Game Theory:
Review of Probability Theory

Branislav L. Slantchev
Department of Political Science, University of California – San Diego

February 3, 2006

Consider any random process that can generate different outcomes. Let S be the set of these outcomes. That is, we assume a *sample space* S and a set of subsets $A \subset S$, $B \subset S$, $C \subset S$, ... We call subsets $A, B, C, \ldots$ *events*. Note that events may not be individual outcomes but rather collections of outcomes. For any events A and B, we shall interpret $A \cap B = AB$ as the event "both A and B occur" (recall that this is what we had in set theory). Similarly, we shall interpret $A \cup B$ as "A or B occurs," and $\overline{A}$ (or $\neg A$) as "A does not occur."

Suppose we conduct an experiment many times over. We may then observe some event or another, in no regular pattern. The probability of an event is the proportion of times it occurs during these repetitions. For example, the experiment may be a coin toss, and the two events "heads" and "tails". We *could* model the process that causes heads or tails to occur, but this would not be necessary for just about anything we're going to need. Instead, if we assume that the coin is fair, we can model the event "heads" as occurring with probability $1/2$, and the event "tails" as occurring with the same probability.

The probability of an event A is denoted by $\Pr(A) = a$, where a is a real number such that $a \in [0, 1]$. We assume that $\Pr(S) = 1$, that is, *some* event in S will occur with certainty. Further, if $A = \bigcup_{i=1}^{\infty} A_i$, where $A_i \subset X$ and $A_i \cap A_j = \varnothing$ for all $i \neq j$, then

$$\Pr(A) = \sum_{i=1}^{\infty} \Pr(A_i).$$

That is, if A is a union of disjoint (mutually exclusive) subsets of S, then the probability of A occurring is the sum of probabilities of all these subsets occurring.

We say that two events A and B are *independent* if $\Pr(A)$ does not depend on whether B occurs or not, and $\Pr(B)$ does not depend on whether A occurs or not. If events A and B are independent, then their joint probability, that is, the probability that both A and B occur is:

$$\Pr(AB) = \Pr(A)\Pr(B)$$

That is, it is the product of the probabilities that either occurs.

For example, suppose we flip a fair coin twice. Let H denote "heads" and T denote "tails." The sample space consists of the four possible outcomes: $S = \{HH, TH, HT, TT\}$. Because the coin is fair, each of these outcomes has an equal probability. Let's compute these probabilities. Because the coin is fair, the probability of "heads" on any flip equals the probability of "tails": $\Pr(H) = \Pr(T) = 1/2$. Because the result of the second flip does not depend on the result of the first flip and vice versa, the flips are independent. Letting s_1 denote the result of the first flip and s_2 denote the result of the second flip, the probability of any $s \in S$ is then $\Pr(s_1 s_2) = \Pr(s_1)\Pr(s_2)$ for $s_1, s_2 \in \{H, T\}$. For instance, if $s_1 = T$ and $s_2 = H$, then $\Pr(TH) = \Pr(T)\Pr(H) = 1/4$.

Clearly then, $\Pr(HH) = \Pr(TH) = \Pr(HT) = \Pr(TT) = 1/4$. These four events are mutually exclusive and their set is exhaustive. Note that this means that $\Pr(S) = \Pr(HH) + \Pr(TH) + \Pr(HT) + \Pr(TT) = 1$. That is, one of these events will occur for sure.

We can define other events in the two flips of a fair coin experiment. For example, the event "heads on both flips" can be denoted by $Y_1 = \{HH\}$. The event "heads on the first flip" can denoted by $Y_2 = \{HH, HT\}$. The event "the coin comes up heads at least once" can be denoted by $Y_3 = \{HH, HT, TH\}$. You should verify that $\Pr(Y_1) = 1/4$, $\Pr(Y_2) = 1/2$, and $\Pr(Y_3) = 3/4$.

1 The Axioms of Probability Theory

Recall that $\Pr(A)$ denotes the probability of an event A occurring while $\Pr(\overline{A})$ is the probability of event A not occurring. Also $\Pr(A \cup B)$ is the probability of event A or event B occurring (the union of the events), and $\Pr(A \cap B)$ is the probability of event A and event B both occurring (the intersection of the events).

