

# Lecture 17: Probability

*10:30 AM, Mar 31, 2009*

## Contents

<b>1</b>	<b>Overview</b>	<b>1</b>
<b>2</b>	<b>Introduction</b>	<b>1</b>
<b>3</b>	<b>Axioms of Probability</b>	<b>2</b>
3.1	Properties of Probability Functions . . . . .	2
<b>4</b>	<b>Conditional Probability</b>	<b>3</b>
4.1	Independence . . . . .	4
<b>5</b>	<b>Random Variables</b>	<b>5</b>
5.1	Expectation and Variance . . . . .	6
5.2	Correlation and Covariance . . . . .	8

## 1 Overview

During the last unit, our investigations of AI were rooted in logic; in particular, we studied “symbolic” AI, in which all statements were assigned binary truth values. We now embark upon a study of what might be called “numeric” AI. Rather than associate discrete values with statements, we associate continuous values, or probabilities. This representational flexibility is intended to model uncertainty in an agent’s environment.

## 2 Introduction

Gambling was the primary force driving the early development of probability theory. As early as the 16th century, gamblers noticed that there are empirical laws which govern the frequencies of the various outcomes in a game of chance, even though the precise outcome cannot generally be predicted in advance. For example, Cardano noticed that for many simple games of chance, each outcome is realized in proportion to the reciprocal of the total number of outcomes. Cardano’s observation applies to rolling dice and flipping coins, for example.

Suppose an experiment (such as spinning a red and black roulette wheel) is repeated  $N$  times. Let  $\#(A)$  denote the number of times the outcome  $A$  (e.g., “landed on red”) is observed. The ratio of  $\#(A)$  to  $N$  is called the **relative frequency** of  $A$ . Empirically, when  $N$  is large, this relative

frequency approximates some  $p \in \mathbb{R}$ : i.e.,

$$\frac{\#(A)}{N} \approx p$$

Since  $p$  depends on  $A$ , it is usually written  $P(A)$ , and it is called the **probability** of  $A$ . Sometimes the symmetric nature of the experiments renders all outcomes are equally likely, as in the games of chance studied by Cardano. In this case,

$$P(A) = \frac{|A|}{|\Omega|}$$

We write  $\Omega$  to denote the **sample space** of possible outcomes, with  $\omega \in \Omega$  and  $A \subseteq \Omega$ .

**Example:** Consider an experiment in which an unbiased coin is flipped twice in succession. In this experiment, the sample space  $\Omega = \{HH, HT, TH, TT\}$ , with each outcome equally likely. If  $A$  is the event “at least one head,” then  $P(A) = 3/4$ .

Similarly, consider an experiment in which an unbiased coin is flipped thrice in succession. In this experiment, the sample space  $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$ , with each outcome equally likely. If  $B$  is the event “at most one tail,” then  $P(B) = 1/2$ .

### 3 Axioms of Probability

The following three axioms characterize the probability function  $P : 2^\Omega \rightarrow \mathbb{R}$ :

1.  $P(\Omega) = 1$
2.  $0 \leq P(A) \leq 1$ , for all events  $A \subseteq \Omega$
3.  $P(A \cup B) = P(A) + P(B)$ , for all events  $A, B \subseteq \Omega$  s.t.  $A \cap B = \emptyset$

The third axiom is often expressed as its inductive equivalent, namely

$$3'. \text{ given the finite set of disjoint events } \{A_1, A_2, \dots, A_n\}, \\ P\left(\bigcup_{i=1}^n A_i\right) = P(A_1) + P(A_2) + \dots + P(A_n)$$

#### 3.1 Properties of Probability Functions

Like probabilities, set theory also provides a foundation for the meaning of logical formulas. If  $\Omega$  is viewed as the set of all “possible worlds” (i.e., interpretations), then the meaning of proposition  $A$  is the set of its models, denoted  $M(A) \subseteq \Omega$ . Moreover,  $M(A \wedge B) = A \cap B$ ,  $M(A \vee B) = A \cup B$ , and  $M(\neg A) = A^c$ , for arbitrary propositions  $A$  and  $B$ . This isomorphism allows us to use our logical toolkit to reason about probabilities.

