 \cap \Upsilon[X] = 1)}{d(\Upsilon[X] = 1)} = \frac{d(\omega_m \times \omega_1)}{d(\omega_m \times \omega_1) + d(\omega_n \times \omega_1)} = \frac{m(\omega_1)}{m(\omega_1) + n(\omega_1)},$$

$$d(\omega_m \times \omega_2 | \Upsilon[X] = 1) = \frac{d(\omega_m \times \omega_2 \cap \Upsilon[X] = 1)}{d(\Upsilon[X] = 1)} = 0,$$

$$d(\omega_n \times \omega_1 | \Upsilon[X] = 1) = \frac{d(\omega_n \times \omega_1 \cap \Upsilon[X] = 1)}{d(\Upsilon[X] = 1)} = \frac{d(\omega_n \times \omega_1)}{m(\omega_1) + n(\omega_1)} = \frac{n(\omega_1)}{m(\omega_1) + n(\omega_1)},$$

$$d(\omega_n \times \omega_2 | \Upsilon[X] = 1) = \frac{d(\omega_n \times \omega_2 \cap \Upsilon[X] = 1)}{d(\Upsilon[X] = 1)} = 0.$$

We can easily see that the conditional probability distribution is normalized. Likewise,

$$d(\omega_m \times \omega_1 | \Upsilon[X] = 2) = \frac{d(\omega_m \times \omega_1 \cap \Upsilon[X] = 2)}{d(\Upsilon[X] = 2)} = 0,$$

$$d(\omega_m \times \omega_2 | \Upsilon[X] = 2) = \frac{d(\omega_m \times \omega_2 \cap \Upsilon[X] = 2)}{d(\Upsilon[X] = 2)} == \frac{m(\omega_2)}{m(\omega_2) + n(\omega_2)} = \frac{1 - m(\omega_1)}{m(\omega_2) + n(\omega_2)},$$

$$d(\omega_n \times \omega_1 | \Upsilon[X] = 2) = \frac{d(\omega_n \times \omega_1 \cap \Upsilon[X] = 2)}{d(\Upsilon[X] = 2)} = 0,$$

$$d(\omega_n \times \omega_2 | \Upsilon[X] = 2) = \frac{d(\omega_n \times \omega_2 \cap \Upsilon[X] = 2)}{d(\Upsilon[X] = 2)} == \frac{n(\omega_2)}{m(\omega_2) + n(\omega_2)} = \frac{1 - n(\omega_1)}{m(\omega_2) + n(\omega_2)}.$$

Our strategy for the guessing game will be to guess m for the identity of the coin-flip probability distribution function if $d(\omega_m \times \omega_1 | \Upsilon[X] = x) > d(\omega_n \times \omega_1 | \Upsilon[X] = x)$ or n if the inequality is reversed (if the conditional probabilities are equal, we can guess either one), where x is the value of $\Upsilon[X](\cdot)$ that we obtain from our single experimental trial. Let us assume that

$$d(\omega_m \times \omega_1 | \Upsilon[X] = 1) \geq d(\omega_n \times \omega_1 | \Upsilon[X] = 1),$$

(without loss of generality - we are just choosing a labeling scheme such that $m(\cdot)$ is the probability distribution function candidate with greater probability for ω_1), and note that this implies

$$d(\omega_n \times \omega_2 | \Upsilon[X] = 2) \geq d(\omega_m \times \omega_2 | \Upsilon[X] = 2),$$

since $m(\omega_1) \geq n(\omega_1)$ implies $1 - m(\omega_1) \leq 1 - n(\omega_1)$. Hence we can arrive at a fixed expression for the *average probability of error*,

$$\boxed{\mathrm{PE} = d(\omega_n \times \omega_1) + d(\omega_m \times \omega_2) = \frac{1}{2}(n(\omega_1) + m(\omega_2)).}$$

