A Few Notes on Sets of Probability Distributions

Fabio G. Cozman
Escola Politécnica, Universidade de São Paulo
fgcozman@usp.br — http://www.poli.usp.br/p/fabio.cozman

June 28, 2008

A person looks at the sky, and tries to assess what she expects about the weather in the near future; another person weighs the costs of a financial application against its expected gain. Such expectations can sometimes be translated into a single number: for instance, the expected return on the financial application may be exactly 100 monetary units. More often, one has less determinate assessments about expectations, such as “I expect a return between 30% and 40% on this investiment.” Or perhaps, “I’d rather bet that it will rain tomorrow than that I get heads in the next toss of this coin.” In general, assessments specify a *set of probability distributions* that represent the uncertainty about a particular situation. This initial chapter introduces some concepts and tools that are useful in dealing with such sets.

1 Possibility space, states, events

The first element of this investigation is the *set of possible states of the world*. Each state is a complete description of all aspects of the world that are of interest; states are mutually exclusive. This set of states, denoted by Ω, is called the *possibility space*. (Often Ω is called the *sample space*, and states are called *samples*, *elements*, *outcomes*, *realizations*. Discussion on terminology can be found in Section 15.)

The following examples present possibility spaces that are binary, finite, countably and uncountably infinite.

> **Example 1.1.** We are interested in the weather forecast for next week. Among the many phenomena that interfere with such a forecast, we focus on just two possibilities: the weather may be *Sunny* or *Rainy*. The possibility space is $\Omega = \{\mathit{Sunny}, \mathit{Rainy}\}$.

Example 1.2. Two coins are tossed; each coin can be heads (H) or tail (T). The possibility space is $\Omega = \{HH, HT, TH, TT\}$.

Example 1.3. A coin is to be tossed n times, and we are interested in the number of heads that we will see in the sequence of tosses. If we take that the sequence may go on forever, the number of heads belongs to $\Omega = \{0, 1, 2, \dots\}$. That is, the possibility space is the set of non-negative integers.

Example 1.4. Suppose we consider infinite sequences of zeros and ones that are perhaps produced by tossing coins, and suppose we understand each sequence ω as the real number $0.\omega$ expressed in the binary numeral system. For instance, the sequence $1, 0, 0, \dots$ is $0.100\dots$, that is, the real number $1/2$. Now the possibility space of all such real numbers is the real interval $\Omega = [0, 1]$.

An *event* is a subset of Ω. The interpretation is that an event A obtains when a state ω in A obtains. There is no fuzziness in our states and events: an event either obtains or does not obtain. For example, the sentence "Mary is young" does not specify an event unless we have a rule that decides exactly when the predicade "is young" is true. Suppose "young" means "less than 18 years old." Then the sentences "Mary is young" and "Mary is less than 18 years old" are equivalent and both define an event.

Example 1.5. Two coins are tossed. Consider three events. Event $A = \{HH\}$ is the event that both tosses produce heads. Event $B = \{HH, TT\}$ is the event that both tosses produce identical outcomes. Event $C = \{HH, TH\}$ is the event that the second toss yields heads.

The superscript c denotes complement with respect to Ω; for example, A^c is the event containing all states not in A. The *intersection* of events A and B is denoted by $A \cap B$; that is, the event containing states both in A and in B. The *union* of events A and B is denoted by $A \cup B$; that is, the event containing states in A or B or both.

When we are dealing with an event A and this event obtains, we often say that we have a *success*; if instead A^c obtains, we say that we have a *failure*.

2 Variables and indicator functions

Once we have a possibility space, we can imagine a function from Ω to the real numbers. Any such function $X : \Omega \to \Re$ is called a *random variable*. The shorter term *variable* is often employed when referring to random variables. The set of all possible values of variable X is denoted by Ω_X.

A random variable is always denoted by capital letters X, Y, Z, etc. A particular value of a variable is not capitalized. For example, x denotes a value of X, and $\{X = x\}$ denotes the event $\{\omega \in \Omega : X(\omega) = x\}$.

If X is a variable, then any function $f : \Re \to \Re$ defines another random variable $f(X)$. Similarly, a function of several random variables is a random variable.

Example 2.1. The age in months of a person ω selected from a population Ω is a variable X. The same population can be used to define a different variable Y where $Y(\omega)$ is the weight (rounded to the next kilogram) of a person ω selected from Ω. Then $Z = X + Y$ defines yet another random variable.

Example 2.2. Suppose n coins are tossed. The possibility space contains 2^n possible sequences of heads and tails. Now define variable $X(\omega)$ as the number of coins that land heads in ω. The value of X ranges from 0 to n. This variable can be used to define interesting events, such as $\{\omega : X(\omega) \leq n/2\}$ and $\{\omega : 5 \leq X(\omega) \leq 10\}$ (if $n < 5$ then the second event is the empty set).

Every event can be associated with a random variable I_A, called its *indicator function,* such that $I_A(\omega) = 1$ if $\omega \in A$ and $I_A(\omega) = 0$ otherwise. Any random variable can be expressed as a linear, perhaps infinite, combination of indicator functions:

$$X = \sum_{\omega \in \Omega} X(\omega) I_\omega. \tag{1}$$

This expression holds because for each $\omega' \in \Omega$ such that $\omega \neq \omega'$, $I_\omega(\omega')$ is zero. So, for every $\omega' \in \Omega$, both sides of Expression (1) refer to $X(\omega')$.

Indicator functions do deserve some attention, as they let Boolean operations be easily expressed as pointwise algebraic operations. For example, the intersection of events A and B has indicator function

$$I_{A \cap B} = \min(I_A, I_B) = I_A I_B.$$

As another example, $I_{A^c} = 1 - I_A$. As a third example, consider unions of events $\{A_i\}_{i=1}^n$. We have $I_{\cup_{i=1}^n A_i} = \max_i I_{A_i}$, and if all events are disjoint, then

$$I_{\cup_{i=1}^n A_i} = \sum_{i=1}^n I_{A_i}.$$

Because $A \cup B = (A \cap B^c) \cup (A^c \cap B) \cup (A \cap B)$ and the latter three events are disjoint,

$$I_{A \cup B} = I_A(1 - I_B) + (1 - I_A) I_B + I_A I_B = I_A + I_B - I_A I_B.$$

Indicator functions are very useful devices, but sometimes they lead to heavy notation with too many subscripts. It is better to use *de Finetti's convention*: use the same symbol for an event and its indicator function. For example, instead of writing $A \cap B = I_A I_B$, we now write $A \cap B = AB$. Likewise,

$$A \cup B = A + B - AB,$$

and if the events $\{A_i\}_{i=1}^n$ are all disjoint, then

$$\cup_{i=1}^n A_i = \sum_{i=1}^n A_i.$$

As another example, consider an expression that mixes set and algebraic notation:

$$(A \cap B^c) \cup (A^c \cap B) = A(1-B) + (1-A)B = (A-B)^2.$$

While such expressions may seem perplexing at first, they do simplify matters. Consider the following important result on general unions, and note how much de Finetti's convention simplifies the notation.

Theorem 2.1. *Given events* $\{A_i\}_{i=1}^n$, *then*

$$\cup_{i=1}^n A_i = \sum_{j=1}^n (-1)^{j+1} A_{j,n}, \tag{2}$$

where $A_{j,n} = \sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq n} A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_j}$.

Proof. The case $n = 2$ has been proven (that is, $A \cup B = A + B - AB$). Now consider an induction argument: suppose the theorem holds for $n-1$ events A_1 to A_{n-1}. Note that $\cup_{i=1}^n A_i = A_n + \cup_{i=1}^{n-1} A_i - A_n \cap \left(\cup_{i=1}^{n-1} A_i\right)$, using the fact that $A \cup B = A + B - AB$. As $A_n \cap \left(\cup_{i=1}^{n-1} A_i\right) = \cup_{i=1}^{n-1} A_i \cap A_n$, we can use the induction hypothesis twice to obtain:

$$\begin{aligned} \cup_{i=1}^n A_i &= A_n + \\ & \left(\sum_{1 \leq i \leq n-1} A_i - \sum_{1 \leq i_1 < i_2 \leq n-1} A_{i_1} \cap A_{i_2} + \sum_{1 \leq i_1 < i_2 < i_3 \leq n-1} A_{i_1} \cap A_{i_2} \cap A_{i_3} - \dots \right) - \\ & \left(\sum_{1 \leq i \leq n-1} A_i \cap A_n - \sum_{1 \leq i_1 < i_2 \leq n-1} A_{i_1} \cap A_{i_2} \cap A_n + \dots \right). \end{aligned} \tag{3}$$

By grouping terms we obtain Expression (2). □

3 Expectations

Suppose one is willing to "buy" a random variable X for any amount α that is strictly smaller than some number $\underline{E}[X]$. The result of paying α and getting X is

$$X - \alpha.$$

Now suppose the person is willing to "sell" X for any amount β that is strictly larger than some number $\overline{E}[X]$. The result of selling X for β is

$$\beta - X.$$

We should expect that $\overline{E}[X] \geq \underline{E}[X]$, for otherwise we might put our subject in a truly embarrassing situation, as follows. Suppose

$$\overline{E}[X] - \underline{E}[X] = -\delta \qquad \text{for some } \delta > 0;$$

then by definition of $\underline{E}[X]$ and $\overline{E}[X]$, the person must be willing to buy X for $\underline{E}[X] - \delta/3$ and then to sell X for $\overline{E}[X] + \delta/3$. The net result of such transactions is

$$\begin{aligned} (X - (\underline{E}[X] - \delta/3)) + ((\overline{E}[X] + \delta/3) - X) &= \overline{E}[X] - \underline{E}[X] + 2\delta/3 \\ &= -\delta/3 < 0. \end{aligned}$$

So, if $\overline{E}[X] \leq \underline{E}[X]$, then the person stands to lose by engaging in two transactions that she, by definition, was willing to accept!

Suppose that for some random variable X we have $\overline{E}[X] = \underline{E}[X]$. We then denote by $E[X]$ the value $\overline{E}[X] = \underline{E}[X]$, and refer to it as the *expectation* of X. The expectation of a random variable X can be interpreted as a "fair price" for X: our subject is willing to buy X for less than $E[X]$, and is willing to sell X for more than $E[X]$.

If we find ourselves in the excepcional circumstance that an unique expectation $E[X]$ is given for every variable X, we can treat $E[\cdot]$ as a single entity, called an *expectation functional*. We assume that any expectation functional should satisfy:

Definition 3.1 (Axioms for expectation functionals)**.**

EU1 For constants α and β, if $\alpha \leq X \leq \beta$, then $\alpha \leq E[X] \leq \beta$.

EU2 $E[X + Y] = E[X] + E[Y]$.

The first axiom is quite reasonable: the fair price for X cannot be smaller than the smallest value of X, and likewise for the largest value of X. The second axiom takes the fair price of $X + Y$ as simply the fair price of X plus the fair price of Y.

Some consequences of the axioms are obvious: for instance, if α is a real number, $E[\alpha] = \alpha$ by EU1. And then, for any X,

$$E[X + (-X)] = E[X] + E[-X] = 0$$

and consequently:

$$E[X] = -E[-X]\,. \tag{4}$$

The axioms have a number of consequences that are worth stating explicitly (note that $X \geq Y$ means $X(\omega) \geq Y(\omega)$ for any $\omega \in \Omega$):

Theorem 3.1. *An expectation functional satisfies, for any real number α and variables X and Y:*

1. $X \geq Y \Rightarrow E[X] \geq E[Y]$.
2. $E[\alpha X] = \alpha X$.

Proof. If $X \geq Y$, then

$$X - Y \geq 0 \Rightarrow E[X - Y] \geq 0 \Rightarrow E[X] + E[-Y] \geq 0 \Rightarrow E[X] \geq -E[-Y] \Rightarrow E[X] \geq E[Y]\,.$$

For the second statement, start by noting that, for any integer $n > 0$, $E[nX] = E[\sum_{i=1}^{n} X] = \sum_{i=1}^{n} E[X]$ using finite induction on EU2. For two integers $n > 0$ and $m > 0$, $E[Y] = (n/n)E[Y] = (1/n)E[nY]$ and $E[Y] = (m/m)E[Y] = (1/m)E[mY]$; thus $(1/n)E[nY] = (1/m)E[mY]$. Take $X = Y/m$ to obtain $E[nX/m] = (n/m)E[X]$. So the second statement holds if α is a positive rational.