Using the axioms of probability and the laws of logic, we now derive several useful properties of probability functions.

Since  $A = A \wedge \top = A \wedge (B \vee \neg B) = (A \wedge B) \vee (A \wedge \neg B)$ , and since the intersection of  $(A \cap B)$  and  $(A \cap B^c)$  is empty, it follows from the third axiom that

$$P(A) = P(A \cap B) + P(A \cap B^c) \quad (1)$$

Letting  $A = \Omega$  yields  $P(\Omega) = P(\Omega \cap B) + P(\Omega \cap B^c) = P(B) + P(B^c)$ , from which we conclude by the first axiom that for all  $B \subseteq \Omega$ ,  $P(B^c) = 1 - P(B)$ . In particular, letting  $B = \Omega$  yields  $P(\emptyset) = P(\Omega^c) = 1 - P(\Omega) = 1 - 1 = 0$ .

For arbitrary propositions  $A$  and  $B$ ,

$$\begin{aligned} A \vee B &= (A \wedge \top) \vee (B \wedge \top) \\ &= (A \wedge (B \vee \neg B)) \vee (B \wedge (A \vee \neg A)) \\ &= (A \wedge B) \vee (A \wedge \neg B) \vee (B \wedge A) \vee (B \wedge \neg A) \\ &= (A \wedge B) \vee (A \wedge \neg B) \vee (B \wedge \neg A) \end{aligned}$$

Using Equation 1, it now follows that for  $A$  and  $B$  not necessarily disjoint,

$$\begin{aligned} P(A \cup B) &= P(A \cap B) + P(A \cap B^c) + P(B \cap A^c) \\ &= P(A \cap B) + (P(A) - P(A \cap B)) + (P(B) - P(A \cap B)) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

**Example:** Once again, let  $\Omega = \{\text{HH, HT, TH, TT}\}$ . If  $A$  is the event “at least one head,” then  $A = \{\text{HH, HT, TH}\}$ ; if  $B$  is the event “at least one tail,” then  $B = \{\text{HT, TH, TT}\}$  and  $B^c = \{\text{HH}\}$ . Now  $A \cap B = \{\text{HT, TH}\}$  and  $A \cap B^c = \{\text{HH}\}$ . Thus, by Equation 1,  $P(A) = 2/4 + 1/4 = 3/4$ . Moreover,  $P(B^c) = 1/4 = 1 - 3/4 = 1 - P(B)$ . Finally,  $P(A \cup B) = 3/4 + 3/4 - 2/4 = 1$ : *i.e.*, with probability 1, the event “at least one head or at least one tail occurs.”

## 4 Conditional Probability

The **conditional probability**  $P(A | B)$  of  $A$  given  $B$  is defined as

$$P(A | B) = \frac{P(A \cap B)}{P(B)} \quad (2)$$

The (unconditional) probability of an event  $A$  is in fact the conditional probability of event  $A$  given sample space  $\Omega$ : *i.e.*,  $P(A | \Omega) = P(A)$ . Conditional probabilities are sometimes called **posterior** probabilities, in which case unconditional probabilities are called **prior** probabilities.

Continuing our running example,  $P(A | B)$  denotes the probability of observing at least one head, given at least one tail. This event occurs with probability  $2/3$ , since the two of the three outcomes in  $B$  which include at least one tail also include at least one head, namely HT and TH.

One important consequence of Equation 2 is the **product rule**:

$$P(A \cap B) = P(A | B)P(B) = P(B | A)P(A) \quad (3)$$

Intuitively, the probability of observing two events is the probability of observing the former, given the latter, times the probability of observing the latter; or it is the probability of observing the latter, given the former, times the probability of observing the former.

In our example,  $P(A \cap B) = P(A | B)P(B) = (2/3)(3/4) = 1/2$ .

