Lecture 6

Finishing Conditional Probability

Review Exercises

1. Let $A, B, C$ be independent events with probabilities $p_A, p_B, p_C$ of occurring, respectively.
   
   (a) What is the probability all 3 events occurring?
   
   (b) What is the probability of none of them occurring?
   
   (c) What is the probability that exactly 1 occurs?

2. You have a 6-sided die that gives the value $i$ with probability $p_i$, where $i = 1, \ldots, 6$. If you roll the die repeatedly, what is the chance you will roll 3 before you get 4?

3. State the definition of events $A, B, C$ being conditionally independent given the event $D$.

4. Assuming the size of a set is roughly proportional to its probability in the following diagrams, which picture most closely depicts independent events $A, B$?

   ![Diagram](image)

Solutions

1. (a) $p_A p_B p_C$
   
   (b) $(1 - p_A)(1 - p_B)(1 - p_C)$
   
   (c) $p_A(1 - p_B)(1 - p_C) + (1 - p_A)p_B(1 - p_C) + (1 - p_A)(1 - p_B)p_C$

2. Using the idea from review exercise 2 in Lecture 5 we have

   \[ P(\text{roll 3}|\text{roll 3 or 4}) = \frac{p_3}{p_3 + p_4}. \]
3. The events $A, B, C$ should be independent with respect to the probability measure $P_D(E) = P(E|D)$. More precisely, we need

$$P(AB|D) = P(A|D)P(B|D), \quad P(AC|D) = P(A|D)P(C|D), \quad P(BC|D) = P(B|D)P(C|D),$$

and

$$P(ABC|D) = P(A|D)P(B|D)P(C|D).$$

4. The left picture.

**Fallacies and Pitfalls**

**Example 1** (Prosecutor’s Fallacy). In a famous trial, a woman was charged with murder when two of her infants died. An expert witness made the claim that the probability of a single infant death to occur randomly is 1 in 8500, so two deaths would be $(1/8500)^2$, or approximately 1 in 73 million. They said this probability is so unlikely, that she must be guilty. The first issue is that the events are definitely not independent. Even if they were, note that the above confuses the probability $P(\text{evidence}|\text{guilty})$ with $P(\text{guilty}|\text{evidence})$, and thus gets an entirely incorrect answer.

**Example 2** (Simpson’s Paradox). Suppose you hear that cigar smokers have a higher mortality rate from smoking related illnesses than cigarette smokers. What does this imply about the health risk of cigars vs. cigarettes? This is a complex issue, but let’s consider one small part of it. Would it surprise you to learn that amongst older people (let’s say above 60) cigars have a lower mortality rate, and the same is true for younger people? Must this data be inconsistent? The answer is no, and we will now see why.

More precisely, let our sample space be smokers in our study (of which there are equally many cigarette and cigar smokers), let $D$ be the event of a person dying from a smoking related illness, let $G$ be the event of being a cigar smoker, and let $O$ be the event of being older. Then

$$P(D|G) > P(D|G^c),$$
$$P(D|G, O) < P(D|G^c, O),$$
$$P(D|G, O^c) < P(D|G^c, O^c).$$

The explanation is that cigar smokers tend to be older than cigarette smokers, and older people die more often than younger people in general. More precisely,

$$P(D|G) = P(D|G, O)P(O|G) + P(D|G, O^c)P(O^c|G),$$
$$P(D|G^c) = P(D|G^c, O)P(O|G^c) + P(D|G^c, O^c)P(O^c|G^c).$$

As a concrete (fabricated) example, suppose there are 100 of each type of smoker with the following data:
The table of mortality information conditioned on cigar smoking is:

<table>
<thead>
<tr>
<th>Cigar</th>
<th>Younger</th>
<th>Older</th>
<th>Cigarette</th>
<th>Younger</th>
<th>Older</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoke Related Death</td>
<td>0</td>
<td>50</td>
<td>Smoke Related Death</td>
<td>30</td>
<td>9</td>
</tr>
<tr>
<td>Not Smoke Related</td>
<td>10</td>
<td>40</td>
<td>Not Smoke Related</td>
<td>60</td>
<td>1</td>
</tr>
</tbody>
</table>

Then cigars vs. cigarettes is 50/100 vs. 39/100, but for younger people it is 0/10 vs. 30/90 and for older people it is 50/90 vs. 9/10.

