Introduction to Probability

COS 341 Fall 2004

Basic Laws of Probability

Definition 1 *A* sample space S *is a nonempty set whose elements are called* outcomes*. The* events *are subsets of* S*.*

Since events are subsets, we can apply the usual set operations to events to obtain new events. For events A and B, the event $A \cap B$ represents the set of outcomes that are in both event A and event B, i.e. $A \cap B$ represents the event A *and* B. Similarly, $A \cup B$ represents the event A *or* B.

Definition 2 *A* probability space *consists of a sample space* S *and a probability function* $\mathbf{Pr}()$*, mapping the events of* S *to real numbers in* $[0,1]$*, such that:*

1. $\mathbf{Pr}(S) = 1$*, and*
2. *If* $A_0, A_1, \ldots$ *is a sequence of disjoint events, then*

$$\mathbf{Pr}\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} \mathbf{Pr}(A_i). \qquad \textit{(Sum Rule)}$$

One consequence of this definition is the following:

$$\mathbf{Pr}(\overline{A}) = 1 - \mathbf{Pr}(\overline{A}). \qquad \text{(Complement Rule)}$$

Several basic rules of probability parallel facts about cardinalities of finite sets:

$$\begin{aligned} \mathbf{Pr}(B - A) &= \mathbf{Pr}(B) - \mathbf{Pr}(A \cap B) && \text{(Difference Rule)} \\ \mathbf{Pr}(A \cup B) &= \mathbf{Pr}(A) + \mathbf{Pr}(B) - \mathbf{Pr}(A \cap B) && \text{(Inclusion-Exclusion)} \end{aligned}$$

An immediate consequence of (Inclusion-Exclusion) is

$$\mathbf{Pr}(A \cup B) \leq \mathbf{Pr}(A) + \mathbf{Pr}(B) \qquad \text{(Boole's Inequality)}$$

Similarly (Difference Rule) impies that

$$\text{If } A \subseteq B\text{, then } \mathbf{Pr}(A) \leq \mathbf{Pr}(B). \qquad \text{(Monotonicity)}$$

Example 1 *Suppose we wire up a circuit containing a total of* n *connections. The probability of getting any one connection wrong is* p*. What can we say about the probability of wiring the circuit correctly ? (The circuit is wired correctly iff all the* n *connections are made correctly.)*

solution: Let A_i denote the event that connection i is made *correctly.* So $\mathbf{Pr}(\overline{A_i}) = p$.

$$\mathbf{Pr}(\text{all connections correct}) = \mathbf{Pr}\left(\cap_{i=1}^{n} A_i\right).$$

Without any additional assumptions (on the dependence of the events A_i), we cannot get an exact answer. However, we can give reasonable upper and lower bounds.

$$\mathbf{Pr}\left(\cap_{i=1}^{n} A_i\right) \le \mathbf{Pr}(A_i) = 1 - p$$

$$\mathbf{Pr}\left(\cap_{i=1}^{n} A_i\right) = 1 - \mathbf{Pr}\left(\overline{\cap_{i=1}^{n} A_i}\right) = 1 - \mathbf{Pr}\left(\cup_{i=1}^{n} \overline{A_i}\right) \ge 1 - \sum_{i=1}^{n} \mathbf{Pr}(\overline{A_i}) = 1 - np$$

Both these bounds are tight, i.e. we can construct situations where the correct answer is equal to the upper bound and those where the correct answer is equal to the lower bound. ∎

Conditional Probability

Definition 3 $\mathbf{Pr}(A|B)$ *denotes the probability of event A given that event B has occured.*

$$\mathbf{Pr}(A|B) ::= \frac{\mathbf{Pr}(A \cap B)}{\mathbf{Pr}(B)}$$

provided $\mathbf{Pr}(B) \ne 0$.

Rearranging terms gives the following:

Rule 1 (Product rule, base case) *Let A and B be events, with $\mathbf{Pr}(B) \ne 0$. Then*

$$\mathbf{Pr}(A \cap B) = \mathbf{Pr}(B) \cdot \mathbf{Pr}(A|B).$$

Rule 2 (Product rule, general case) *Let $A_1, A_2, \ldots, A_n$ be events.*

$$\mathbf{Pr}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbf{Pr}(A_1) \cdot \mathbf{Pr}(A_2|A_1) \cdot \mathbf{Pr}(A_3|A_1 \cap A_2) \cdot \ldots \cdot \mathbf{Pr}(A_n|A_1 \cap \cdots \cap A_{n-1}).$$