If α is not rational, recall that the rationals are dense in the reals (that is, there is always a rational between two distinct reals). Assume $X \geq 0$. If $E[X] = 0$, then $\alpha E[X] = E[\alpha X] = 0$ and the desired result obtains). Assume then $E[X] > 0$, and produce a contradiction by assuming $\alpha E[X] < E[\alpha X]$. Find a rational r such that $\alpha < r < E[\alpha X]/E[X]$. Then: (a) $\alpha < r \Rightarrow \alpha X < rX \Rightarrow E[\alpha X] < E[rX]$; (b) $r < E[\alpha X]/E[X] \Rightarrow rE[X] < E[\alpha X] \Rightarrow E[\alpha X] > E[rX]$. So we get a contradiction. A similar contradiction obtains if we assume that $\alpha E[X] > E[\alpha X]$. Thus $\alpha E[X] = E[\alpha X]$ if $X \geq 0$.

If X is not always nonnegative, then write X as $Y - Z$, where $Y = \max(X, 0)$ and $Z = \max(-X, 0)$, and use the fact that $E[-Z] = -E[Z]$ to produce the result as follows: $E[\alpha X] = E[\alpha Y - \alpha Z] = E[\alpha Y] - E[\alpha Z] = \alpha E[Y] - \alpha E[Z] = \alpha E[X]$. □

Thus, expectation functionals are *linear* because $E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y]$, and are *monotone* because $X \geq Y$ implies $E[X] \geq E[Y]$.

Example 3.1. Suppose $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and take the four variables in Table 1. Suppose we know that a person has bounds on her fair prices such that $E[X_i] \in [\mu_i, \nu_i]$, where μ_i and ν_i are also indicated in Table 1.

Take an additional variable X_5 such that $X_5(\omega_1) = 3$, $X_5(\omega_2) = -1$, $X_5(\omega_3) = 2$. Axiom EU1 requires that $E[X_5] \in [-1, 3]$. As $X_5 = X_1 + X_2 + 2X_3$, axioms EU1-EU2 imply

$$E[X_5] = E[X_1] + E[X_2] + 2E[X_3]\,.$$

There is no unique fair price for X_5 given the assessments in Table 1. All we can say is that fair prices for X_5 must belong to an interval $[\mu_5, \nu_5]$. Certainly

$$\mu_5 \geq \mu_1 + \mu_2 + 2\mu_3 = 1/3,$$

and

$$\nu_5 \leq \nu_1 + \nu_2 + 2\nu_3 = 10/3.$$

Actually, the best bounds on $E[X_5]$ are $[11/18, 11/6]$; an easy method to obtain such bounds is discussed in Example 6.2.

X_i	$X_i(\omega_1)$	$X_i(\omega_2)$	$X_i(\omega_3)$
X_1	1	0	0
X_2	2	-1	0
X_3	0	0	1
X_4	-1	0	1

μ_i	ν_i
0	$2/3$
0	2
$1/6$	$1/3$
-1	0

Table 1: Variables and bounds on their prices (Example 3.1).

It remains to examine the relationship between fair prices $E[X]$, the "maximum buying price" $\underline{E}[X]$, and the "minimum selling price" $\overline{E}[X]$. Consider the following rationale:

- If one holds a set of expectations for X, then this person must be willing to buy X for any amount smaller than $\inf E[X]$.
- Likewise, the person must be willing to sell X for any amount larger than $\sup E[X]$.

For instance, in Example 3.1 we should take each μ_i as the lower expectation $\underline{E}[X_i]$, and likewise ν_i should be the upper expectation $\overline{E}[X_i]$. We are thus led to the definition:

Definition 3.2. The *lower* expectation and the *upper* expectation of variable X are respectively

$$\underline{E}[X] = \inf E[X] \qquad \text{and} \qquad \overline{E}[X] = \sup E[X] \,.$$

For any variable, $\underline{E}[X] = -\overline{E}[-X]$ because

$$\underline{E}[X] = -\sup E[-X] = -\overline{E}[-X] \,. \tag{5}$$

If $\alpha \leq X \leq \beta$, then clearly

$$\alpha \leq \underline{E}[X] \leq \beta, \qquad \alpha \leq \overline{E}[X] \leq \beta.$$

Also, for $\alpha > 0$ (note: α is nonnegative!),

$$\underline{E}[\alpha X] = \alpha \underline{E}[X] \,, \qquad \overline{E}[\alpha X] = \alpha \overline{E}[X] \,.$$

Moreover, for any two variables X and Y,

$$\underline{E}[X+Y] = \inf E[X+Y] \geq \inf E[X] + \inf E[Y] = \underline{E}[X] + \underline{E}[Y]$$

and

$$\overline{E}[X+Y] = \sup E[X+Y] \leq \sup E[X] + \sup E[Y] = \underline{E}[X] + \underline{E}[Y]$$

The discussion so far has implicitly assumed that $\underline{E}[X]$ and $\overline{E}[X]$ are finite; otherwise, we may run into difficulties whenever $\underline{E}[X] = \overline{E}[X] = \infty$. In this chapter we adopt the following assumption, which is enough to avoid problems given axioms EU1-EU2 and Definition 3.2:

Assumption 3.1. *Every random variable is bounded.*

4 Some consequences of the axioms

The assessment of expectations for some variables constrains the values of other assessments through axioms EU1-EU2. Here we examine a few such constraints.

4.1 Simple variables and their expectations

A random variable is *simple* if it may assume only a finite number of different values. A simple variable X can always be written as follows:

$$X = \sum_{i=1}^{n} \alpha_i A_i,$$

for a set of numbers α_i and indicator functions A_i. Axioms EU1-EU2 imply:

$$E[X] = E\left[\sum_{i=1}^{n} \alpha_i A_i\right] = \sum_{i=1}^{n} E[\alpha_i A_i] = \sum_{i=1}^{n} \alpha_i E[A_i]\,.$$

Obviously, if the possibility space is finite then any variable is simple. In this case we can use Expression (1) and write:

$$\begin{aligned} E[X] &= E\left[\sum_{\omega\in\Omega} X(\omega) I_\omega\right] \\ &= \sum_{\omega\in\Omega} X(\omega) E[I_\omega]\,. \end{aligned} \tag{6}$$

In words: if the possibility space is finite, every expectation functional is defined uniquely by the expectations $E[I_w]$ for each $\omega \in \Omega$.

> **Example 4.1.** A six-sided die is rolled. The possibility space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. Variable X maps $\{1, 3, 5\}$ to zero and $\{2, 4, 6\}$ to one (X is the indicator function of "an even number obtains"). Variable Y maps a face with i pips to $2i$. Suppose the expectations for $E[\omega_i]$ have been assessed, for $i = 1 \ldots 6$. Note that we use $E[w_i]$ instead of the more cumbersome notation $E[I_{w_i}]$. We have $E[X] = E[\omega_2] + E[\omega_4] + E[\omega_6]$ and $E[Y] = \sum_{i=1}^{6} 2iE[\omega_i]$. If $W = XY^2$, then $E[W] = 16E[\omega_2] + 64E[\omega_4] + 144E[\omega_6]$.

We might try to use simple variables to help in computing the expectation for variables in infinite possibility spaces. For a variable X, consider the set of all simple variables Y such that $X \geq Y$. For all those Y, we have $E[X] \geq Y$. It is reasonable to expect that the supremum of $E[Y]$ for all such Y will approximate $E[X]$ rather well:

$$E[X] \approx \sup\left(E[Y] : Y \text{ is simple and } X \geq Y\right).$$

This strategy is indeed successful, in the sense that $E[X]$ is equal to $\sup E[Y]$, in many situations.

4.2 Inequalities: Jensen, Hölder, Cauchy-Scharwz

Inequalities are quite important because they allow us to bound quantities, sometimes quite accurately, even when we have few values of expectations at hand. Some inequalities are so useful that they receive special names.

Theorem 4.1 (Jensen inequality)**.** *If $f(X)$ is a convex function,* $f(E[X]) \leq E[f(X)]$.

Proof. If $f(X)$ is convex, there is a line through the point $(E[X], f(E[X]))$ such that $f(X)$ lies above the line; that is, $f(X) \geq f(E[X]) + \mu(X - E[X])$ for some μ. Use monotonicity: $E[f(X)] \geq E[f(E[X])] + E[\mu(X - E[X])] = f(E[X]) + \mu E[X - E[X]] = f(E[X])$. □

Monotonicity can be used to turn pointwise inequalities into inequalities for expectations. For instance, note that if $\alpha, \beta > 1$ are such that $1/\alpha + 1/\beta = 1$, then for any $a, b > 0$:

$$ab = \exp(\log ab) = \exp((1/\alpha)\log a^\alpha + (1/\beta)\log b^\beta) \leq \exp(\log(a^\alpha/\alpha + b^\beta/\beta)) = a^\alpha/\alpha + b^\beta/\beta.$$

When $a = 0$ or $b = 0$, the inequality $ab \leq a^\alpha/\alpha + b^\beta/\beta$ also holds. Thus we conclude that for any two variables X and Y,

$$E[|X||Y|] \leq E[|X|^\alpha]/\alpha + E\left[|Y|^\beta\right]/\beta \quad \text{whenever } \alpha, \beta > 1 \text{ and } 1/\alpha + 1/\beta = 1.$$

This inequality can in turn be used to prove:

Theorem 4.2 (Hölder inequality)**.** *For real numbers α, β such that $\alpha, \beta > 1$ and $1/\alpha + 1/\beta = 1$,* $E[|XY|] \leq \sqrt[\alpha]{E[|X|^\alpha]}\,\sqrt[\beta]{E[|Y|^\beta]}$.

Proof. Suppose $E[|X|^\alpha] = 0$. As $\alpha > 1$, if $E[|X|^\alpha] = 0$ then $E[|X|^\alpha] \geq E[|X|]^\alpha$ by Jensen inequality and then $E[|X|] = 0$. As $|X|\sup|Y| \geq |XY|$, we have $0 = E[|X|]\sup|Y| \geq E[|XY|] \geq 0$ (the supremum is finite by Assumption 3.1). Likewise, $E[|XY|] = 0$ if $E\left[|Y|^\beta\right] = 0$. So Hölder inequality holds in these cases. Now suppose $E[|X|^\alpha] > 0$ and $E\left[|Y|^\beta\right] > 0$; then:

$$\frac{E[|XY|]}{\sqrt[\alpha]{E[|X|^\alpha]}\,\sqrt[\beta]{E[|Y|^\beta]}} = E\left[\frac{|X|}{\sqrt[\alpha]{E[|X|^\alpha]}}\frac{|Y|}{\sqrt[\beta]{E[|Y|^\beta]}}\right] \leq \frac{1}{\alpha}E\left[\frac{|X|^\alpha}{(\sqrt[\alpha]{E[|X|^\alpha]})^\alpha}\right] + \frac{1}{\beta}E\left[\frac{|Y|^\beta}{(\sqrt[\beta]{E[|Y|^\beta]})^\beta}\right],$$

and the last expression is equal to $1/\alpha + 1/\beta = 1$ by assumption. □

Hölder inequality leads to another famous inequality (by taking $\alpha = \beta = 2$):

Theorem 4.3 (Cauchy-Schwarz inequality)**.** $E[|XY|] \leq \sqrt{E[X^2]\,E[Y^2]}$.

5 Moments, variance and covariance

Expectations of powers of variables are ubiquitous quantities that have special names:

Definition 5.1. The ith *moment* of X is the expectation $E[X^i]$.

Definition 5.2. The ith *central moment* of X is the expectation $E[(X - E[X])^i]$.

Definition 5.3. The *variance* $V[X]$ of X is second central moment of X.

Note that $V[X] \geq 0$ for any X. Moreover,

$$\begin{aligned} V[X] &= E\big[(X - E[X])^2\big] \\ &= E\big[X^2 - 2XE[X] + E[X]^2\big] \\ &= E\big[X^2\big] - E[X]^2 . \end{aligned}$$

Definition 5.4. The *covariance* of variables X and Y is $\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$.