Axiom 1. The probability of any event A is a real number between zero and one:

$$0 \leq \Pr(A) \leq 1.$$

Axiom 2. The probability of a certain event, S, is one:

$$\Pr(S) = 1.$$

Axiom 3. The probability of an event which is the union of two mutually exclusive events is the sum of the probabilities of the two:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) \quad \text{if} \quad A \cap B = \emptyset.$$

Given these axioms, we can derive a bunch of useful identities:

1. For any event A,
$$\Pr(\overline{A}) = 1 - \Pr(A).$$

2. For any two events (i.e. not just mutually exclusive ones):
$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$$

Recalling that $\Pr(A \cap B) = 0$ for mutually exclusive events, we get our axiom above. Rewriting this gives us a formula for the intersection:

$$\Pr(A \cap B) = \Pr(A) + \Pr(B) - \Pr(A \cup B).$$

3. Given any two events (this is very useful and is known as the total probability theorem):
$$\Pr(A) = \Pr(A \cap B) + \Pr(A \cap \overline{B}).$$

4. DeMorgan's Laws apply here as well:

$$\Pr(\overline{A} \cap \overline{B}) = \Pr(\overline{A \cup B})$$
$$\Pr(\overline{A} \cup \overline{B}) = \Pr(\overline{A \cap B})$$

2 Conditional Probability and Bayes Rule

We shall mostly deal with probabilities when we want to represent some player's belief about another player. For example, consider arms reduction negotiations. One player may possess private information about his preferences: He may be "tough" and prefer no deal to even small concessions, or "weak" and prefer significant concessions to no deal at all. The other player has a **prior belief** about this negotiator (possibly based on previous experience, etc.) but in the course of bargaining, as new information accumulates, it is reasonable that this prior belief will change. We shall call the belief updated in the light of new evidence the **posterior belief**.

The classic illustration is from a medical case. Suppose 1% of the population carries the virus X, and so the prior probability that I carry the virus is 0.01. There is an imperfect test for the presence of X: it is positive in 90% of the subjects who carry X and in 20% of the subjects who do not carry X. If I test positive, what is my posterior belief about carrying the virus?

To answer this question, we need to deal with *conditional probability*, that is, the probability of an event A given that event B has occurred. The probability of event A occurring conditional on event B having occurred is the joint probability of the two events occurring divided by the probability of B occurring:

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \tag{1}$$

In words, take the probability of two events occurring jointly, and then incorporate the information that one of them did, in fact, occur. Note that if $\Pr(B) = 0$, then $\Pr(A|B)$ is undefined. You cannot condition on zero-probability events. This problem is going to crop up later, so make sure you remember it. Further, we can get another formula for the intersection of events by rearranging terms in (1):

$$\Pr(A \cap B) = \Pr(A|B) \cdot \Pr(B),$$

which you can extend to an arbitrary number of events. Here's an example with three:

$$\Pr(A \cap B \cap C) = \Pr(A|(B \cap C)) \cdot \Pr(B \cap C) = \Pr(A|(B \cap C)) \cdot \Pr(B|C) \cdot \Pr(C).$$

We can also get a version of the *Total Probability Theorem:*

$$\Pr(A) = \Pr(A|B) \cdot \Pr(B) + \Pr(A|\overline{B}) \cdot \Pr(\overline{B}),$$

which we can generalize to an arbitrary number of mutually exclusive and exhaustive events $\{B_i : i = 1, \ldots, N\}$:

$$B_1 \cup B_2 \cup B_3 \cup \cdots \cup B_N = S, \quad \text{with } B_i \cap B_j = \emptyset \text{ for all } i \neq j.$$

Here's the general version of the theorem:

$$\Pr(A) = \sum_{i=1}^{N} \Big[\Pr(A|B_i) \cdot \Pr(B_i) \Big]. \qquad \text{(Total Probability)}$$

With these results, we can now approach the problem above. Let A be the event that I carry X and let B be the event that I test positive. We are thus given:

$$\begin{aligned} \Pr(A) &= 0.01 \\ \Pr(B|A) &= 0.90 \\ \Pr(B|\overline{A}) &= 0.20 \end{aligned}$$

We do not know either $\Pr(A \cap B)$ or $\Pr(B)$. However, **Bayes' Rule** gives us a way to solve this by expressing the conditional probability $\Pr(A|B)$ in terms of $\Pr(A)$, $\Pr(B|A)$, and $\Pr(B|\overline{A})$:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)\Pr(A) + \Pr(B|\overline{A})\Pr(\overline{A})}. \qquad (2)$$

Here's how you can obtain the formula in (2) from (1). Note that $\Pr(A \cap B) = \Pr(B|A)\Pr(A)$ from (1) by exchanging A and B. This yields the numerator in Equation 2. To see that the denominator equals $\Pr(B)$, note that this is just an application of the total probability theorem. We add (i) the probability of B occurring conditional on A having occurred times the probability of A occurring and (ii) the probability of B occurring conditional on A having not occurred times the probability that A does not occur. Since A can either occur or not (but not both), these are mutually exclusive and exhaustive events, the sum equals the probability of B occurring. That is, this is the total probability theorem.[1]

[1] There may be more possible events. For example, suppose I can only get to the office in only three possible ways: by scooter, shuttle, or bike. Then the probability of "I got to my office" equals to the probability that either "I got to my office on my bike" or "I got to my office on my scooter," or "I got to my office on the shuttle." The point is that since there are three states of the world in which I can get to the office (bike, scooter, or shuttle), summing the probabilities of getting to the office in any of these states yields the probability of getting to the office without reference to any particular state.