Rewriting the product rule yields **Bayes' rule**:

$$P(B | A) = \frac{P(A | B)P(B)}{P(A)} \quad (4)$$

Equations 1 and 3 together imply  $P(A) = P(A | B)P(B) + P(A | B^c)P(B^c)$ , which we can use to reformulate Bayes' rule:

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

For example, consider the probability of Disease given Behavior:

$$P(\text{Disease} | \text{Behavior}) = \frac{P(\text{Behavior} | \text{Disease})P(\text{Disease})}{P(\text{Behavior})}$$

where  $P(\text{Behavior}) = P(\text{Behavior} | \text{Disease})P(\text{Disease}) + P(\text{Behavior} | \text{Disease}^c)P(\text{Disease}^c)$ .

More specifically, consider a medical clinic in which 10% of the patients have cancer, 25% of the patients are smokers, and 75% of cancer patients smoke. According to Bayes' rule, the likelihood that a patient who smokes has cancer is equal to  $(.75)(.1)/.25 = .3$ .

In general, let  $\{A_1, \dots, A_n\}$  be a disjoint set of events such that  $\bigcup_{i=1}^n A_i = \Omega$ . By the definition of conditional probability,

$$P(A_i | B) = \frac{P(B \cap A_i)}{P(B)}$$

By the product rule,  $P(B \cap A_i) = P(B | A_i)P(A_i)$ . Now since the  $A_i$ 's are disjoint,  $B = \bigcup_{j=1}^n (B \cap A_j)$ , and moreover,  $P(B) = \sum_{j=1}^n P(B \cap A_j)$ . Again, by the product rule,

$$P(B) = \sum_{j=1}^n P(B | A_j)P(A_j)$$

And now Bayes' rule in all its generality:

$$P(A_i | B) = \frac{P(B | A_i)P(A_i)}{\sum_{j=1}^n P(B | A_j)P(A_j)} \quad (5)$$

## 4.1 Independence

$A$  and  $B$  are **independent** events iff  $P(A | B) = P(A)$ . If  $P(A | B) = P(A)$ , then by Equation 4,  $P(B | A) = P(B)$ . Moreover,

$$P(A \cap B) = P(A | B)P(B) = P(A)P(B) \quad (6)$$

**Example:** If  $A$  is the event "head on the first toss" and  $B$  is the event "head on the second toss," then  $A = \{HH, HT\}$ ,  $B = \{HH, TH\}$ , and  $A \cap B = \{HH\}$ .  $A$  and  $B$  are independent events, since  $P(A \cap B) = 1/4 = (1/2)(1/2) = P(A)P(B)$ . On the other hand, if  $B$  is the event "at least one

tail,” then  $B = \{\text{HT,TH,TT}\}$ , and the events  $A$  and  $B$  are *not* independent:  $P(A \cap B) = 1/4$  but  $P(A)P(B) = (1/2)/(3/4) = 3/8$ .

**Exercise:** Show that if  $A$  and  $B$  are independent events, then the pairs of events  $A$  and  $B^c$ ,  $A^c$  and  $B$ , and  $A^c$  and  $B^c$  are all independent.

$A$  and  $B$  are **conditionally independent** with respect to  $C$  iff  $P(A | B \cap C) = P(A | C)$ . This definition also implies that  $P(B | A \cap C) = P(B | C)$ , since

$$\begin{aligned} P(B | A \cap C) &= \frac{P(A \cap C | B)P(B)}{P(A \cap C)} \\ &= \frac{P(A | B \cap C)P(C | B)P(B)}{P(A \cap C)} \\ &= \frac{P(A | C)P(C | B)P(B)}{P(A \cap C)} \\ &= \frac{P(C | B)P(B)}{P(C)} \\ &= P(B | C) \end{aligned}$$

Moreover, if  $A$  and  $B$  are conditionally independent of  $C$ , then

$$P(A \cap B | C) = P(A | B \cap C)P(B | C) = P(A | C)P(B | C) \quad (7)$$

**Exercise:** Show that if  $A$  and  $B$  are independent events, then  $A$  and  $B$  are conditionally independent of  $C$  for all events  $C$ .