**Fallacy Exercises**

The table of mortality information conditioned on cigar smoking is:

<table>
<thead>
<tr>
<th>Cigar</th>
<th>Younger</th>
<th>Older</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoke Related Death</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>Not Smoke Related</td>
<td>10</td>
<td>40</td>
</tr>
</tbody>
</table>

Compute the following quantities based on the table where \( G \) denotes cigar smoking, \( O \) denotes older, and \( D \) denotes smoke related death.

1. \( P(D \cap O|G) \), \( P(D \cap O^c|G) \)
2. \( P(D^c \cap O|G) \), \( P(D^c \cap O^c|G) \)
3. \( P(O|G) \), \( P(O^c|G) \)
4. \( P(D|G, O) \), \( P(D|G, O^c) \)
5. \( P(D|G) \)

**Solutions**

1. 
   \[ P(D \cap O|G) = \frac{50}{100}, \quad P(D \cap O^c|G) = \frac{0}{100}. \]

2. 
   \[ P(D^c \cap O|G) = \frac{40}{100}, \quad P(D^c \cap O^c|G) = \frac{10}{100}. \]

3. 
   \[ P(O|G) = \frac{90}{100}, \quad P(O^c|G) = \frac{10}{100}. \]

4. 
   \[ P(D|G, O) = \frac{50}{90}, \quad P(D|G, O^c) = \frac{0}{100}. \]

5. 
   \[ P(D|G) = \frac{50}{100}. \]
Discrete Random Variables

Functions

A function \( f : X \rightarrow Y \) assigns to each element of \( X \) (the domain) exactly one element of \( Y \) (the codomain). The image (or range) of a function is the set \( f(X) \) defined by

\[
f(X) = \{ f(x) : x \in X \}.
\]

We say that a function is injective or one-to-one if \( f(x_1) = f(x_2) \) implies \( x_1 = x_2 \) for all \( x_1, x_2 \in X \). In the language of calculus/precalculus, injective functions pass the horizontal line test, and thus can be inverted (i.e., we get a function \( f^{-1} : f(X) \rightarrow X \)). We say a function is onto or surjective if \( f(X) = Y \), or in other words, every element of the codomain is assigned to some element of the domain by \( f \). Given any function (even non-invertible functions) we can form something called the preimage. More precisely, if \( A \subset Y \) then we have

\[
f^{-1}(A) = \{ x \in X : f(x) \in A \},
\]

called the preimage of \( A \) under \( f \). The functions we will deal will be real-valued (they may also take values in \( \mathbb{R}^n \), but this can just be looked at as a list of \( n \) real-valued functions). As such, we can perform algebra on functions in a natural way. For instance, if \( f, g : X \rightarrow \mathbb{R} \) then

\[
(f + g)(x) = f(x) + g(x) \quad \text{and} \quad (f^2)(x) = f(x)^2.
\]

Function Exercises

1. Let \( |X| = 5 \) and \( |Y| = 7 \).
   (a) How many functions \( f : X \rightarrow Y \) are there?
   (b) How many of them are injective?
   (c) How many are surjective?
   (d) (**) How many have images of size 3?

2. Let \( f : \mathbb{R} \rightarrow \mathbb{R} \) be defined by \( f(x) = x^2 \). What is \( f^{-1}([1, 2]) \)? In other words, what is \( \{ x : f(x) \in [1, 2] \} \)?

3. Let \( X = \{(d_1, d_2) : 1 \leq d_i \leq 6 \} \) and \( Y = \mathbb{R} \). If \( f, g : X \rightarrow Y \) with \( f(x_1, x_2) = x_1 \) and \( g(x_1, x_2) = x_2 \), then what is \( (f + g)^{-1}(\{7\}) = \{ x \in X : (f + g)(x) = 7 \} \)?

Solutions

1. (a) There are \( 7^5 \) possible functions.
   (b) There are \( 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \) possible injective functions.
   (c) There are no possible surjective functions.
(d) We first solve a slightly simpler problem. Suppose \(|Z| = 3\). How many surjective functions \(f : X \to Z\) are there? The total number of functions from \(X\) to \(Z\) is \(3^5\). The number of non-surjective functions from \(X\) to \(Z\) is \((\binom{3}{2})^2 5 - (\binom{3}{1})^1 5\).