We now apply Bayes' Rule to our little medical problem (using the fact that $\Pr(\overline{A}) = 1 - \Pr(A) = 0.99$):

$$\Pr(A|B) = \frac{(0.90)(0.01)}{(0.90)(0.01) + (0.20)(0.99)} = \frac{0.009}{0.009 + 0.198} \approx 0.04,$$

and so we conclude that if I test positive, I have approximately 4% chance of carrying the virus X. This is how the new information from the test has resulted in me updating my **prior belief** that I had the virus (1%) into my **posterior belief** that I have the virus (4%) after receiving the information from the test with the given accuracy.

More generally, let $A_1, A_2, \ldots, A_n$ be a collection of events, exactly one of which must occur, and let B be another event. The probability of A_k conditional on B is

$$\Pr(A_k|B) = \frac{\Pr(B|A_k)\Pr(A_k)}{\sum_{i=1}^{n}\Pr(B|A_i)\Pr(A_i)}. \qquad \text{(Bayes' Rule)}$$

We shall call event B a "signal" and events A_i "states" with every conditional probability $\Pr(B|A_i)$ being either 0 or 1, depending on whether the state A_i generates the signal B or not. In our negotiator example, suppose the signal B is "quit angrily." The two states are "opponent is tough" and "opponent is weak." We shall be interested in the conditional probabilities "weak opponent walks out" (that is, the probability of walking out conditional on the opponent being weak) and "tough opponent walks out" (that is, the probability of walking out conditional on the opponent being tough). We want to be able to calculate the posterior probability so that a player who observes a walk-out can update his prior beliefs given the new information.

Because Bayes rule is essential, we shall now go through several examples of its application. These are taken from the Gintis book.

2.1 The Bolt Factory

In a bolt factory, machines A, B, and C manufacture 25%, 35%, and 40% of the total output, and have defective rates of 5%, 4%, and 2%, respectively. A bolt is chosen at random and is found to be defective. What are the probabilities that it was manufactured by each of the three machines?

First, let's translate the worded problem into symbols to see exactly what information we have. The probability that a randomly chosen bolt was manufactured by machine A is $\Pr(A) = 0.25$, and similarly $\Pr(B) = 0.35$, and $\Pr(C) = 0.40$. Let D denote the event "defective bolt". We now have $\Pr(D|A) = 0.05$, $\Pr(D|B) = 0.04$, and $\Pr(D|C) = 0.02$. We want to know each of $\Pr(A|D), \Pr(B|D)$, and $\Pr(C|D)$. Let's do one, the other two are analogous. By Bayes rule, we have

$$\Pr(A|D) = \frac{\Pr(D|A)\Pr(A)}{\Pr(D|A)\Pr(A) + \Pr(D|\neg A)\Pr(\neg A)}$$

because $\neg A \equiv B \vee C$, we have

$$= \frac{\Pr(D|A)\Pr(A)}{\Pr(D|A)\Pr(A) + \Pr(D|B)\Pr(B) + \Pr(D|C)\Pr(C)}$$
$$= \frac{(0.05)(0.25)}{(0.05)(0.25) + (0.04)(0.35) + (0.02)(0.40)}$$
$$= 0.3623$$

We conclude that given a defective bolt chosen at random, the probability that it came from machine A is 36.23%. You should calculate the other two probabilities and then verify that the three sum to 1. Why should they?

2.2 Murder and Abuse

Let A be the event that a "woman has been abused by her husband", let M be the event that the "woman was murdered", and let H be the event "murdered by her husband." We know (from sociological research) that (a) 5% of women are abused by their husbands, (b) 0.5% of women are murdered, (c) 0.25% of women are murdered by their husbands, (d) 90% of women who are murdered by their husbands have been abused by them, (e) a woman who is murdered but not by her husband is neither more nor less likely to have been abused by her husband than a randomly selected woman.[2]

Suppose a woman is found murdered and the prosecution has ascertained that she was abused by her husband. What is the probability that she was murdered by her husband?

We have to be careful how we ask this question. Are we interested in the probability that a husband murders his wife given that he has abused her? Or are we interested in the probability that a husband has murdered his wife given that he has abused her and that she was murdered? The answer depends on whether you are working for the defense team or the prosecution. Let's calculate both probabilities.