An easy consequence of the Cauchy-Schwarz inequality is

$$\mathrm{Cov}(X, Y) \leq \sqrt{V[X]V[Y]}.$$

If two variables X and Y are such that $\mathrm{Cov}X, Y = 0$, then X and Y are *uncorrelated*.

As an exercise, suppose we have variables $X_1, \ldots, X_n$, all with expectations $E[X_i]$ in the interval $[\underline{\mu}, \overline{\mu}]$. Additionally, suppose X_i and X_j are uncorrelated for any $i \neq j$. If we define the *mean*

$$Y = \frac{\sum_{i=1}^n X_i}{n},$$

then its expectation is

$$E[Y] = \frac{E[\sum_{i=1}^n X_i]}{n} = \frac{\sum_{i=1}^n E[X_i]}{n} \in [\underline{\mu}, \overline{\mu}] \tag{7}$$

and its variance is

$$\begin{aligned} V[Y] &= E\left[\left(\frac{\sum_{i=1}^n X_i}{n} - \frac{\sum_{i=1}^n E[X_i]}{n}\right)^2\right] = (1/n^2)E\left[\left(\sum_{i=1}^n X_i - E[X_i]\right)^2\right] \\ &= (1/n^2)\sum_{i=1}^n E\big[(X_i - E[X_i])^2\big] + (1/n^2)\sum_{i\neq j} E[(X_i - E[X_i])(X_j - E[X_j])] \\ &= (1/n^2)\sum_{i=1}^n V[X_i]. \end{aligned} \tag{8}$$

6 Probabilities

The expectation of an indicator function is be easily interpreted: For any event A, the expectation $E[A]$ indicates how much we expect A to "happen." That is, if we were to get 1 in case the event A obtains and 0 otherwise, how much would the indicator function A be worth?

Such an interesting kind of expectation deserves a special name:

Definition 6.1. The *probability* of event A, denoted by $P(A)$, is equal to $E[A]$.

We can likewise define lower and upper probabilities:

Definition 6.2. The *lower* and *upper* probabilities of event A are respectively

$$\underline{P}(A) = \underline{E}[A] \quad \text{and} \quad \overline{P}(A) = \overline{E}[A]\,.$$

Thus, $\underline{P}(A) = \inf P(A)$ and $\overline{P}(A) = \sup P(A)$. Note that for any event, $\underline{P}(A) = 1 - \overline{P}(A^c)$ because

$$\begin{aligned} \underline{P}(A) &= \underline{E}[A] = -\overline{E}[-A] = -\overline{E}[A^c - 1] = -(\overline{E}[A^c] - 1) \\ &= 1 - \overline{E}[A^c] = 1 - \overline{P}(A)\,. \end{aligned} \tag{9}$$

A *probability measure* is a function that assigns a probability to each event. A probability measure satisfies:

PU1 For any event A, $P(A) \geq 0$.

PU2 The space Ω has probability one: $P(\Omega) = 1$.

PU3 If events A and B are disjoint (that is, $A \cap B = \emptyset$), then $P(A \cup B) = P(A) + P(B)$.

These properties are direct consequences of EU1-EU2.

Proof. For any event, $A \geq 0$; thus EU1 implies $P(A) = E[A] \geq 0$. The indicator function of Ω is identically one, so EU1 leads to $P(\Omega) = 1$. Finally, $P(A \cup B) = E[A \cup B] = E[A + B] = E[A] + E[B] = P(A) + P(B)$ when $A \cap B = \emptyset$ by EU2. Note the simplifying power of de Finetti's convention. □

In the remainder of this section these definitions are illustrated in the context of *finite* possibility spaces, where we obtain from Expression (6):

Theorem 6.1. *If the possibility space is finite, then for any variable X,*

$$E[X] = \sum_{\omega \in \Omega} X(\omega) P(\omega)\,, \tag{10}$$

and for any event A,

$$P(A) = E[A] = \sum_{\omega \in \Omega} A(\omega) P(\omega) = \sum_{\omega \in A} P(\omega)\,. \tag{11}$$

> **Example 6.1.** A six-sided die is rolled. Suppose all states of Ω are assigned precise and identical probability values. We must have $\sum_{\omega \in \Omega} P(\omega) = P(\Omega) = E[\Omega] = E[1] = 1$, thus we have $P(\omega) = 1/6$ for all states. The event $A = \{1, 3, 5\}$ (outcome is odd) has probability $P(A) = 1/2$. The event $B = \{1, 2, 3, 5\}$ (outcome is prime) has probability $P(B) = 2/3$.

These results emphasize a point that may not be immediately obvious from the axioms EU1-EU2: we need only $n - 1$ numbers to completely specify an expectation functional over a possibility space with n states. This follows from the fact that $n - 1$ probabilities are sufficient to specify a probability measure over a possibility space with n states. This simple fact can be used to great effect.

> **Example 6.2.** Consider again Example 3.1, where a person announces her intentions concerning purchase and sale of variables. The person agrees to purchase X_i for up to μ_i; we interpret this to mean $E[X_i] \geq \mu_i$ and write
>
> $$\sum_{\omega} X_i(\omega) P(\omega) \geq \mu_i.$$
>
> Likewise, the person agrees to sell X_i for ν_i or more; we interpret this to mean $E[X_i] \leq \nu_i$ and write
>
> $$\sum_{\omega} X_i(\omega) P(\omega) \leq \nu_i.$$
>
> We have $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and the variables and assessments in Table 1. Using p_i to denote $P(\omega_i)$, we have
>
> $$0 \leq p_1 \leq 2/3, \quad 0 \leq 2p_1 - p_2 \leq 2, \quad 1/6 \leq p_3 \leq 1/3, \quad -1 \leq p_3 - p_1 \leq 0. \tag{12}$$
>
> Each probability measure over Ω can be viewed as a point (p_1, p_2, p_3). All these points must satisfy $p_i \geq 0$ and $\sum_i p_i = 1$, so they live in the simplex depicted in Figure 1. Inequalities (12) specify a set of measures, indicated in the figure by the hatched region. This set of probability measures is *convex* — when P_1 and P_2 belong to the set, their convex combination $\alpha P_1 + (1 - \alpha) P_2$ also belongs to the set for $\alpha \in [0, 1]$.
>
> Take variable X_5 as in Example 3.1. The set of possible values of $E[X_5]$ is the interval
>
> $$\left[\min_{p_1, p_2, p_3} (3p_1 - p_2 + 2p_3), \max_{p_1, p_2, p_3} (3p_1 - p_2 + 2p_3) \right],$$
>
> where the minimum and maximum are subject to inequalities in (12). An exercise in linear programming produces the interval $[11/18, 11/6]$. Thus a transaction $X_5 - \mu_5$ is acceptable for $\mu_5 < 11/18$; likewise, a transaction $\nu_5 - X_5$ is acceptable for $\nu_5 > 11/6$.

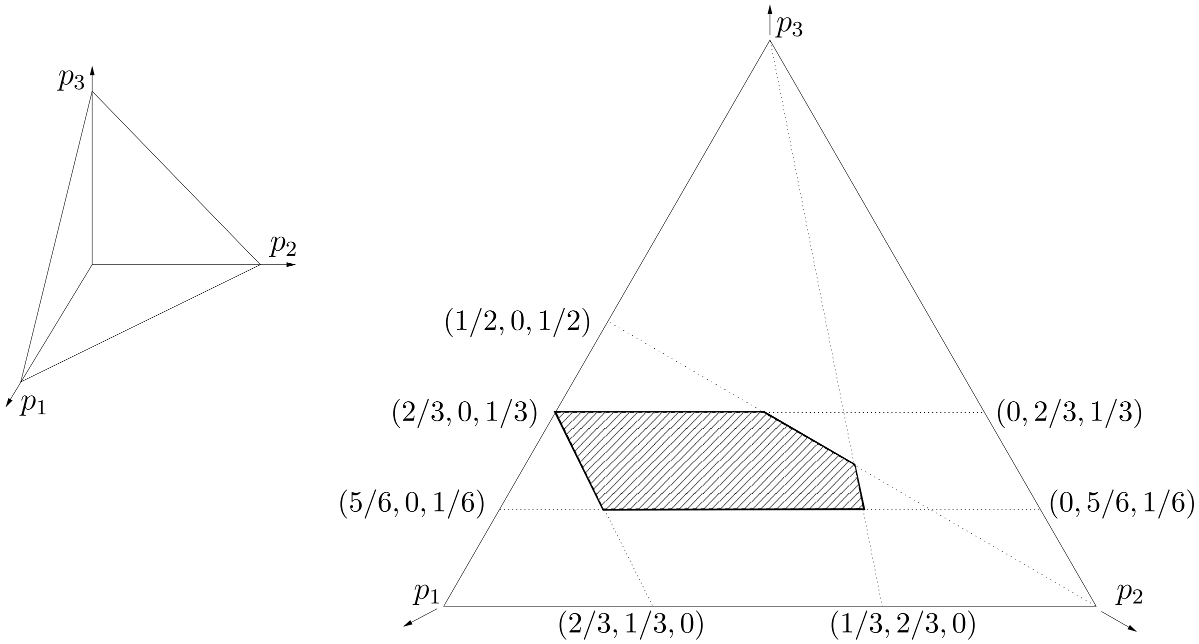

Figure 1: Left: the simplex containing all probability measures in a possibility space with three states. Right: a view from the point (1,1,1) of the simplex and the set of probability measures discussed in Example 6.2.

The right drawing in Figure 1 uses *baricentric coordinates*. These coordinates are quite valuable in the study of sets of probability measures, as they can be used to represent the simplex of all probability measures over three disjoint and exhaustive events. Each vertex of the triangle represents a measure assigning probability one to an event (and probability zero to the other two events). Figure 2 shows a few valuable geometric relations concerning baricentric coordinates. Suppose we have a probability "point" (p_1, p_2, p_3) and we wish to represent it in baricentric coordinates. Calculations are simplified if we write this point as $(\alpha(1-\beta), (1-\alpha)(1-\beta), \beta)$. Clearly $\beta = p_3$. To obtain α, consider two cases. When $p_1 + p_2 = 0$, we have the point $(0, 0, 1)$ and we can choose any α; for instance $\alpha = 1/2$. When $p_1 + p_2 > 0$, then $\alpha = p_1/(p_1 + p_2)$. By placing axes as in the right drawing in Figure 2, we reduce our problem to: find the two-dimensional point (μ, η) that corresponds to point $(\alpha(1-\beta), (1-\alpha)(1-\beta), \beta)$. Relations in the triangles yield $\mu = \sqrt{2}(1/2 - \alpha)(1-\beta)$ and $\eta = \beta\sqrt{3/2}$.

Suppose we have points drawn in baricentric coordinates. To read the coordinates of a point in the triangle, imagine three lines bissecting the angles of the triangle, and read the coordinates on these lines. The coordinates of a point are read by projecting the points to the bissecting lines. Figure 3 illustrates the process.

At this point we digress for a moment, and comment on alternatives in approaching probability theory, still assuming a finite possibility space. We have adopted axioms EU1-EU2 and derived properties PU1-PU3 from Definition 6.1. We could do it differently. We could take

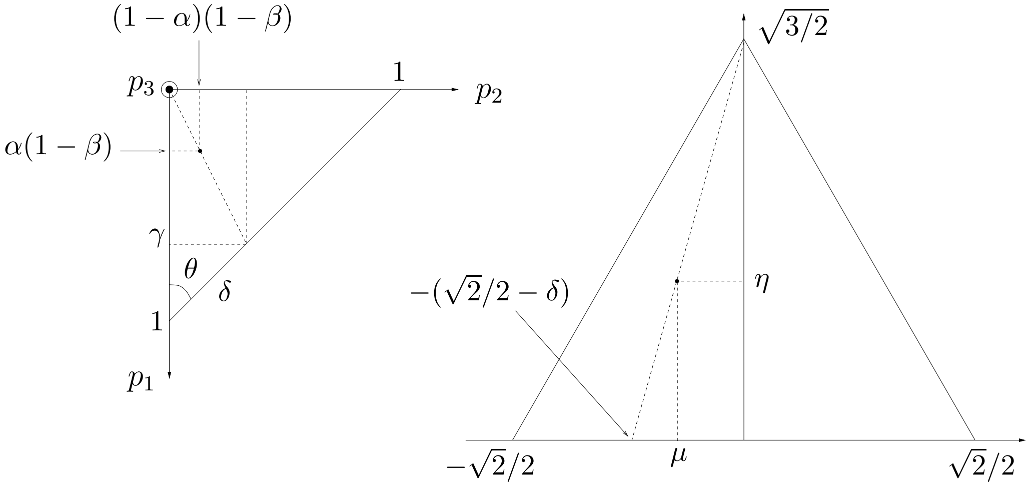

Figure 2: Relationships in baricentric coordinates. Consider a point $(\alpha(1-\beta),(1-\alpha)(1-\beta),\beta)$ and its representation (μ,η). In the left we see a top view of the probability simplex; we have $\alpha(1-\beta)/\gamma=(1-\alpha)(1-\beta)/(1-\gamma)$ and then $\alpha=\gamma$. Using the fact that $\theta=\pi/4$ radians, $\delta=\sqrt{2}(1-\alpha)$. In the larger triangle, we have $-(\sqrt{2}/2-\delta)/\mu=\sqrt{3/2}/(\sqrt{3/2}-\eta)$ and then $(\sqrt{2}(1-\alpha)-\sqrt{2}/2)/\mu=1/(1-\beta)$ because $\eta=\beta\sqrt{3/2}$.