## 5 Random Variables

A **simple probability space**  $\langle \Omega, 2^\Omega, P \rangle$  is a (finite) sample space  $\Omega$  together with a function  $P : 2^\Omega \rightarrow \mathbb{R}$  that assigns real values to events  $A \subseteq \Omega$  and satisfies the axioms of probability. Given a simple probability space  $\langle \Omega, 2^\Omega, P \rangle$ , a (discrete) **random variable**  $X$  is a map  $X : \Omega \rightarrow \{x_1, x_2, \dots\}$ . The real-valued function  $P(\{\omega | X(\omega) = x_i\})$ , abbreviated  $P(X = x_i)$ , gives rise to a **probability mass function**  $p_X : \{x_1, x_2, \dots\} \rightarrow \mathbb{R}$  given by  $p_X(x_i) \equiv P(X = x_i)$ .

**Example:** If  $\Omega = \{\text{HH,HT,TH,TT}\}$ , a possible random variable  $X$  is the total number of heads, in which case the range of  $X$  is  $\{0, 1, 2\}$ . Now  $p_X(0) = P(\{\text{TT}\}) = 1/4$ ,  $p_X(1) = P(\{\text{HT,TH}\}) = 2/4$ , and  $p_X(2) = P(\{\text{HH}\}) = 1/4$ . Similarly, if  $Y$  is the random variable representing the total number of tails, then  $p_Y(0) = P(\{\text{HH}\}) = 1/4$ ,  $p_Y(1) = P(\{\text{HT,TH}\}) = 2/4$ , and  $p_Y(2) = P(\{\text{TT}\}) = 1/4$ .

These probability functions can be fully specified using one-dimensional tables:

$X$	$p(X)$	$Y$	$p(Y)$
0	1/4	0	1/4
1	1/2	1	1/2
2	1/4	2	1/4

Let  $X$  and  $Y$  be random variables with ranges  $\{x_1, x_2, \dots\}$  and  $\{y_1, y_2, \dots\}$ , respectively. The random variable  $X \times Y$  has as its range the set of ordered pairs  $\{x_1, x_2, \dots\} \times \{y_1, y_2, \dots\}$ . The

**joint probability mass function**  $p_{X \times Y}(x_i, y_j) \equiv P(X = x_i, Y = y_j)$  is defined by  $P(\{\omega \mid X(\omega) = x_i, Y(\omega) = y_j\})$ . For example, the joint probability mass function on the random variables  $X$  and  $Y$  in the above example is specified in the following two-dimensional table:

$X, Y$	0	1	2
0	0	0	1/4
1	0	1/2	0
2	1/4	0	0

In general, it requires space exponential in the number of random variables to specify a joint probability mass function. However, when certain in/dependence criteria are satisfied, space requirements can be reduced.

Random variables  $X$  and  $Y$  with respective ranges  $\{x_1, x_2, \dots\}$  and  $\{y_1, y_2, \dots\}$  are **independent** iff the events  $X = x_i$  and  $Y = y_j$  are independent: i.e.,  $P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$ . Assuming independence, the joint probability mass function  $P(X = x_i, Y = y_j)$  can be specified using two one-dimensional tables, rather than one two-dimensional table. In our example, however, the variables  $X$  and  $Y$  are *not* independent;  $X$  and  $Y$  are perfectly correlated. Indeed, one one-dimensional table suffices to describe both variables.

The definitions conditional probability and conditional independence for random variables, like the definition of independence for random variables, translate directly from the respective definitions for events. The corresponding notions of the product rule and Bayes' rule are given by:

$$\begin{aligned} P(X = x_i, Y = y_j) &= P(X = x_i \mid Y = y_j)P(Y = y_j) \\ &= P(Y = y_j \mid X = x_i)P(X = x_i) \end{aligned}$$

and

$$P(X = x_i \mid Y = y_j) = \frac{P(Y = y_j \mid X = x_i)P(X = x_i)}{P(Y = y_j)}$$

## 5.1 Expectation and Variance

The **expected value**  $\mathbb{E}[X]$  of a random variable  $X$  is defined as:

$$\mathbb{E}[X] = \sum_i p(x_i)x_i$$

The expected value is also called the mean, in which case it is denoted  $\mu_X$ , or  $\mu$ , whenever the random variable over which it is defined is clear from context.