First, what is the probability that an abusive husband murders his wife?

$$\Pr(H|A) = \frac{\Pr(A|H)\Pr(H)}{\Pr(A|H)\Pr(H) + \Pr(A|\neg H)\Pr(\neg H)}.$$

We now need to calculate the various components of this expression. We have $\Pr(A|H) = 0.9$ from (d), $\Pr(H) = 0.0025$ from (c), $\Pr(A|\neg H) = 0.05$ from (a) and (e), and $\Pr(\neg H) = 1 - \Pr(H) = 0.9975$. Plugging in these values yields $\Pr(H|A) = 0.0432$. That is, the probability that an abusive husband murders his

[2] The original version (in Gintis's book) of (e) stated "a woman who is not murdered by her husband is neither more nor less likely to have been abused by her husband than a randomly selected woman." Thanks to Boris Demeshev from the Higher School of Economics (Moscow, Russia) for spotting a mistake in this. The problem is that the assumptions above imply $P(A) = P(A|H)$ because $P(H) > 0$. This is from $P(A) = P(A|H)P(H) + P(A|\neg H)P(\neg H)$, which from (e) gives us $P(A) = P(A|H)P(H) + P(A)P(\neg H)$, which simplifies to $P(A)P(H) = P(A|H)P(H)$, yielding the contradiction.

wife is measly 4.32%, which would seem like a strong argument in the husband's presumed innocence.

However, this ignores one piece of additional information we have. Namely, the fact that event M has occurred. What, then, is the probability that a husband murders his wife given that he has abused her and that she was murdered?

$$
\begin{aligned}
\Pr(H|AM) &= \frac{\Pr(AMH)}{\Pr(AM)} \\
&= \frac{\Pr(AMH)}{\Pr(AMH) + \Pr(AM\neg H)} \\
&= \frac{\Pr(A|MH)\Pr(MH)}{\Pr(A|MH)\Pr(MH) + \Pr(A|M\neg H)\Pr(M\neg H)} \\
&= \frac{\Pr(A|MH)\Pr(H|M)\Pr(M)}{\Pr(A|MH)\Pr(H|M)\Pr(M) + \Pr(A|M\neg H)\Pr(\neg H|M)\Pr(M)} \\
&= \frac{\Pr(A|MH)\Pr(H|M)}{\Pr(A|MH)\Pr(H|M) + \Pr(A|M\neg H)\Pr(\neg H|M)}
\end{aligned}
$$

since from (e) we know that $\Pr(A|M\neg H) = \Pr(A)$, this yields

$$
= \frac{\Pr(A|MH)\Pr(H|M)}{\Pr(A|MH)\Pr(H|M) + \Pr(A)(1 - \Pr(H|M))}
$$

because $H \subset M \Rightarrow M \cap H = MH = H$, this yields

$$
= \frac{\Pr(A|H)\Pr(H|M)}{\Pr(A|H)\Pr(H|M) + \Pr(A)(1 - \Pr(H|M))}
$$

and since $\Pr(H|M) = \Pr(HM)/\Pr(M) = \Pr(H)/\Pr(M)$, this yields

$$
= \frac{\Pr(A|H)\Pr(H)}{\Pr(A|H)\Pr(H) + \Pr(A)(\Pr(M) - \Pr(H))}.
$$

As before, we know that $\Pr(A|H) = 0.9$ from (d), and that $\Pr(H) = 0.0025$ from (c). Further, $\Pr(A) = 0.05$ from (a), and $\Pr(M) = 0.005$ from (b). Calculating the probability gives us $\Pr(H|AM) = 0.9474$. That is, the probability that an abusive husband has killed his murdered wife is a whooping 94.74% which would seem like a strong argument about the husband's guilt.

2.3 The Monty Hall Game

Let's just apply Bayes' Rule to an interesting example. You are a contestant in a game show. There are three closed doors, with a car behind one and dead goats behind the other two. You may choose any door. Since you have no information other than that, your prior about where the car is (we assume you prefer the car to a dead goat) is $^1/_3$ probability for each door. So you pick door 1. Monty (the show host) opens door 2 and shows you that there is a dead goat behind it. He then asks, "Would you now like to change your choice?" What should you do?[3]

There are really two cases depending on what you want to assume about the way Monty makes his choices. We shall show that the answer to the question depends on whether you assume that he chooses *randomly* from all doors except the one you picked, or *non-randomly* by never choosing to open the door with the car behind it.

2.3.1 The Three-Door Example

Consider the contest with just three doors and suppose (without loss of generality) that the contestant chooses door 1. Monty opens door 2 and shows him the goat. What next? We want to know if the contestant can gain from switching his choice to door 3. So we have to compare the probability of winning a car by staying with door 1, and the probability of winning by switching to door 3.