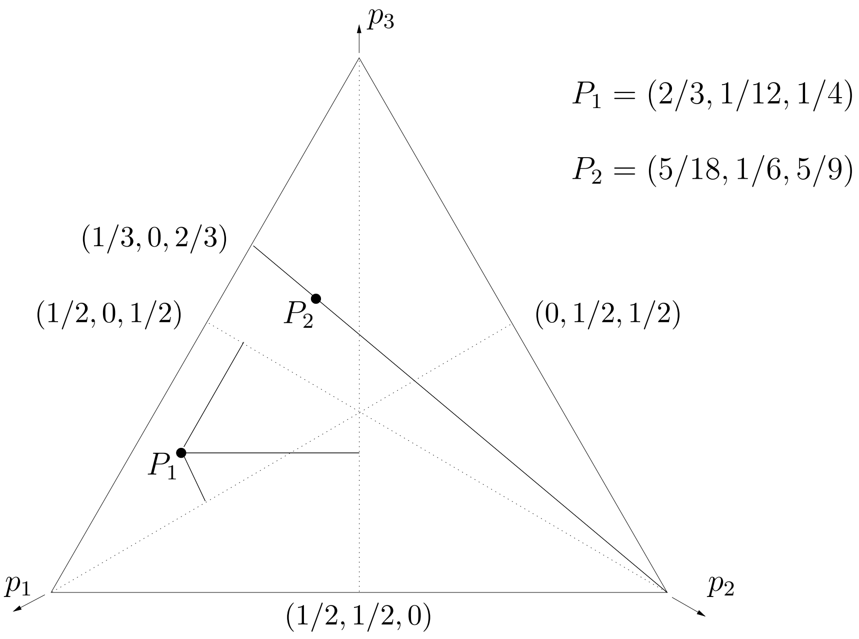

Figure 3: Baricentric coordinates: each point in the triagle represents a probability measure over a possibility space with three states. Each vertex of the triangle assigns probability one to an event; the coordinates of a point are read on the lines bissecting the angles to the triangle.

PU1-PU3 as *axioms*, and *define* expectation as

$$E[X] = \sum_{\omega \in \Omega} X(\omega)P(\omega)\,.$$

That is, probability is the primitive concept and expectation is a derived concept. In this case EU1-EU2 are direct consequences of PU1-PU3 (if $\alpha \leq X \leq \beta$, then $\alpha \leq \sum_{\omega\in\Omega} X(\omega)P(\omega) \leq \beta$; if $Z = X + Y$, then $E[Z] = \sum_{\omega\in\Omega}(X(\omega) + Y(\omega))P(\omega) = E[X] + E[Y]$).

In short: in finite spaces one may start from axioms EU1-EU2 and derive properties PU1-PU3, or one may start from *axioms* PU1-PU3 and derive *properties* EU1-EU2. This interchangeability of axioms becomes more delicate in infinite spaces, depending on the assumptions one is willing to take regarding limit operations.

7 Some properties of probabilities

As the results in this section show, even a few expectations and assumptions can significantly constrain probabilities of interest. In practice we find ourselves relying on a variety of assessments, be them point-valued, interval-valued, or set-valued. Depending on these assessments and on our computational abilities, other probabilities and expectations may be calculated or bounded.

7.1 Complements, unions, intersections

As A and A^c are disjoint and $A \cup A^c = \Omega$, we have $P(A) + P(A^c) = P(\Omega) = 1$ and then

$$P(A) = 1 - P(A^c)\,. \tag{13}$$

Consequently:

$$P(\emptyset) = 1 - P(\emptyset^c) = 1 - P(\Omega) = 0. \tag{14}$$

By applying the finite additivity axiom $n - 1$ times we get:

$$P(\cup_{i=1}^n B_i) = \sum_{i=1}^n P(B_i) \qquad \text{whenever events } B_i \text{ are disjoint.} \tag{15}$$

If events $\{B_i\}$ form a partition of Ω (that is, B_i are mutually disjoint and their union is equal to Ω):

$$P(A) = P(\cup_{i=1}^n A \cap B_i) = \sum_i P(A \cap B_i)\,. \tag{16}$$

An important consequence of the axioms is an expression for $P(A \cup B)$ that does not require A and B to be disjoint:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)\,. \tag{17}$$

The proof is simple: start from $A \cup B = A + B - AB$; use the linearity of expectations and remember that $AB = A \cap B$.

Much more interesting is the following result on arbitrary unions of sets (not necessarily disjoint), directly obtained from Theorem 2.1. The following notation is useful: for a set of events $\{A_i\}_{i=1}^n$, define $S_{j,n}$ as the summation $\sum P\big(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_j}\big)$, where the summation is over all $1 \leq i_1 < i_2 < \cdots < i_j \leq n$. Thus we have $S_{1,n} = \sum_{i=1}^n P(A_i)$, $S_{2,n} = \sum_{1 \leq i_1 < i_2 \leq n} P(A_{i_1} \cap A_{i_2})$, and so on. We take $S_{j,n} = 0$ whenever $j > n$.

Theorem 7.1. *Given events* $\{A_i\}_{i=1}^n$,

$$P(\cup_{i=1}^n A_i) = \sum_{j=1}^n (-1)^{j+1} S_{j,n}. \tag{18}$$

Proof. From Theorem 2.1, we have $\cup_{i=1}^n A_i = \sum_{j=1}^n (-1)^{j+1} A_{j,n}$; taking expectations on both sides we obtain Expression (18) because $S_{j,n} = E[A_{j,n}]$. □

7.2 Inequalities: Fréchet, Bonferroni, Markov, Chebyshev

There are many situations where probability values (or combinations of probability values) cannot be precisely specified. Suppose for example that one assesses the probabilities of two events A and B, but no precise value is given to $P(A \cap B)$. In the absence of further information, all that can be stated is that there is a set of probability values that are consistent with the assessments, so that:

$$\max\left(P(A) + P(B) - 1, 0\right) \leq P(A \cap B) \leq \min\left(P(A), P(B)\right). \tag{19}$$

These are special cases of the *Fréchet bounds*. To obtain the upper bound in Expression (19), note that either A or B may contain the other. To obtain the lower bound, note that events A and B can be disjoint whenever $P(A) + P(B) \leq 1$; if $P(A) + P(B) \geq 1$, we obtain the minimum of $P(A \cap B)$ by setting $P(A \cup B) = 1$ and using Expression (17).

Consider a similar situation, where one deals with a set of events $\{A_i\}_{i=1}^n$ and specifies some of the summations $S_{k,n}$ used in Theorem 7.1. For example, one specifies only $S_{1,n}$ and $S_{2,n}$. The following bounds are examples of *Bonferroni inequalities*, and are used in a large number of fields (actually, the "classic" Bonferroni inequalities are in fact more general in that they bound the probability that any number of successes regarding the B_i).

Theorem 7.2. *Given n events A_i, then for any $m > 0$,*

$$\sum_{j=1}^{2m} (-1)^{j+1} S_{j,n} \leq P(\cup_{i=1}^n A_i) \leq \sum_{j=1}^{2m-1} (-1)^{j+1} S_{j,n}. \tag{20}$$

Proof. The case $n = 2$ is immediate because $P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$ by Expression (17). Now consider an induction argument: suppose the theorem holds for $n -$

1 events. The upper bound on $P(\cup_{i=1}^n A_i)$ is generated by taking expectations on Expression (3) and by keeping only the $2m-1$ terms inside the first parenthesis and the $2m-2$ terms inside the second parenthesis. The induction hypothesis guarantees that terms remaining in the first parenthesis are larger than $P\big(\cup_{i=1}^{n-1} A_i\big)$ and that terms remaining in the second parenthesis are smaller than $P\big(\cup_{i=1}^{n-1} A_n \cap A_i\big)$. The result is certainly larger than $P(\cup_{i=1}^n A_i)$. A similar procedure (keeping $2m$ terms inside the first parenthesis and $2m-1$ inside the second parenthesis) generates the lower bound. □

Expression (19) and Theorem 7.2 deal with situations where some probability assessments are precisely specified, but they are not enough to pin down the value of other probabilities — they only define a set of consistent values. In practice it may even be the case that initial assessments are not precise.

> **Example 7.1.** Suppose $P(A) \in [0.5, 0.6]$ and $P(B) \in [0.6, 0.7]$. These two interval-valued assessments imply $\max(0.5 + 0.6 - 1, 0) = 0.1 \leq P(A \cap B) \leq 0.6 = \min(0.6, 0.7)$.

Clever manipulation of axioms EU1-EU2 can lead to useful inequalities connecting expectations and probabilities. As an example, consider the following celebrated result.

Theorem 7.3 (Markov inequality)**.** *For a nonnegative variable X and a real number $t > 0$,*

$$P(X \geq t) \leq \frac{E[X]}{t}. \tag{21}$$

Proof. If $X(\omega) < t$, then $\{X \geq t\}(\omega) = 0$; thus $X(\omega)/t \geq \{X \geq t\}(\omega)$ because $X(\omega) > 0$ and $t > 0$. If $X(\omega) \geq t$, then $X(\omega)/t \geq 1 = \{X \geq t\}(\omega)$. Consequently, $X/t \geq \{X \geq t\}$ and then $E[X]/t \geq E[X \geq t]$. □

Thus if an assessment $E[X] = \alpha$ is given, we can infer that $P(X \geq t) \leq \alpha/t$ for any $t > 0$ without considering any other assessment.