This formula is computed via inclusion-exclusion:

\[
\sum_{A \subset Z} |F_A| - \sum_{A,B \subset Z} |F_A F_B|
\]

where \(F_A\) is the number of functions \(g : X \to A\). Thus the number of surjective functions is \(3^5 - (\binom{3}{2})^2 5 + (\binom{3}{1})^1 5\). As there are \((\binom{7}{3})\) choices for \(Z\) the final answer is

\[
\binom{7}{3} \left(3^5 - (\binom{3}{2})^2 5 + (\binom{3}{1})^1 5\right).
\]

2. \([-\sqrt{2}, 1] \cup [1, \sqrt{2}]\).

3. \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}.

Random Variables

Having built up the structure of probability theory, we need a way to represent random quantities (like the number of heads in a sequence of flips, etc.). We do this using random variables. Formally, a random variable is a function \(X : S \to \mathbb{R}\) where \(S\) is our sample space. We typically use capital letters near the end of the alphabet for random variables. [Technically a random variable is a measurable function on the sample space, but we are going to ignore this point, since we aren’t developing measure theory.] You can think of a random variable as a label on each point in the sample space.

Example 3 (Coin Flips). Suppose we flip a fair coin 3 times. Let \(S\) be our standard associated sample space: \(S = \{(f_1, f_2, f_3) : f_i \in \{H, T\}\}\).

Define \(X : S \to \mathbb{R}\) to be the number of heads that occurred. Then \(X\) is a random variable.

Random Variable Exercises

1. Let \(X\) be a random variable, and let \(A \subset \mathbb{R}\). What does \(P(X \in A)\) mean? If \(x \in \mathbb{R}\), what does \(P(X = x)\) mean?

2. Suppose we flip \(n\) coins that are heads with probability \(p\), and let \(X\) denote the number of heads we get. What are \(P(X = 0)\), \(P(X = 1)\), \(P(X = 2)\), and \(P(X = 3)\)?

3. Roll 2 fair 6-sided dice. Let \(X\) represent the first die, and \(Y\) the second. What is \(P(X + Y = 7)\)?
Solutions

1. We define $P(X \in A)$ by

$$P(X \in A) = P(\{s \in S : X(s) \in A\}) = P(X^{-1}(A)).$$

Note that $P(X = x)$ is the same as $P(X \in \{x\})$. Thus we can write

$$P(X = x) = P(\{s \in S : X(s) = x\}) = P(X^{-1}(\{x\})).$$

In both of these cases we have take preimages to turn statements about the value of $X$ into events (subsets of $S$) whose probability can be determined.

2. The first step is to state the sample space

$$S = \{(f_1, \ldots, f_n) : f_i \in \{H, T\}\}.$$

We use a general finite sample space with outcome probability given by

$$P(\{(f_1, f_2, \ldots, f_n)\}) = p^\text{# of heads}(1 - p)^\text{# of tails}.$$

Then $X : S \to \mathbb{R}$ is defined by

$$X((f_1, \ldots, f_n)) = \# \text{ of } f_i \text{ that are } H.$$

We then have

$$P(X = 0) = (1 - p)^n,$$

$$P(X = 1) = \binom{n}{1} p(1 - p)^{n-1},$$

$$P(X = 2) = \binom{n}{2} p^2(1 - p)^{n-2},$$

$$P(X = 3) = \binom{n}{3} p^3(1 - p)^{n-3}.$$

As an example, we explicitly calculate $P(X = 2)$ using the definition:

$$P(X = 2) = P(\{(f_1, \ldots, f_n) : \# \text{ of } f_i \text{ that are } H \text{ is } 2\}) = \binom{n}{2} p^2(1 - p)^{n-2}.$$

Here there are $\binom{n}{2}$ different sequences of $n$ flips with 2 heads and each has a probability of $p^2(1 - p)^{n-2}$ assigned to it.

3. Using question 3 from the previous set of exercises, we have $P(X + Y = 7) = \frac{6}{36} = \frac{1}{6}$. 

6