Let's get some notation to facilitate exposition. Let A be the event "car is behind door 1," let C be the event "car is behind door 3," and let B be the event "Monty picks door 2 and there is a goat behind it." Since from the contestant's initial perspective, the car is equally likely to be behind any of the three doors, $\Pr(A) = \Pr(C) = 1/3$. We want to know $\Pr(A|B)$ and $\Pr(C|B)$. Bayes rule tells us that:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} = 1/3 \times \frac{\Pr(B|A)}{\Pr(B)}$$
$$\Pr(C|B) = \frac{\Pr(B|C)\Pr(C)}{\Pr(B)} = 1/3 \times \frac{\Pr(B|C)}{\Pr(B)}.$$

First, suppose Monty randomly picks one of the remaining doors. The probability that he picks door 2 is then $1/2$. The probability that there's a goat behind any given door is $1 - 1/3 = 2/3$, so the probability that Monty picks door 2 and it has a goat behind it is $\Pr(B) = 1/2 \times 2/3 = 1/3$ because the two events are independent by our assumption that Monty picks randomly. We now need to determine $\Pr(B|A)$ and $\Pr(B|C)$. If the car is behind door 1, then door 2 certainly has a goat behind it, and hence the probability that Monty picks it and it has a goat behind it will equal the probability that Monty picks it: $\Pr(B|A) = 1/2$. Similarly, if the car is behind door 3, then door 2 also has a goat behind it for sure, yielding $\Pr(C|A) = 1/2$. Putting all this together yields:

$$\Pr(A|B) = \frac{1/3 \times 1/2}{1/3} = 1/2$$
$$\Pr(C|B) = \frac{1/3 \times 1/2}{1/3} = 1/2.$$

[3]This problem caused quite a controversy on the Internet a (long) while back until it was finally banned from the newsgroups. The problem was that most people who posted it did not realize there are really two cases they should deal with. Fortunately, we do realize that (an example of strategic thinking), so we shall cover both of them.

Hence, the contestant cannot gain from switching. This is not surprising: if Monty makes a random choice, his action reveals nothing.

Suppose now that Monty would never pick a door with a car behind it (much more likely in a real show). As before, $\Pr(B|A) = 1/2$ because if the car is behind door 1, then both doors 2 and 3 certainly have goats behind them, and so Monty will pick randomly between them. However, $\Pr(B|C) = 1$ because if the car is behind door 3, then door 2 certainly has a goat behind it, but since Monty never reveals the car (and cannot open door 1 because the contestant picked it), he will certainly pick door 2. Finally, because Monty will never pick a door with a car behind it, $\Pr(B) = 1/2$. You can use the total probability theorem to obtain this number. Let "2" denote the event "the car is behind door 2." Then:

$$\begin{aligned}
\Pr(B) &= \Pr(B|A)\Pr(A) + \Pr(B|2)\Pr(2) + \Pr(B|C)\Pr(C) \\
&= 1/2 \times 1/3 + 0 \times 1/3 + 1 \times 1/3 \\
&= 1/6 + 1/3 = 1/2.
\end{aligned}$$

We used the fact that the probability of the event "Monty picks door 2 and it has a goat behind it" is zero if there is a car behind that door: $\Pr(B|2) = 0$. This now yields:

$$\begin{aligned}
\Pr(A|B) &= \frac{1/2 \times 1/3}{1/2} = 1/3 \\
\Pr(C|B) &= \frac{1 \times 1/3}{1/2} = 2/3.
\end{aligned}$$

Since $\Pr(C|B) > \Pr(A|B)$, the contestant should definitely switch. The reason is that Monty's informed choice reveals additional information which the contestant should incorporate in his beliefs.

2.3.2 The General Case

Suppose there are $n \geq 3$ doors. Let A_k be the event "car is behind door k." Let B_k be the event "Monty chooses door k, which contestant did not choose, and door k has a goat behind it." The prior probability is $\Pr(A_k) = 1/n$ for all k. Bayes' Rule then gives

$$\Pr(A_i|B_k) = \frac{\Pr(B_k|A_i)\Pr(A_i)}{\Pr(B_k)} = \left(\frac{1}{n}\right)\frac{\Pr(B_k|A_i)}{\Pr(B_k)} \tag{3}$$

Without loss of generality, suppose that the contestant chooses door 1, and so he wins if A_1. That is, $\Pr(\text{Win}) = 1/n$.

Random Choice. Suppose Monty chooses randomly from all remaining doors $k > 1$. He will open a particular door with probability $1/(n-1)$, and since the probability that it has a goat behind it is $1 - (1/n) = (n-1)/n$, it means that $\Pr(B_k) = [(n-1)/n][1/(n-1)] = 1/n$ for all $k > 1$. If A_1, then door k certainly has a goat behind it, and so $\Pr(B_k|A_1) = 1/(n-1)$. From (3), we have

$$\Pr(A_1|B_k) = \frac{1/(n-1)}{n(1/n)} = \frac{1}{n-1} \text{ for all } k > 1.$$

Since Monty picks randomly (as if he did not know what door had the car behind it), he chooses doors $2, 3, \ldots, n$ with equal probability. Moreover, whatever door he chooses certainly has a goat behind it since $k \neq i$. From $\Pr(B_k|A_i) = 1/(n-1)$ and (3) we get

$$\Pr(A_i|B_k) = \frac{1/(n-1)}{n(1/n)} = \frac{1}{n-1} \text{ for } i \neq k, i > 1.$$

That is, if A_i for $i > 1$, then for $k \neq 1, i$, the probability of B_k conditional on A_i is $1/(n-1)$, which is exactly the same as $\Pr(A_1|B_k)$. The contestant cannot gain from switching.