The Markov inequality leads to an important inequality connecting the expectation and variance of a variable:

Theorem 7.4 (Chebyshev inequality)**.** *For $t > 0$,*

$$P(|X - E[X]| \geq t) \leq \frac{V[X]}{t^2}.$$

Proof. Take $Y = (X - E[X])^2$. As $Y \geq 0$, we can apply the Markov inequality with Y and t^2:

$$P\big((X - E[X])^2 \geq t^2\big) \leq \frac{E\big[(X - E[X])^2\big]}{t^2} = \frac{V[X]}{t^2}$$

Now note that $P\big((X - E[X])^2 \geq t^2\big) = P(|X - E[X]| \geq t)$. □

7.3 Weak laws of large numbers

The Chebyshev inequality can be used to prove the following nice result:

Theorem 7.5. *If variables $X_1, X_2, \ldots, X_n$ have expectations $E[X_i] \in [\underline{\mu}, \overline{\mu}]$ and variances $V[X_i] = \sigma_i^2$, and X_i and X_j are uncorrelated for every $i \neq j$, then for any $\epsilon > 0$ there is $\delta > 0$ such that*

$$\underline{P}\left(\underline{\mu} - \epsilon < \frac{\sum_i X_i}{n} < \overline{\mu} + \epsilon\right) \geq 1 - \frac{\delta}{n}.$$

Proof. Define $Y = (1/n)\sum_i X_i$; then the expectation and variance of Y are μ and $\sum_{i=1}^n \sigma_i^2/n$ respectively (by Expressions (7) and (8)). Apply the Chebyshev inequality: $P(|Y - E[Y]| \geq \epsilon) \leq \sum_i \sigma_i^2/(n\epsilon)^2 \leq (\max_i \sigma_i^2)/(n\epsilon^2)$; that is, $P(E[Y] - \epsilon < Y < E[Y] + \epsilon) \geq 1 - \delta/n$ for $\delta = (\max_i \sigma_i^2)/\epsilon^2$. The result follows as $P\big(\underline{\mu} - \epsilon < Y < \overline{\mu} + \epsilon\big) \geq P(E[Y] - \epsilon < Y < E[Y] + \epsilon)$, because $\underline{\mu} - \epsilon \leq E[Y] - \epsilon$ and $\overline{\mu} + \epsilon \geq E[Y] + \epsilon$. As the result holds for every probability measure satisfying the constraints on expectation and variance, we obtain the bound on the lower probability. □

The message of this theorem is simple yet powerful: if one assesses the expectation of all X_i as μ, then one is forced to believe that a long sequence of X_i will have a mean that is close to μ. This statement is made more dramatic by taking limits:

Corollary 7.1 (Very weak law of large numbers)**.** *If variables $X_1, X_2, \ldots, X_n$ have expectations $E[X_i] \in [\underline{\mu}, \overline{\mu}]$ and variances $V[X_i] = \sigma_i^2$, and X_i and X_j are uncorrelated for every $i \neq j$, then for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \underline{P}\left(\underline{\mu} - \epsilon < \frac{\sum_i X_i}{n} < \overline{\mu} + \epsilon\right) = 1.$$

Proof. For any $\delta' > 0$, choose integer $N > \max_i \sigma_i^2/(\delta'\epsilon^2)$; then $\underline{P}(\overline{\mu} - \epsilon < (1/n)\sum_i X_i < \overline{\mu} + \epsilon) > 1 - \delta'$ for $n > N$ as desired. □

If we are exceptionally knowledgeable about the variables X_i, to the point that we can assess identical and precise expectations $E[X_i] = \mu$ for all of them, we have:

Corollary 7.2 (Weak law of large numbers)**.** *If variables $X_1, X_2, \ldots, X_n$ have expectations $E[X_i] = \mu$ and variances $V[X_i] = \sigma_i^2$, and X_i and X_j are uncorrelated for every $i \neq j$, then for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \underline{P}\left(\left|\frac{\sum_i X_i}{n} - \mu\right| < \epsilon\right) = 1.$$

Note that even when all expectations are precisely specified, we may have more than one probability measure satisfying the given set of assessments. Thus the law still refers to a lower probability.

8 Conditioning

We now wish to define a *conditional* expectation of X given event B, that is to encode our expectation about X upon learning that B has obtained. We adopt:

Definition 8.1. The conditional expectation of X given B, denoted by $E[X|B]$, is

$$E[X|B] = \frac{E[BX]}{P(B)} \quad \text{when } P(B) > 0.$$

How can we justify this definition? Consider an auxiliary variable Y that yields $E[X|B]$ when B obtains and X otherwise:

$$Y = BE[X|B] + (1 - B)X; \qquad \text{that is, } Y(\omega) = \begin{cases} E[X|B] & \text{if } \omega \in B, \\ X(\omega) & \text{otherwise.} \end{cases}$$

Now suppose that $E[Y]$ is equal to $E[X]$. This is a reasonable assumption if $E[X|B]$ is to represent the value of X contingent on B. This simple assumption implies:

$$\begin{aligned} E[X] &= E[Y]\,, \\ E[BX + (1 - B)X] &= E[BE[X|B] + (1 - B)X]\,, \\ E[BX] + E[(1 - B)X] &= E[B]\,E[X|B] + E[(1 - B)X]\,, \\ E[BX] &= P(B)\,E[X|B]\,. \end{aligned} \tag{22}$$

Expression (22) yields

$$E[X|B] = E[BX]\,/P(B) \qquad \text{when } P(B) > 0.$$

When instead $P(B) = 0$, Expression (22) does not constrain $E[X|B]$ because both $E[BX]$ and $P(B)$ are zero.

Given Definition 8.1, it is natural to define conditional probability as follows:

Definition 8.2. The *conditional probability* of event A given event B is defined only when $P(B) > 0$, and it is then equal to $E[A|B]$.

The following celebrated result is a straightforward consequence of Definition 8.2:

Theorem 8.1 (Bayes rule)**.** *If* $P(B) > 0$, *then*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{23}$$

Conditional moments and conditional central moments are defined in the obvious way; for example, the conditional variance of X given event B such that $P(B) > 0$ is

$$V[X|B] = E\big[(X - E[X|B])^2|B\big] = E\big[X^2|B\big] - E[X|B]^2 \,.$$

For two variables X and Y, the conditional expectation $E[X|Y = y]$ can be viewed as a function of Y; for each y such that $P(Y = y) > 0$ we get a real number. That is, $E[X|Y]$ is a variable that is well-defined except when $P(Y = y) = 0$. We can consider the expectation of this variable, $E[E[X|Y]]$. In the special case of a finite possibility space, a nice result is:

Theorem 8.2. *For a finite possibility space,* $E[E[X|Y]] = E[X]$.

Proof. $E[E[X|Y]] = \sum_{Y:P(Y=y)>0} \left(\sum_X xp(x|y)\right) p(y) = \sum_{X,Y:P(Y=y)>0} xp(x|y)\, p(y) = E[X]$. □

9 Properties and examples of conditional probabilities

We now examine a few consequences of the definition of conditional probability through Bayes rule (Expression (26)).

Suppose $\{B_i\}_{i=1}^n$ are events; then the following decomposition is obtained by repeated application of Bayes rule, assuming relevant events have positive probability:

$$\begin{aligned} P(B_1 \cap B_2 \cap \cdots \cap B_n) &= P(B_1) \times P(B_2|B_1) \times \ldots P(B_n|B_{n_1} \cap \cdots \cap B_2 \cap B_1) \\ &= P(B_1) \prod_{i=2}^{n} P\big(B_i| \cap_{j=1}^{i-1} B_j\big) \,. \end{aligned} \tag{24}$$

The following easy result is often called the *total probability theorem*. If events $\{B_i\}$ form a partition of Ω such that all $P(B_i) > 0$, then, using Expression (16):

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)\, P(B_i) \,. \tag{25}$$

From previous derivations we obtain

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B)}{P(A \cap B) + P(A^c \cap B)} = \frac{P(B|A)\, P(A)}{P(B|A)\, P(A) + P(B|A^c)\, P(A^c)},$$

and more generally, if the $\{B_i\}$ form a partition such that $P(B_i) > 0$, then

$$P(B_i|A) = \frac{P(A|B_i)\, P(B_i)}{\sum_i P(A|B_i)\, P(B_i)}. \tag{26}$$

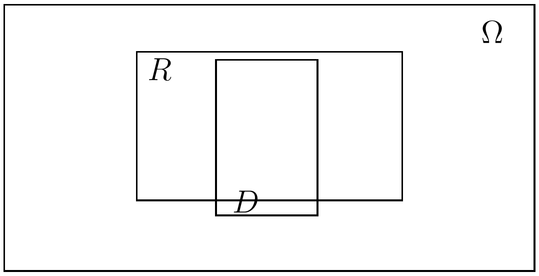

Figure 4: The events in Example 9.1; areas are proportional to probability values.

Example 9.1. A rather pedestrian example can illustrate basic facts about conditioning. Suppose some individuals in an office have a disease D. There is a test to detect the disease; the test produces either result R or R^c. The probability of a result R conditional on the presence of D is $P(R|D) = 9/10$ (the medical literature calls the probability of positive test given disease the *sensitivity* of the test). Also, the probability of a result R^c conditional on the absence of D is $P(R^c|D^c) = 4/5$ (this is called the *specificity* of the test). Suppose also that $P(D) = 1/9$. What it the probability the person is sick if the test produces R? Using Bayes rule:

$$P(D|R) = \frac{P(R|D)\,P(D)}{P(R|D)\,P(D) + P(R|D^c)\,P(D^c)} = \frac{9/10 \times 1/9}{9/10 \times 1/9 + 1/5 \times 8/9} = 9/25.$$

Figure 4 shows the various events in this story, showing their relative measures. It is apparent that the "size" of $D \cap R$ "covers" about one third of the "size" of R.

The medical literature often discusses the following apparent paradox involving conditional probabilities.

Example 9.2. Suppose a person takes a test for a serious disease D where $P(D) = 0.01$. The test has sensitivity and specificity equal to 0.99, so it seems to be a fairly good test. Still, if one takes the test and the result is positive (indicating disease), the probability of actually having a disease is only 0.5! That happens because $P(D|R) = (0.99 \times 0.01)/(0.99 \times 0.01 + 0.01 \times 0.99) = 0.5$. Thus a test for a rare event must have very high sensitivity and specificity in order to be really effective. Drawing a diagram similar to Figure 4 may clarify the situation.

In many practical circumstances one cannot specify conditional probabilities precisely. The following example, describing the *three-prisoners problem*, presents one such situation.

Example 9.3. There are three prisoners, Teddy, Jay, and Mark, waiting to be executed. The prisoners learn that the governor will select one of them to be freed, and that each one of them is equally probable to be selected. Prisoner Teddy learns that the warden knows the governor's decision, and asks the warden about it. The warden does not want to tell Teddy about Teddy's fate, but Teddy convinces the warden to say the name of one of his fellow inmates who will be executed. As one of Jay or Mark is bound to be executed, apparently the warden is not disclosing anything of importance. The warden is known to be an honest person: if the warden says that someone is to be executed, this event will happen for sure. Then the warden says that Jay is to be executed, and suddenly Teddy is happy: he had a chance in three to be freed, and now he has a chance in two (as the decision is between Teddy and Mark). But the same rationale could be applied if the warden had said that Mark were to be executed; so it seems that an irrelevant piece of information is bound to increase Teddy's chance of survival! There is something strange with the conclusion that $P(\text{Teddy freed}) = 1/3$ and $P(\text{Teddy freed}|\text{warden says Jay}) = 1/2$.

Consider the following analysis of the three-prisoners problem. The possibility space has four states:

$$\Omega = \left\{ \begin{array}{c} \text{Teddy freed} \cap \text{warden says Jay}, \\ \text{Teddy freed} \cap \text{warden says Mark}, \\ \text{Jay freed} \cap \text{warden says Mark}, \\ \text{Mark freed} \cap \text{warden says Jay} \end{array} \right\}.$$

We know that

$$P(\text{Teddy freed}) = P(\text{Jay freed}) = P(\text{Mark freed}) = 1/3.$$

Looking at the events in Ω, we note that

$$P(\text{Jay freed} \cap \text{warden says Mark}) = P(\text{Jay freed}) \,,$$

and likewise,

$$P(\text{Mark freed} \cap \text{warden says Jay}) = P(\text{Mark freed}) \,.$$

Concerning Teddy's freedom and the warden's behavior,

$$P(\text{Teddy freed} \cap \text{warden says Jay}) = P(\text{warden says Jay}|\text{Teddy freed})\, P(\text{Teddy freed}) \,,$$

and likewise for $P(\text{Teddy freed} \cap \text{warden says Mark}))$. Thus, to completely specify probabilities over Ω, it is only necessary to assess

$$P(\text{warden says Jay}|\text{Teddy freed}) \,.$$

How would the warden behave if Teddy is to be freed?

Suppose the warden has equal probability of selecting Jay and Mark when Teddy is to be freed; that is, suppose

$$P(\text{warden says Jay}|\text{Teddy freed}) = P(\text{warden says Mark}|\text{Teddy freed}) = 1/2.$$

Then we have:

$$P(\text{Teddy freed} \cap \text{warden says Jay}) = 1/6, \quad P(\text{Teddy freed} \cap \text{warden says Mark}) = 1/6,$$

$$P(\text{Jay freed} \cap \text{warden says Mark}) = 1/3, \quad P(\text{Mark freed} \cap \text{warden says Jay}) = 1/3.$$

Hence

$$P(\text{Teddy freed}|\text{warden names Jay}) = \frac{1/6}{1/3 + 1/6} = 1/3.$$

This is the usual analysis of the three-prisoners problem, yielding the result $P(\text{Teddy freed}) = P(\text{Teddy freed}|\text{warden names Jay}) = P(\text{Teddy freed}|\text{warden names Mark}) = 1/3$.