We check that for the orthogonal distributions $m(\omega_1) = 1,\ n(\omega_1) = 0,\ \mathrm{PE} \to 0,$ whereas for identical distributions $m(\bullet) = n(\bullet)$ we find $\mathrm{PE} \to 1/2.$ For the case

$$\boxed{m(\omega_1) = 0.99, \quad m(\omega_2) = 0.01, \quad n(\omega_1) = 0.01, \quad n(\omega_2) = 0.99,}$$

considered above, $\mathrm{PE} \to 0.01$, and for

$$\boxed{m(\omega_1) = 0.501, \quad m(\omega_2) = 0.499, \quad n(\omega_1) = 0.499, \quad n(\omega_2) = 0.501,}$$

we have $\mathrm{PE} \to 0.499$. In general it is possible to show that

$$\boxed{\mathrm{PE}[m(\bullet), n(\bullet)] = \frac{1}{2} \sum_{\omega_i \in \Omega} \min\{m(\omega_i), n(\omega_i)\}.}$$

For more on PE and the other measures discussed below, see for example C. A. Fuchs and J. van de Graaf, IEEE Transactions on Information Theory, Vol. 45, p. 1216 (1999).

In the above discussion we have assumed that the person trying to distinguish $m(\bullet)$ versus $n(\bullet)$ has access to the precise outcome of the coin flip, or equivalently to a one-to-one random variable. If we expand our perspective to think a bit about larger sample spaces such as that of a single role of a six-sided die, it becomes interesting to contemplate scenarios in which we have access only to 'partial information' observables that are not one-to-one. For example, suppose I roll a die that is weighted according to either $m(\bullet)$ or $n(\bullet)$ and you must guess which on the basis of a single roll, but the only information I will give you is the value of one observable that takes values in the set $\{0, 1\}$ (an indicator function/projector). You can pick whatever random variable you like from this class, and clearly you now have a compound optimization problem to solve: first you must pick an observable and then you must design a policy for deciding based on the measurement result. In some cases this is easy. For example if $m(\bullet)$ and $n(\bullet)$ are orthogonal, you can pick an observable that corresponds to an indicator function on the zeros of one of the candidate probability distribution functions. But the general case is less obvious, and you might worry that you would just need to compute the probability of error for every possible binary observable and then pick the best one. We'll revisit this problem in the quantum setting next week.

The probability of error is one fundamental measure of distingushability between probability distribution functions (or states), which is motivated by this measure-once-and-guess scenario, but there are others we could consider as well. There is for example the *Kolmogorov distance*,

$$\boxed{\mathrm{K}[m(\bullet), n(\bullet)] \equiv \frac{1}{2} \sum_{\omega_i \in \Omega} |m(\omega_i) - n(\omega_i)|,}$$

which has the advantage of satisfying a triangle inequality since in matrix notation

$$\boxed{\mathrm{K}[m(\bullet), n(\bullet)] \to \frac{1}{2} \mathrm{Tr}|\rho_m - \rho_n|,}$$

and the RHS is just the trace-norm distance on operators. The Kolmogorov distance is closely related to the probability of error, as

$$\boxed{\mathrm{PE}[m(\bullet), n(\bullet)] = \frac{1}{2} - \frac{1}{2}\mathrm{K}[m(\bullet), n(\bullet)].}$$

Another measure, whose quantum generalization is quite popular, is the Bhattacharya

coefficient

$$\mathrm{B}[m(\cdot), n(\cdot)] \equiv \sum_{\omega_i \in \Omega} \sqrt{m(\omega_i) n(\omega_i)}\,,$$

which is a sort of geometric overlap between probability distribution functions. The Bhattacharya coefficient is unfortunately neither a distance function nor related to any known statistical inference problem, but it does simply capture the basic idea of 'near-orthogonality' with which we started this discussion.

Further measures of distinguishability such as the Kullback-Leibler distance and the statistical distance are motivated by ideas from classical information theory, but these are beyond the scope of our short classical probability review.

Hopefully by now it is perfectly clear that in classical probability theory we have *configurations*, which by definition are perfectly distinguishable from one another, as well as *states*, which in general are not. The problem naturally arises of designing statistical analysis procedures to distinguish optimally between/among states given a finite number of samples, and when the configuration is not directly accessible we can ask about optimal measurements for the purpose of state discrimination. Next week we'll begin to examine quantum versions of these basic problems in measurement theory...