Non-Random Choice. Of course, one would expect Monty to know where the car is and not reveal it when he opens the door for it would spoil the fun. Suppose therefore that he never picks the door with the car behind it. We now have $\Pr(B_k|A_1) = 1/(n-1)$, as before, but $\Pr(B_k)$ is now also $1/(n-1)$ because Monty will certainly pick one of the $n-1$ doors without a car behind it. From (3) we have

$$\Pr(A_1|B_k) = \frac{1/(n-1)}{n[1/(n-1)]} = \frac{1}{n} \text{ for all } k > 1.$$

For $i > 1$ and $k \neq i$, $\Pr(B_k|A_i) = 1/(n-2)$ since Monty must now choose randomly among all doors except 1 and i. From (3) we get

$$\Pr(A_i|B_k) = \frac{1/(n-2)}{n[1/(n-1)]} = \frac{n-1}{n(n-2)} \text{ for } i \neq k, i > 1.$$

Since $(n-1)/[n(n-2)] > 1/n$, the contestant should switch. The idea is that when acting strategically (i.e. non-random selection), Monty conveys additional information by opening a door. Since Monty *knows* there's no car behind the door he picks, the contestant is able to update his estimate about where the car actually is. A random pick by Monty, of course, does not reveal anything the contestant did not know before.

3 Probability Distributions

Suppose we have three events, A, B, and C. The assignment of probabilities to events is called a **probability distribution**. For example, one such distribution may assign probabilities $(0, 1, 0)$ to these events. That is, according to this distribution, event B occurs with certainty, and events A and C do not occur for sure. Another probability distribution may be $(1/3, 1/3, 1/3)$. According to it, these events occur with equal probability. Clearly, there is an infinite number of possible probability distributions.

In each probability distribution, the sum of probabilities of all possible events is 1 (recall that this implies that one of the events must occur). Further, for each event, the probability assigned by the distribution must be, as before, some real number $a \in [0, 1]$.

We shall be dealing extensively with probability distributions. For example, suppose a decision-maker faces an uncertain situation in which outcomes are probabilistic. As we shall see next time, this defines a "lottery" over the outcomes, which is just another probability distribution. Also, suppose the decision-maker has three different options to choose from. The choice of each option is called a "pure strategy." An important concept is the "mixed strategy" where the decision-maker randomizes among the pure strategies. That is, instead of choosing one of them with certainty, the decision-maker chooses among them according to some probability distribution. Mixed strategies are but probability distributions over the space of pure strategies.

As we shall see, it is often too cumbersome to work with outcomes and events directly, and it is much more convenient to associate numbers with them. A numerical representation is called a **random variable.**[4] That is, a random variable X is a function that maps all possible outcomes to real numbers, or $X : S \to \mathbb{R}$. That is, $X(s) = x$ means that the real number x corresponds to outcome $s \in S$. Events then become sets of real numbers, and we can define the probability distributions over random variables. Letting A denote any subset of $\mathbb{R}$, and $\Pr(X \in A) = \Pr(s|X(s) \in A)$ denote the probability that X is in this set, the probability distribution specifies $\Pr(X \in A)$ for all A.

The most common way of specifying a probability distribution is with a **probability function**. When a random variable can take only a finite number of outcomes, we say that it has a **discrete distribution**. When it has a discrete distribution, the probability function is defined as $f(x) = \Pr(X = x)$ for any $x \in \mathbb{R}$. If A is the set of all possible values that X can take, then $f(x) = 0$ for any $x \notin A$. We also require that $f(x) \in [0, 1]$ and $\sum_{x \in A} f(x) = 1$.

For example, consider an experiment that consists of five tosses of a biased coin that comes up heads with probability $1/3$ and tail with probability $2/3$. The sample space has $2^5 = 32$ elements, so I won't enumerate it. We are interested in the total number of heads, so let X denote that number. For example, $X(HTTHH) = 3$. Clearly, the set of all possible realizations of the random variable is $A = \{0, 1, 2, 3, 4, 5\}$. The probability function takes any $x \in A$ and returns the probability associated with that realization of X.