However, note that the statement of the three-prisoners problem does not say anything about the behaviour of the warden in cases where Teddy is to be freed. The description of the problem does not justify $P(\text{warden names Jay}|\text{Teddy freed}) = 1/2$. It might be the case that the warden will select Jay whenever Teddy is to be freed; that is, $P(\text{warden names Jay}|\text{Teddy freed}) = 1$. Or it might be that the warden will select Mark whenever Teddy is to be freed; that is, $P(\text{warden names Jay}|\text{Teddy freed}) = 0$. Thus all that is really known is

$$P(\text{warden names Jay}|\text{Teddy freed}) \in [0, 1]$$

and consequently

$$P(\text{Teddy freed}|\text{warden names Jay}) \in \left[\frac{0}{0 + 1/3}, \frac{1/3}{1/3 + 1/3}\right] = [0, 1/2].$$

10 Digression: probability zero and conditioning

At this point we should pause and discuss the fact that expectations/probabilities are not constrained given an event of zero probability. Recall that, if $P(B) > 0$, then $P(A|B)$ is clearly defined to be $P(A \cap B) / P(B)$; but if $P(B) = 0$, then $P(A|B)$ is left "undefined." We should be clear on what this means, and on what this *does not* mean.

It is useful to look at a seemingly similar situation that arises in arithmetic, namely, division by zero. If we operate with the real numbers, we can understand the meaning of α/β for any two real numbers; but if β is zero, we cannot figure out the meaning of α/β. One might try to evaluate $\alpha/0$ as some sort of infinity; but, staying within the real numbers, we cannot define $\alpha/0$. So $\alpha/0$ is left "undefined."

One might think that $P(A|B)$ is similarly "undefined" when $P(B) = 0$; however, this is not the case. What "undefined" means here is that *no constraints on $P(A|B)$ are defined when* $P(B) = 0$. It is important to focus on Expression (22):

$$E[BX] = E[X|B]\, P(B)\,.$$

Note that, if $P(B) = 0$, this expression does not define any constraint on $P(A|B)$. Of course one might try to define $P(A|B)$ to be some arbitrary quantity in such cases. but it seems pointless to impose a value on $P(A|B)$ when we have no information to base it. One might give constraints on $P(A|B)$, but they do not affect the underlying *unconditional* measure P, because the feasible values of $P(A|B)$ will be multiplied by zero in the end.

There is, however, an entirely different way to proceed. The idea is to treat conditional probabilities as primitive concepts that are subject to some specific axioms, barring only those conditioning events that are empty sets (those are *really impossible* events).

That is, we again start with Expression (22), but we impose some general constraints that any conditional probability should abide by. The most obvious constraint is that conditional expectations should behave as expectation functionals, with the appropriate modifications (in what follows the conditioning events are always assumed to be nonempty):

EC1 For constants α and β, if $\alpha B \leq BX \leq \beta B$, then $\alpha B \leq E[X|B] \leq \beta B$.

EC2 $E[X+Y|B] = E[X|B] + E[Y|B]$.

By identifying $P(A|B)$ with $E[A|B]$, we then obtain a set of axioms for conditional probability that may be easier to understand than EC1-EC2:

PC1 For any event A, $P(A|B) \geq 0$.

PC2 The space A has conditional probability one: $P(\Omega|A) = 1$.

PC3 If events A and B are disjoint (that is, $A \cap B = \emptyset$), then $P(A \cup B|C) = P(A|C) + P(B|C)$.

Now given these axioms, there is no more use for any "unconditional" expectation functional or probability measure: to obtain one of these, just take the corresponding entity conditioned on the whole space Ω. That is, write $E[X]$ as an abbreviation for $E[X|\Omega]$, and likewise write $P(A)$ as an abbreviation for $P(A|\Omega)$.

All of this would give a rather different status to conditional expectation (and conditional probability), because they would not always be derived from their unconditional counterparts. They would be *primitive* concepts, satisfying their own axioms; the ensuing theory would avoid the annoying clauses "if $P(B) > 0$" because expectations and probabilities would be defined given any nonempty event. The theory would be mathematically quite elegant when it comes to define sets of conditional probability measures. However, this kind of fully-conditional theory does face some difficulties because it clashes with assumptions that are often adopted for *infinite* possibility spaces.

11 Distributions

Suppose we have a possibility space Ω and a probability measure over Ω. Suppose also that we have a variable $X : \Omega \to \Re$. Then there is a possibility space Ω_X containing every possible value of X. The probability measure on Ω induces a measure over subsets of Ω_X by assigning to any event $A \subseteq \Omega_X$:

$$P(X \in A) = P(\{\omega \in \Omega : X(\omega) \in A\}) \,. \tag{27}$$

This definition does produce a function that satisfies PU1-PU3, as can be easily verified. The measure on Ω_X is usually called the *distribution* of X. Note: a distribution is always a measure; it is just a convenient way to stress that a measure is defined over the values of some variable.

A measure over a possibility space Ω induces a single distribution for any variable X. The reverse is not true: a distribution for X generally induces a set of probability measures over Ω:

Example 11.1. Take $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and variable X such that

$$X(\omega_1) = 0, \quad X(\omega_2) = X(\omega_3) = 1.$$

Suppose $P(X = 0) = 1/3$ and $P(X = 1) = 2/3$. This implies:

$$P(\omega_1) = 1/3, \quad P(\omega_2) + P(\omega_3) = 2/3,$$

thus defining a *set* of probability measures over Ω. One measure in this set assigns $P(\omega_1) = 1/3$, $P(\omega_2) = 2/3$, $P(\omega_3) = 0$; another measure assigns $P(\omega_1) = 1/3$, $P(\omega_2) = 0$, $P(\omega_3) = 2/3$.

The *conditional distribution* of variable X given event B is the probability measure over Ω_X such that, for any event $A \subseteq \Omega_X$:

$$P(X \in A|B) = P(\{\omega \in \Omega : X(\omega) \in A\}|B)\,.$$

If $P(B) = 0$, then the right hand side is left “unspecified” (no constraints on it), and likewise for the left hand side.

12 Finite possibility spaces: probability mass functions

In this section the possibility space Ω is assumed finite; a few useful concepts can be defined in this case.

Given a variable X, the *probability mass function* of X, denoted by $p(X)$, is simply

$$p(x) = P(\{X = x\})\,.$$

The probability of any event $A \subseteq \Omega_X$ can be computed using the probability mass function of X:

$$P(X \in A) = \sum_{x \in A} p(x)\,.$$

Example 12.1. Consider a variable X with k values. The *uniform distribution* for X assigns $p(x) = 1/k$ for every value x of X.

Example 12.2. Consider a binary variable X with values 0 and 1. The *Bernoulli distribution with parameter* p for X takes two values: $P(\{X = 0\}) = 1 - p$ and $P(\{X = 1\}) = p$. The expectation of X is $E[X] = 0(1 - p) + 1p = p$. The variance of X is $V[X] = E\big[(X - p)^2\big] = E\big[X^2 - 2pX + p^2\big] = p(1 - p)$.

The expectation of a variable X can be easily calculated using the probability mass function of X:

Theorem 12.1. $E[X] = \sum_X xp(x)$.

Proof. We partition Ω by running through the values of X; thus a summation over $\omega \in \Omega$ can be split into a summation over values of X and an "inner" summation over subsets of Ω:

$$E[X] = \sum_{\omega \in \Omega} X(\omega)P(\omega) = \sum_X \sum_{\omega \in \Omega : X(\omega)=x} xP(\omega) = \sum_X xP(X = x) = \sum_X xp(x)\,.$$

□

Consider a function $Y = f(X)$. Then Y is a function from Ω_X to $\Re$. As Y and X can be combined to take values from Ω to $\Re$, the probability mass function of Y can be computed either by

$$p(y) = P(\{Y = y\}) = \sum_{x \in \Omega_X, Y(x)=y} p(x)\,,$$

or by

$$p(y) = P(\{Y = y\}) = \sum_{\omega \in \Omega, f(X(\omega))=y} P(\omega)\,.$$

We usually simplify the notation by implicitly assuming Ω_X:

$$p(y) = \sum_{Y(x)=y} p(x)\,,$$

as it is clear that we should sum over the possible values of X.

> **Example 12.3.** Consider a variable X with a uniform distribution over the integers from $-k$ to k. Thus Ω_X is the set of integers from $-k$ to k, and $p(x) = 1/(2k+1)$ for $k \in \Omega_X$. Consider now a variable Y such that $Y = X^2$. Then Ω_Y contains the integers from 0 to k^2 that are squares of integers from 0 to k. The distribution function $p(Y)$ is such that $P(y) = \sum_{x \in [-k,k], x^2=y} p(x)$, so we have:
>
> $$p(y) = \begin{cases} \frac{1}{2k+1} & \text{for } y = 0; \\ \frac{2}{2k+1} & \text{for } y \in \Omega_Y,\, y \neq 0. \end{cases}$$

A useful consequence of convexity is:

Theorem 12.2. $E[f(X)] = \sum_X f(x)p(x)$.

Proof. $E[f(X)] = \sum_X \sum_{\omega \in \Omega : X(\omega)=x} f(x)P(\omega) = \sum_X f(x)p(x)$. □

Thus if we are only interested in expectations for a variable X and functions of X, we need not be concerned with the possibility space Ω: all we need is the probability mass function $p(X)$. The possibility space is hidden behind the calculations.

Given two variables X and Y, we can define events such as $\{\{X \in A\} \cap \{Y \in B\}\}$. To simplify the computation of probabilities for such "joint" events, we can use the *joint probability mass function* $p(X, Y)$ of X and Y, defined as:

$$p(x, y) = P(\{X = x\} \cap \{Y = y\})\,.$$

Given a joint probability mass function $p(X, Y)$, we can easily compute the probability mass functions $p(X)$ and $p(Y)$:

$$p(x) = P(\{X = x\}) = \sum_{y \in \Omega_Y} P(\{X = x\} \cap \{Y = y\}) = \sum_{y \in \Omega_Y} p(x, y)\,,$$

and likewise

$$p(x) = P(\{X = x\}) = \sum_{y \in \Omega_Y} p(x, y)\,.$$

Example 12.4. Consider two variables X and Y, with three values each, and with joint probability mass function:

$p(x,y)$	$y = 1$	$y = 2$	$y = 3$
$x = 1$	$1/10$	$1/25$	$1/20$
$x = 2$	$1/20$	$1/5$	$1/25$
$x = 3$	$1/10$	$1/50$	$2/5$

Using this table, we compute $P(\{X \leq 2\} \cap \{Y \geq 2\}) = 1/25 + 1/20 + 1/5 + 1/25 = 33/100$. The marginal probability mass function $p(X)$ is given by $p(1) = 1/10 + 1/25 + 1/20 = 19/100$, $p(2) = 1/20 + 1/5 + 1/25 = 29/100$, and $p(3) = 1/10 + 1/50 + 2/5 = 26/50$.

The *conditional probability mass function* $p(X|B)$ is a function defined only if $P(B) > 0$, as

$$p(x|B) = P(\{X = x\}|B)\,.$$

The *joint conditional probability mass function* for variables $X_1, \dots, X_n$ given an event B such that $P(B) > 0$ is then:

$$p(x_1, \dots, x_n|B) = P(\{X_1 = x_1\} \cap \dots \cap \{X_n = x_n\}|B)\,.$$

Often the conditioning event B is an assignment of values to variables; for example, we may be interested in $p(x_1, x_2|\{X_3 = x_3\} \cap \{X_4 = x_4\})$. Notation is then simplified by omiting references to variables and just writing $p(x_1, x_2|x_3, x_4)$ whenever possible.

13 Credal sets

Given a collection of assessments concerning expectations and probabilities specify, we may form the set of all probability measures that comply with the assessments — as we have already done in several examples.

Definition 13.1. A *credal set* is a set of probability measures.

Denote the set of distributions for variable X by $K(X)$. We can now write the expressions for lower/upper expectations/probabilities using this new notation:

$$\underline{E}[X] = \inf_{P \in K} E[X] \,, \quad \overline{E}[X] = \sup_{P \in K} E[X] \,,$$

$$\underline{P}(A) = \inf_{P \in K} P(A) \,, \quad \overline{P}(A) = \sup_{P \in K} P(A) \,.$$

As we are only dealing with closed credal sets, we can always replace infima by minima and suprema by maxima in these expressions.