For example, if  $I$  is a random variable with range  $\{0, 1, \dots, n\}$ , and if  $I$  is uniformly distributed (i.e.,  $P(I = i) = 1/(n + 1)$  for all  $0 \leq i \leq n$ ), then

$$\mathbb{E}[I] = \sum_{i=0}^n i \frac{1}{n+1} = \frac{1}{n+1} \sum_{i=0}^n i = \frac{1}{n+1} \frac{n(n+1)}{2} = \frac{n}{2}$$

**Exercise:** Compute  $\mathbb{E}[X]$  and  $\mathbb{E}[Y]$  in our running example.

The following properties hold of expectation:

- If  $X$  and  $Y$  are independent random variables, then  $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ .
- Linearity of expectation: *i.e.*,  $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ .
- For arbitrary constant  $c \in \mathbb{R}$ ,  $\mathbb{E}[cX] = c\mathbb{E}[X]$ .

**Exercise:** Prove the stated properties of expectation.

Expectation suffices to predict an average data point. The expected value of the two sequences  $0, 0, \dots$  and  $-1, 1, -1, 1, \dots$  both equal zero, however. Variance captures the variability in a series of data points.

Given random variable  $X$ , the **variance**  $\text{var}(X) = \mathbb{E}[(X - \mu_X)^2]$ :

$$\begin{aligned} \text{var}(X) &= \mathbb{E}[(X - \mu_X)^2] \\ &= \mathbb{E}[X^2 - 2\mu_X X + \mu_X^2] \\ &= \mathbb{E}[X^2] - 2\mu_X \mathbb{E}[X] + \mu_X^2 \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \end{aligned}$$

Variance is often denoted  $\sigma_X^2$ , or simply  $\sigma^2$ , whenever  $X$  is clear from context. The value  $\sigma$  is called the standard deviation.

If  $I$  is as above, a random variable with range  $\{0, 1, \dots, n\}$  and  $p(i)$  is uniformly distributed, then

$$E[I^2] = \sum_{i=0}^n i^2 \frac{1}{n+1} = \frac{1}{n+1} (0^2 + 1^2 + 2^2 + \dots + n^2) = \frac{n(2n+1)}{6}$$

Thus,

$$\text{var}(I) = \mathbb{E}[I^2] - (E[I])^2 = \frac{n(2n+1)}{6} - \left(\frac{n}{2}\right)^2 = \frac{n(n+2)}{12}$$

The following properties hold of variance:

- For arbitrary constant  $a \in \mathbb{R}$ ,  $\text{var}(aX) = a^2 \text{var}(X)$ .
- For arbitrary constant  $c \in \mathbb{R}$ ,  $\text{var}(X + c) = \text{var}(X)$ .

**Exercise:** Prove the stated properties of variance.

**Exercise:** Let  $Z$  represent the number of heads after three successive coin tosses. Given  $p(Z)$  as follows, compute  $\mathbb{E}[Z]$  and  $\text{var}(Z)$ .

$Z$	$p(Z)$
0	1/8
1	3/8
2	3/8
3	1/8

## 5.2 Correlation and Covariance

Correlation and covariance provide numerical measurements of the strength of the relationship between two random variables.

Given random variables  $X$  and  $Y$  with respective means  $\mu_X$  and  $\mu_Y$ , the covariance  $\text{cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$ :

$$\begin{aligned}
 \text{cov}(X, Y) &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\
 &= \mathbb{E}[(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y)] \\
 &= \mathbb{E}[XY] - \mu_X \mathbb{E}[Y] - \mu_Y \mathbb{E}[X] + \mu_X \mu_Y \\
 &= \mathbb{E}[XY] - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y \\
 &= \mathbb{E}[XY] - \mu_X \mu_Y
 \end{aligned}$$

The correlation coefficient  $\rho_{XY}$  is defined in terms of covariance and standard deviation as follows:

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \quad (8)$$

The following properties hold of covariance and correlation:

- For independent random variables  $X$  and  $Y$ ,  $\text{cov}(X, Y) = \rho_{XY} = 0$ .
- For arbitrary random variables  $X$  and  $Y$ ,  $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2 \text{cov}(X, Y)$ .