What is $f(5)$? There is only one event with 5 heads, and it requires that the coin comes up heads in each and every toss. The probability of this event is $(1/3)^5 = 1/243$. Similarly, the probability of the event no heads at all is $f(0) = (2/3)^5 = 32/243$. You could enumerate the rest, or you could use the well-known formula for the *binomial distribution* that returns the probability of x successes in n trials, where each trial results in success with probability p and failure with probability $1 - p$:

$$f(x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

[4]The usage of "random" does not correspond to the everyday one. Colloquially, we take "random" to be synonymous with "equally likely." With a random variable, we mean that it takes values in a non-deterministic way. That is, when we repeat an experiment, we do not know for sure what the outcome will be. There is no requirement that the possible outcomes are equally likely.

for all $x = 0, 1, \ldots, n$. Recalling that $\binom{n}{x} = \frac{n!}{x!(n-x)!}$, we can easily compute other values of f. For example,

$$f(2) = \binom{5}{2} (1/3)^2 (2/3)^3 = 10 \times 1/9 \times 8/27 = 80/243.$$

Let's check. There are $\binom{5}{2} = 10$ outcomes with exactly two heads: HHTTT, HTHTT, HTTHT, HTTTH, THHTT, THTHT, THTTH, TTHHT, TTHTH, TTTHH. Each of these has the same probability of occurring, and it is the probability of getting two heads and three tails, which is $(1/3)^2 \times (2/3)^3 = 8/243$. Since these 10 outcomes are mutually exclusive, we just add the 10 individual probabilities and obtain the answer above.

What happens if X can take on an infinite number of values? In this case, we say it it is a *continuous random variable* or that it has a **continuous distribution**. Defining the probability function is a bit trickier now because $\Pr(X = x) = 0$. Instead, the non-negative function f is such that for any interval $A = [a, b]$,

$$\Pr(X \in A) = \Pr(a \le X \le b) = \int_a^b f(x)\,\mathrm{d}x.$$

That is, the **probability density function** (pdf) f is some function such that the area under its curve between two points a and b in the range of X equals the probability that X will take on some value between these points. It is important to realize that the pdf f may *not* return values that are probabilities because unlike the discrete case, it does not assign probabilities directly. Rather, its integral assigns these probabilities. Hence, the two requirements for a pdf are $f(x) > 0$ and $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$.

One very common and convenient distribution is the *uniform (rectangular) distribution* defined on some interval $[a, b]$. When X is distributed uniformly, the probability that it lies in some subset of the interval with a given length equals the probability that it lies in any other subset that has the same length. This means that the pdf must have the same value, say $f(x) = c$, for any point $x \in [a, b]$ and 0 everywhere else. We now have:

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x = \int_a^b c\,\mathrm{d}x = 1.$$

We now need to find the precise value of c, which we can obtain by solving the equation:

$$\int_a^b c\,\mathrm{d}x = 1 \Leftrightarrow c(b-a) = 1 \Leftrightarrow c = \frac{1}{b-a}.$$

Hence, the uniform pdf is defined as follows:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise.} \end{cases}$$

It is worth repeating (yet again!) that the pdf does not give us a probability value, its integral does. One very common and convenient interval we shall see repeatedly is $[0, 1]$, in which case the pdf reduces to $f(x) = 1$ for all $x \in [0, 1]$.

3.1 Cumulative Distributions

Since we are usually interested in probabilities of events, the probability density function is not directly useful because it does not return probabilities. Obviously, we will need to use its integral for this purpose. The most convenient definition involves calculating the probability that X takes on a value that is at most equal to some x. The **cumulative distribution function** (cdf) F takes a number x and returns the probability that X will take on a value lower than this number:

$$F(x) = \Pr(X \le x).$$

For example, going back to our discrete case where X denoted the number of heads from 5 tosses of the biased coin, we have:

$$\begin{aligned}
F(0) &= \Pr(X \le 0) = f(0) = 1 \times {}^{32}/_{243} = {}^{32}/_{243} \\
F(1) &= \Pr(X \le 1) = F(0) + f(1) = {}^{32}/_{243} + 5 \times {}^{16}/_{243} = {}^{112}/_{243} \\
F(2) &= \Pr(X \le 2) = F(1) + f(2) = {}^{112}/_{243} + 10 \times {}^{8}/_{243} = {}^{192}/_{243} \\
F(3) &= \Pr(X \le 3) = F(2) + f(3) = {}^{192}/_{243} + 10 \times {}^{4}/_{243} = {}^{232}/_{243} \\
F(4) &= \Pr(X \le 4) = F(3) + f(4) = {}^{232}/_{243} + 5 \times {}^{2}/_{243} = {}^{242}/_{243} \\
F(5) &= \Pr(X \le 5) = F(4) + f(5) = {}^{242}/_{243} + 1 \times {}^{1}/_{243} = {}^{243}/_{243} = 1.
\end{aligned}$$

In general, if X is a discrete random variable, then:

$$F(x) = \sum_{u \le x} f(u).$$

The continuous case is analogous except that instead of summing we take the integral:

$$F(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u.$$

The definition of the cdf yields an alternative definition of the pdf:

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}.$$

In other words, the pdf $f(x)$ is the first derivative of a cdf $F(x)$, with respect to x where the derivative exists. Hence, a continuous random variable can be represented either by the pdf or the cdf.