The following result is quite useful:

Theorem 13.1. *For closed convex sets, lower and upper expectations are attained at vertices.*

Proof. Suppose a lower/upper expectation is attained at a non-vertex P_α of the credal set $K(X)$. Then P_α can be written as $\alpha P_0 + (1-\alpha)P_1$ for $\alpha \in (0,1)$ and two vertices P_0 and P_1 of the credal set. Denote by $E_\alpha[X]$ the expectation with respect to P_α, and likewise for $E_0[X]$ and $E_1[X]$. Then $E_\alpha[X] = \alpha E_0[X] + (1-\alpha)E_1[X]$. If $E_0[X] = E_1[X] = E_\alpha[X]$ then the vertices also attain the lower/upper expectation. The case $E_0[X] > E_1[X]$ is impossible for it implies that $E_0[X] > E_\alpha[X]$ contradicting the assumption that $E_\alpha[X]$ attains a minimum/maximum, and likewise for the case $E_0[X] < E_1[X]$. □

A credal set $K(X)$ induces a set of conditional distributions $K(X|B)$ by elementwise conditioning; that is, by conditioning every distribution in $K(X)$. The set $K(X|B)$ is called a *conditional credal set*. There are two alternatives, depending on how we treat zero probabilities for B:

Definition 13.2. *Strict conditioning* only defines the conditional credal set K given B when $\underline{P}(B) > 0$; in this case K is the set of all conditional measures $P(\cdot|B)$.

Definition 13.3. *Regular conditioning*, or simply *conditioning*, only defines the conditional credal set K given B when $\overline{P}(B) > 0$; in this case K is the set of all conditional measures $P(\cdot|B)$ such that $P(B) > 0$.

Note that in regular conditioning conditional credal set $K(X|B)$ may be defined even when $\underline{P}(B)$ is equal to zero.

Conditional lower and upper expectations are defined in the obvious way:

Definition 13.4. The *conditional lower* and *conditional upper* expectations of variable X given event B are respectively

$$\underline{E}[X|B] = \inf_{P \in K} E[X|B] \quad \text{and} \quad \overline{E}[X|B] = \sup_{P \in K} E[X|B]$$

whenever $K(X|B)$ is defined.

Likewise for conditional upper and lower probability:

Definition 13.5. The *conditional lower* and *conditional upper* probabilities of event A given event B are respectively

$$\underline{P}(A|B) = \inf_{P \in K} P(A|B) \quad \text{and} \quad \overline{P}(A|B) = \sup_{P \in K} P(A|B)$$

whenever $K(X|B)$ is defined.

Example 13.1. Take a variable X with three values x_1, x_2 and x_3. Suppose $K(X)$ is the convex hull of two distributions:

$$P_1(X = x_1) = 1, \quad P_1(X = x_2) = P_1(X = x_3) = 0;$$

$$P_2(X = x_1) = P_2(X = x_2) = P_2(X = x_3) = 1/3.$$

Regular conditioning leads to

$$\underline{P}(X = x_2|\{X = x_2 \cup X = x_3\}) = \overline{P}(X = x_2|\{X = x_2 \cup X = x_3\}) = 1/2.$$

Strict conditioning leaves these quantities undefined.

There is an analogue of Theorem 13.1 for conditional lower and upper expectations:

Theorem 13.2. *For closed convex sets, conditional lower and upper expectations are attained at vertices.*

Proof. Suppose a conditional lower/upper expectation is attained at a non-vertex P_α of the credal set $K(X)$. Then P_α can be written as $\alpha P_0 + (1 - \alpha)P_1$ for $\alpha \in (0, 1)$ and two vertices P_0 and P_1 of the credal set. Denote by $E_\alpha[X|B]$ the conditional expectation with respect to P_α, and likewise for $E_0[X|B]$ and $E_1[X|B]$ (whenever they are defined). Depending on how the credal set $K(X|B)$ is defined, there are two possible situations.

First, suppose $P_0(B) > 0$ but $P_1(B) = 0$; in this case $E_\alpha[X|B] = E_0[X|B]$. Likewise if $P_0(X) = 0$ and $P_1(X) > 0$ then $E_\alpha[X|B] = E_1[X|B]$.

Second, suppose $P_0(B) > 0$ and $P_1(B) > 0$. Then $E_\alpha[X|B] = \beta E_0[X] + (1 - \beta)E_1[X]$ for $\beta = \alpha P_0(B)/(\alpha P_0(B) + (1 - \alpha)P_1(B))$. If $E_0[X] = E_1[X] = E_\alpha[X]$ then the vertices also attain the lower/upper expectation. The case $E_0[X] > E_1[X]$ is impossible for it implies that $E_0[X] > E_\alpha[X]$ contradicting the assumption that $E_\alpha[X]$ attains a minimum/maximum, and likewise for the case $E_0[X] < E_1[X]$. □

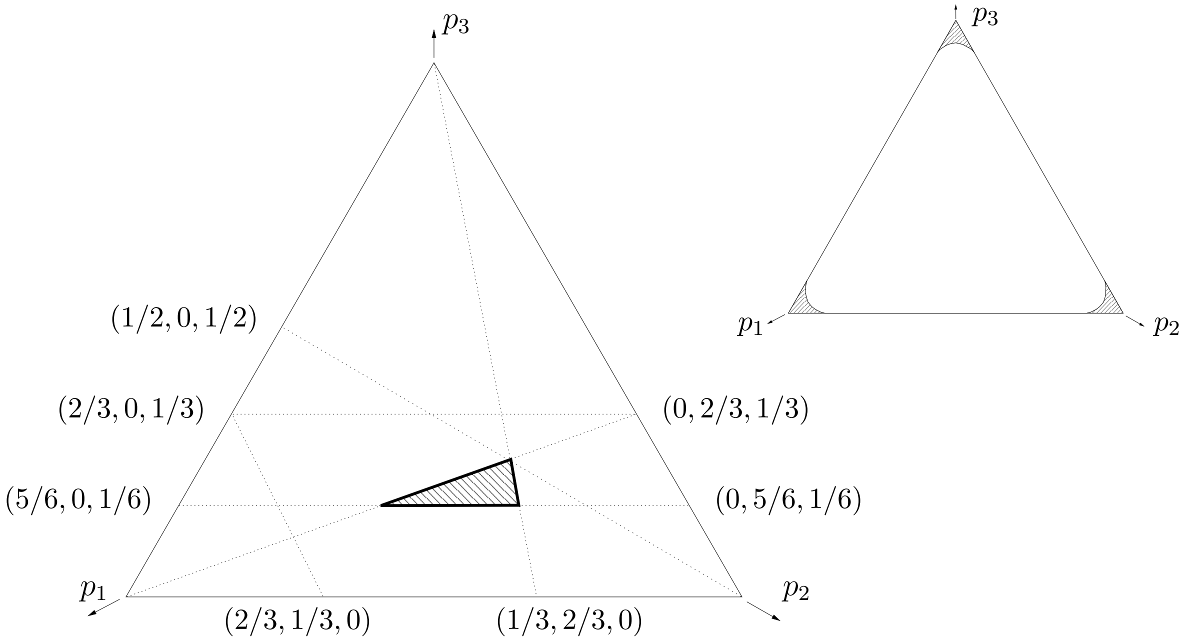

Figure 5: Left: the set of probability measures discussed in Example 13.2: an assessment $E[X|\omega_2 \cup \omega_3] \geq 1$ produces a linear constraint from point $(1, 0, 0)$ to point $(0, 2/3, 1/3)$. Right: the set of probability measures discussed in Example 13.3.

At this point the following picture is complete. One starts with assessments over probability values $P(A_i|B_i)$ and expectations $E[X_j|B_j]$, for various events and variables (where B_j may be Ω). We are then interested in the largest credal set that complies with these assessments; this credal set contains every distribution that is consistent with the available information — hence it is a faithful representation for the assessments.

One may also find that, given a set of assessments, no probability distribution satisfies them. For instance, the assessments $P(A|B) = \mu$, $P(B) = \nu$ and $P(A) = 0$ are inconsistent when $\mu > 0$, $\nu > 0$.

Example 13.2. Consider Example 6.2. Suppose that $E[X_6|\omega_2 \cup \omega_3] \geq 1$ is assessed for a variable X_6 such that $X_6(\omega_1) = 1$, $X_6(\omega_2) = 3$ and $X_6(\omega_3) = -3$. We then have $3p_2/(p_2 + p_3) - 3p_3/(p_2 + p_3) \geq 1$, as $(p_2 + p_3) > 0$ for every valid measure (given the other assessments). Thus the constraint is $3p_2 - 3p_3 \geq p_2 + p_3$; that is, $p_2 \geq 2p_3$. The credal set produced by all assessments is shown in Figure 5. Suppose we want to compute the value of $P(\omega_1|\omega_1 \cup \omega_3)$. This probability cannot be obtained precisely, as every value in the interval $[1/2, 3/4]$ is allowed by the assessments; the lower bound is produced by the point $(1/2, 1/3, 1/6)$. That is, $\underline{P}(\omega_1|\omega_1 \cup \omega_3) = 1/2$ and $\overline{P}(\omega_1|\omega_1 \cup \omega_3) = 3/4$. Now if the assessment $p_2 \geq 5p_3$ were made, the assessements would become inconsistent as no probability measure satisfies all of them.

Example 13.3. To close the chapter, consider a nonconvex credal set. The *entropy* of a variable X is

$$H(X) = -\sum_{x:p(x)>0} p(x) \log_2 (p(x)) \,.$$

This quantity finds application in several fields, from Engineering to Economics, as it tries to capture the amount of "information" in a variable X. The idea is that the information conveyed by a value x of X is $\log_2 (p(x))$: likely values transmit less information. The entropy is then the expectation of information: $H(X) = E[p(X)]$. The entropy of a variable X is zero iff we are practically certain about the value of X; that is, iff for some value x', we have $p(x') = 1$. Very little is transmitted by sending the result of a trial with X, because after all we know that x' will obtain with probability one.

Suppose that a person observes a variable X with three values for a while, and then declares: "The entropy of this variable is very low, because I receive very little information by reading X; in fact, the entropy of X is smaller than or equal to $1/2$." The credal set containing all distributions for X such that $H(X) \leq 1/2$ is depicted in Figure 5. This credal set is not convex. Also note that $\underline{P}(x) = 0$ and $\overline{P}(x) = 1$ for all values of x.

14 Summary

The discussion so far can be conveniently summarized as follows. From a given set of assessments, one then tries to obtain constraints on expectations and probabilities of interest, perhaps reaching the point where a single expectation functional or probability measure can be selected. The idea of an "assessment" has been used quite vaguely so far; an assessment is simply a linear or nonlinear equality or inequality involving probabilities and expectations.

Definition 14.1. A *possibility space* Ω is a set; elements of Ω are called *states*, and subsets of Ω are called *events*.

Definition 14.2. A *random variable* is a function $X : \Omega \to \Re$. The set of values of X is denoted by Ω_X.

Definition 14.3. An *expectation* functional $E[\cdot]$ assigns real numbers to random variables, satisfying: (EU1) For constants α and β, if $\alpha \leq X \leq \beta$, then $\alpha \leq E[X] \leq \beta$; (EU2) $E[X + Y] = E[X] + E[Y]$.

Definition 14.4. The ith *moment* of X is the expectation $E[X^i]$.The ith *central moment* of X is the expectation $E[(X - E[X])^i]$. The *variance* $V[X]$ of X is second central moment of X. The *covariance* $\mathrm{Cov}(X, Y)$ of X and Y is the expectation $E[(X - E[X])(Y - E[Y])]$.

Definition 14.5. A *probability measure* is a set-function $P(\cdot)$ from events to real numbers, satisfying: (PU1) For any event A, $P(A) \geq 0$; (PU2) $P(\Omega|B) = P(B|B) = 1$; (PU3) If $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$.

Definition 14.6. If $P(B) > 0$, the *conditional expectation* of X given B is $E[X|B] = E[BX] / P(B)$.

Definition 14.7. If $P(B) > 0$, the *conditional probability* of A given B is $P(A|B) = P(A \cap B) / P(B)$.