Returning the uniform distribution example, consider the cumulative probability distribution function F, for which $F(a) = 0$ (i.e. the probability of a number less than a is 0) and $F(b) = 1$ (the probability of a number at most equal to b is 1). So, $F(x)$ represents the probability that X will take on a value at most equal to x with the property that subsets of the interval that have the same size have the same probability. We use the pdf we derived above and obtain:

$$F(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u = \int_{a}^{x} \frac{1}{b-a}\,\mathrm{d}u = \frac{x-a}{b-a}.$$

Suppose $a = 0$ and $b = 2$. The probability of $x \in [.25, .35] = F(.35) - F(.25) = .35/2 - .25/2 = .05$, which is the same as the probability that $x \in [1.85, 2]$, which is, of course, why the distribution is called "uniform." When the interval is $[0, 1]$ (i.e. $a = 0$ and $b = 1$), then $F(x) = x$.

When the number of events is infinite, the probability of any particular event is 0, and so we must rely on a cumulative distribution function to describe probabilities of sets of events. To recover the probability for x from F, all we need to do is calculate $F(x) - F(x')$, i.e. the difference between values of F at x and the next smaller event x'. In the continuous case, we have $\Pr(X = x) = \lim_{\epsilon \to 0} F(x + \epsilon) - F(x)$.

3.2 Properties of the cdf and Expectation

Let's recap. Every random variable has a distribution function. There are a few things about these functions that are useful to remember. First, the cumulative distribution function (cdf) is a probability, so its values always lie in the interval between zero and one:

$$0 \le F(x) \le 1.$$

Also, the probability must be non-decreasing:

$$F(x) \le F(y) \quad \text{for} \quad x < y.$$

Finally, because it cannot exceed 1 or be less than zero, we have:

$$\lim_{x \to \infty} F(x) = 1 \quad \text{and} \quad \lim_{x \to -\infty} F(x) = 0.$$

The density function, on the other hand, is *not* necessarily a probability! That is, it can easily take on values that exceed one. However, it cannot be negative, and so we have:

$$f(x) \ge 0 \quad \text{for all} \quad x.$$

The one important property is that the area under the curve defined by this function must be exactly one (this immediately follows from the definition of the cdf):

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1.$$

One crucial concept we shall see over and over again should already be familiar to you if you've taken the intro to statistics class. This is the idea of *expectation.* The **expected value** of a random variable generalizes the mean value (or, in everyday parlance, the average). If X is a continuous variable with density $f(\cdot)$, then its expected value is:

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x,$$

and if X is a discrete variable, it is

$$\mathrm{E}[X] = \sum x f(x).$$

For example, go back to the number of heads from five tosses of a biased coin. What is the expected number of heads? Letting x_i denote i heads:

$$\begin{aligned} \mathrm{E}[X] &= \sum_{i=0}^{5} x_i f(x_i) \\ &= 0(^{32}/_{243}) + 1(^{80}/_{243}) + 2(^{80}/_{243}) + 3(^{40}/_{243}) + 4(^{10}/_{243}) + 5(^{1}/_{243}) \\ &= {}^{405}/_{243}. \end{aligned}$$

We should expect to get fewer than 2 heads (approximately 1.67) in five consecutive tosses of the biased coin.

We can generalize the idea of expectation to functions, and so the **expected value of a function** is:

$$\mathrm{E}[g(x)] = \int_{-\infty}^{\infty} g(x) f(x)\,\mathrm{d}x$$

for the continuous case, where $g(x)$ is some function of the random variable whose density is given by $f(x)$. The analogous case for discrete variables (which we shall be dealing with almost exclusively) is:

$$\mathrm{E}[g(x)] = \sum g(x) f(x).$$

The expectation of a constant is just the constant:

$$\mathrm{E}[a] = a.$$

The expectation operator is linear, and so the following always holds:

$$\mathrm{E}[aX + b] = a\,\mathrm{E}[X] + b.$$

for any two constants a and b. A bit more generally, you should remember that

$$\mathrm{E}[aX + bY + c] = a\,\mathrm{E}[X] + b\,\mathrm{E}[Y] + c.$$

This also works for multiplication as follows:

$$\mathrm{E}[aX \cdot bY] = ab\,\mathrm{E}[XY].$$

Letting $\mu_x = \mathrm{E}[X]$ and $\mu_y = \mathrm{E}[Y]$, in general it is **not** the case that $\mathrm{E}[XY] = \mu_x \mu_y$. This only holds if X and Y are independent random variables (you should review what statistical independence means). Finally, for linear combinations of many variables, we have:

$$\mathrm{E}\left[\sum_{i=1}^{N} \alpha_i X_i\right] = \sum_{i=1}^{N} \alpha_i\,\mathrm{E}[X_i].$$

This you can get by applying the summation rules and the fact that the expectation is linear. All of these will come in handy very soon when we prove the expected utility theorem.