Definition 14.8. A *credal set* is a set of probability measures. *Strict conditioning* only defines the conditional credal set K given B when $\underline{P}(B) > 0$; in this case K is the set of all conditional measures $P(\cdot|B)$. *Regular conditioning*, or simply *conditioning*, only defines the conditional credal set K given B when $\overline{P}(B) > 0$; in this case K is the set of all conditional measures $P(\cdot|B)$ such that $P(B) > 0$.

Definition 14.9. The *lower* and *upper* expectations of variable X are respectively $\underline{E}[X] = \inf_{P \in K} E[X]$ and $\overline{E}[X] = \sup_{P \in K} E[X]$; the *lower* and *upper* probabilities of event A are respectively $\underline{P}(A) = \inf_{P \in K} P(A)$ and $\overline{P}(A) = \sup_{P \in K} P(A)$.

Definition 14.10. The *conditional lower* and *conditional upper* expectations of variable X given event B are respectively $\underline{E}[X|B] = \inf_{P \in K} E[X|B]$ and $\overline{E}[X|B] = \sup_{P \in K} E[X|B]$ whenever $K(X|B)$ is defined; the *conditional upper* and *conditional lower* probability of event A given event B are respectively $\underline{P}(A|B) = \inf_{P \in K} P(A|B)$ and $\overline{P}(A|B) = \sup_{P \in K} P(A|B)$ whenever $K(X|B)$ is defined.

15 Bibliographic notes

Events and random variables are the basic concepts of probability theory. Detailed discussions of these concepts can be found in several excellent textbooks [3, 18]. States are also called elements, points, outcomes, configurations. Usually the term *sample space* is used for the set of states; the term *possibility space* has been proposed by Walley [22] and seems a more accurate description. Possibility spaces should not be viewed as fixed objects; they are often revised (enlarged, reduced, modified) when analyzing a problem. The term *random variable* is somewhat unfortunate as any random variable is actually defined in very deterministic terms. Other possible names are *random quantity* [2], *gamble* [22] and *bet* [11].

The use of identical symbols for events and indicator functions has been pionereed by de Finetti [2]; a nice discussion of this issue is given by Pollard [20]. There is another convention proposed by de Finetti that is quite appealing. His second convention is to use the same symbol P for probability and for expectation; that is, $P(X)$ denotes the expectation of variable X. However appealing, this notation may be confusing when mixed with $p(X)$ (for probability mass functions and densities). For this reason it is not adopted here.

Axioms PU1-PU3 offer a standard approach to probability in finite spaces [3, 18]; axioms EU1-EU2 are their obvious "expectation version" [23]. They are adopted here for any possibility space, finite or infinite. Textbooks usually introduce probability axioms first, and expectation is then defined; here the order is reversed (as in de Finetti's [2] and Whittle's [23] books). There are many reasons to start with expectations. First, expectations are arguably more intuitive from a practical point of view (they are obviously representing averages). Second, the move to infinite spaces is relatively simple for expectations, but very complex for probabilities (particularly when conditioning is discussed on infinite spaces). Finally, many applications deal just with expectations, or are just interested in expectations.

The definition of conditioning follows the usual Kolmogorovian theory [3, 13]; that is, no conditioning on events of zero probability. It would be better to free conditioning from such an unpleasant constraint.

The three-prisoners problem is described in many articles and books [1, 19].

This text emphasizes the need to represent assessments that do not yield a single probability measure. In this case one must employ credal sets. Early advocacy for credal sets can be found in work by Good [10], Kyburg [14], Dempster [4, 5], Shafer [21], and others. The most vocal early proponent of general sets of probability measures was Levi [16, 17], who proposed the term *credal state* to indicate a state of uncertainty that can only be translated into a set of probability measures. Work by Williams [24, 25, 26], Giron and Rios [9] and Huber [12] offered axiomatizations of credal sets. Sets of probability measures were then adopted in the field of robust Statistics and in a number of approaches to uncertainty within the field of Artificial Intelligence. A solid foundation was concocted by Walley [22] and the theory has grown steadily since then. The device of baricentric coordinates appears already in Levi's work [17] and is extensively used in the literature; Lad offers a detailed discussion [15].

The terms "probability mass function" and "joint probability mass function" can be replaced by the more general term *density* [20, Chapter 3]. The presentation here follows a traditional scheme where densities are used for distinct purposes [3].

The literature offers vast material on Markov, Chebyshev, and similar bounds [3, 18, 7], Frechet bounds [6] and Bonferroni bounds [8].

---

16 A few exercises

16.1 Two salesmen are trying to sell different products. The probability that the first salesman will not sell his products is 0.2. The probability that the second salesman will not sell his product is 0.4. What is the probability that neither salesmen will sell anything? What is the probability that both will sell something? What is the probability that exactly one will not sell anything?

16.2 There are four pairs of objects in a table. The first two pairs contain two spoons each. The third pair contains two pencils. A robot selects a pair, then takes the pair and selects an object from the pair. The probability that any of the pairs is selected is 0.25, and the conditional probability that either object is selected given a pair is 0.5.

(a) Suppose the fourth pair contains a spoon and a pencil. What is the possibility space? What is the probability measure on the possibility space? If the selected object is a spoon, what is the probability that the other animal is a spoon as well?

(b) Suppose now the fourth pair of objects contains *either* a spoon and a pencil *or* two pencils. If the selected object is a spoon, what is the lower probability that the other object is a spoon as well?

16.3 A doctor has a client that may have a disease D. The doctor performs a test T that can be positive ($T+$) or negative ($T-$). The test has sensitivity 90% ($P(T+|D) = 0.9$) and specificity 90% ($P(T-|D^c) = 0.9$).

- From medical journals, the doctor assumes that 20% of the population have disease D. Determine $P(D|T+)$.
- Suppose the doctor does not have statistics about the population but thinks that $P(D)$ must be larger than 0.01 but smaller than 0.3. Determine the interval of values of $P(D|T+)$. Can the doctor establish the status of the client? Suppose the doctor uses the interval [0.15, 0.25] for $P(D)$; can he say something definite about the client?

16.4 Two dice are to be thrown; the possibility space contains the 36 possible outcomes. Denote by W the number of pips turned up on the first die and by X the number of pips turned up on the second die. Define $Y = W + X$ and $Z = \min(W, X)$.

- Suppose first all 36 possible outcomes are given equal probability, and determine $p(Y)$, $p(Z)$, $p(Y, Z)$ and $p(Y|Z)$.
- Now suppose the 6 possible values of Z are given equal probability, with no further assessments. What is the lower probability $\underline{P}(Y = 4|Z = 3)$?

16.5 Consider a variable X with 3 possible values x_1, x_2 and x_3, and suppose the following assessments are given: $p(x_1) \leq p(x_2) \leq p(x_3)$; $p(x_i) \geq 1/20$ for $i \in \{1, 2, 3\}$; and $p(x_3|x_2 \cup x_3) \leq 3/4$. Show the credal set determined by these assessments in baricentric coordinates. Obtain lower and upper bounds for the probability $P(x_1|x_1 \cup x_2)$.

17 More exercises

17.1. Prove that $\underline{E}[X + \alpha] = \underline{E}[X] + \alpha$ for any real number α.

17.2. Prove the *Triangle Inequality*:

$$\sqrt{E[(X + Y)^2]} \leq \sqrt{E[X^2]} + \sqrt{E[Y^2]}.$$

(Hint: Note that $E[XY] \leq E[|XY|]$; apply the Cauchy-Scharwz inequality to show that $E[(X+Y)^2] \leq E[X^2] + 2\sqrt{E[X^2]\,E[Y^2]} + E[Y^2]$ and then note that the last expression is equal to $(\sqrt{E[X^2]} + \sqrt{E[Y^2]})^2$.)

17.3. Show that for any two events A and B, the probability that exactly one of them will occur is $P(A) + P(B) - 2P(A \cap B)$. Generalize: show that given any events $\{B_i\}$, the probability that exactly one of them obtains is

$$\sum_{i=1}^{n} (-1)^{i+1} i S_{i,n}.$$

(Hint: use an induction argument inspired by the proof of Theorem 2.1 and take expectations.)

17.4. Suppose that n events $\{B_i\}$ are given together with assessments $P(B_i)$ (and no other assessments). Prove the following bounds:

$$\max\left(\left(\sum_i P(B_i))\right) - (n-1), 0\right) \leq P(\cap_{i=1}^{n} B_i) \leq \min_i P(B_i)\,.$$

17.5. The Kullback-Leibler divergence between distributions P and Q is

$$D(P, Q) = \sum_{x:Q(X=x)>0} P(X = x) \log \frac{P(X = x)}{Q(X = x)}.$$

Show that $D(P, Q) \geq 0$.
(Hint: Use Jensen inequality.)

17.6. Suppose that $g(X)$ is a non-negative non-decreasing function; show that for $t > 0$,

$$P(|X| > t) \leq \frac{E[g(|X|)]}{g(t)}. \qquad (28)$$

This is a generalization of the Markov inequality. Now consider a binary variable X with probability $P(X = 1) = t$. Show that the Markov inequality is in fact an equality in this case. (That is, we cannot improve on Markov inequality without further assumptions on X.) Construct a variable for which the inequality in Expression (28) is an equality (for fixed t).

17.7. Suppose that variable X assumes integer values from 0 to k. Show that

$$E[X] = \sum_{i=0}^{k} P(X > i)\,. \qquad (29)$$

Now show: $\underline{E}[X] \geq \sum_{i=0}^{k} \underline{P}(X > i)$.
(Hint: write X as a sum of indicator functions.)

References

[1] Gert de Cooman and Marco Zaffalon. Updating incomplete observations. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 142–150, San Francisco, California, 2003. Morgan Kaufmann.

[2] Bruno de Finetti. *Theory of probability, vol. 1-2*. Wiley, New York, 1974.

[3] M. H. DeGroot. *Probability and Statistics*. Addison-Wesley, 1986.

[4] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[5] A. P. Dempster. A generalization of Bayesian inference. *Journal Royal Statistical Society B*, 30:205–247, 1968.

[6] M. Denuit, J. Dhaene, M. Goovaerts, and R. Kass. *Actuarial Theory for Dependent Risks*. John Wiley & Sons, Ltd, 2005.

[7] Luc Devroye, Laszlo Gyorfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer Verlag, New York, 1996.

[8] Janos Galambos and Italo Simonelli. *Bonferroni-type inequalities with applications*. Springer-Verlag, New York, 1996.

[9] F. J. Giron and S. Rios. Quasi-Bayesian behaviour: A more realistic approach to decision making? In J. M. Bernardo, J. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics*, pages 17–38. University Press, Valencia, Spain, 1980.

[10] I. J. Good. *Good Thinking: The Foundations of Probability and its Applications*. University of Minnesota Press, Minneapolis, 1983.

[11] J. A. Hartigan. *Bayes theory*. Springer-Verlag, New York, 1983.

[12] P. J. Huber. *Robust Statistics*. Wiley, New York, 1980.

[13] Andrei Nikolaevich Kolmogorov. *Foundations of the theory of probability*. Chelsea Pub. Co. (translation edited by Nathan Morrison, original from 1933), New York, 1950.

[14] H. E. Kyburg Jr. *The Logical Foundations of Statistical Inference*. D. Reidel Publishing Company, New York, 1974.

[15] Frank Lad. *Operational subjective statistical methods: a mathematical, philosophical, and historical, and introduction*. John Wiley, New York, 1996.

[16] Isaac Levi. On indeterminate probabilities. *Journal of Philosophy*, 71:391–418, 1874.

[17] Isaac Levi. *The Enterprise of Knowledge*. MIT Press, Cambridge, Massachusetts, 1980.

[18] A. Papoulis. *Probabilities, Random Variables and Stochastic Processes*. McGraw-Hill, New York, 1991.

[19] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, California, 1988.

[20] David Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press, 2001.

[21] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[22] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[23] Peter Whittle. *Probability*. Penguin, Harmondsworth, 1970.

[24] P. M. Williams. Coherence, strict coherence and zero probabilities. *Fifth Int. Congress of Logic, Methodology and Philos. Sci.*, VI:29–30, 1975.

[25] P. M. Williams. Notes on conditional previsions. Technical report, School of Math. and Phys. Sci., University of Sussex, 1975.

[26] P. M. Williams. Indeterminate probabilities. *Formal Methods in the Methodology of Empirical Sciences*, 1976.