Case Analysis

Theorem 1 (Total Probability) *If a sample space is the disjoint union of events $B_1, B_2, \ldots$, then for all events A,*

$$\mathbf{Pr}(A) = \sum_{i \in \mathbb{N}} \mathbf{Pr}(A \cap B_i).$$

Corollary 1 (Total Probability) *If a sample space is the disjoint union of events $B_1, B_2, \ldots$, then for all events A,*

$$\mathbf{Pr}(A) = \sum_{i \in \mathbb{N}} \mathbf{Pr}(A|B_i)\,\mathbf{Pr}(B_i).$$

Independence

Definition 4 *Suppose A and B are events, and B has positive probability. Then A is* independent *of B iff*

$$\mathbf{Pr}(A|B) = \mathbf{Pr}(A)$$

The above definition does not apply when $\mathbf{Pr}(B) = 0$. We will extend the definition to the zero probability case as follows:

Definition 5 *If A and B are events and $\mathbf{Pr}(B) = 0$, then A is defined to be independent of B.*

Now we can define independence in an alternate way:

Theorem 2 *Events A and B are independent iff*

$$\mathbf{Pr}(A \cap B) = \mathbf{Pr}(A) \cdot \mathbf{Pr}(B). \qquad \textit{(Independence Product Rule)}$$

Note that *disjoint* events are not the same as *independent* events. In general disjoint events are not independent.

Random Variables

Informally, a random variable is the value of a measurement associated with an experiment, e.g. the number of heads in n tosses of a coin. More formally, a random variable is defined as follows:

Definition 6 *A* random variable *over a sample space is a function that maps every sample point (i.e. outcome) to a real number.*

An *indicator random variable* is a special kind of random variable associated with the occurence of an event. The indicator random variable I_A associated with event A has value 1 if event A occurs and has value 0 otherwise. In other words, I_A maps all outcomes in the set A to 1 and all outcomes outside A to 0.

Random variables can be used to define events. In particular, any predicate involving random variables defines the event consisting of all outcomes for which the predicate is true. e.g. for random variables R_1, R_2, $R_1 = 1$ is an event, $R_2 \leq 2$ is an event, $R_1 = 1 \wedge R_2 \leq 2$ is an event.

Events derived from random variables can be used in expressions involving conditional probability as well. e.g.

$$\mathbf{Pr}(R_1 = 1 | R_2 \leq 2) = \frac{\mathbf{Pr}(R_1 = 1 \wedge R_2 \leq 2)}{\mathbf{Pr}(R_2 \leq 2)}$$

Independence of Random Variables

Definition 7 *Two random variables R_1 and R_2 are* independent*, if for all $x_1, x_2 \in \mathbb{R}$, we have:*

$$\mathbf{Pr}(R_1 = x_1 \wedge R_2 = x_2) = \mathbf{Pr}(R_1 = x_1) \cdot \mathbf{Pr}(R_2 = x_2)$$

An alternate definition is as follows:

Definition 8 *Two random variables R_1 and R_2 are* independent*, if for all $x_1, x_2 \in \mathbb{R}$, such that $\mathbf{Pr}(R_2 = x_2) \neq 0$, we have:*

$$\mathbf{Pr}(R_1 = x_1 | R_2 = x_2) = \mathbf{Pr}(R_1 = x_1)$$

In order to prove that two random variables are *not* independent, we need to exhibit a pair of values x_1, x_2 for which the condition in the definition is violated. On the other hand, proving independence requires an argument that the condition in the definition holds for all pairs of values x_1, x_2.

Mutual Independence

Definition 9 *Random variables $R_1, R_2, \ldots, R_t$ are mutually independent if, for all $x_1, x_2, \ldots, x_t \in \mathbb{R}$,*

$$\mathbf{Pr}\left(\bigcap_{i=1}^{t} R_i = x_i\right) = \prod_{i=1}^{t} \mathbf{Pr}(R_i = x_i).$$

Definition 10 *A collection of random variables is said to be k-wise independent if all subsets of k variables are mutually independent.*

Consider a sample space consisting of bit sequences of length 2, where all 4 possible two bit sequences are equally likely. Random variable B_1 is the value of the first bit, B_2 is the value of the second bit and B_3 is $B_1 \oplus B_2$. Here the variables B_1, B_2, B_3 are 2-wise independent, but they are not mutually independent.

Pairwise independence is another name for 2-wise independence, i.e. when we say that a collection of variables is pairwise independent, we mean that they are 2-wise independent.

Probability Density Functions

Probability density functions are used to describe the distribution of a random variable, i.e. the set of values a random variable takes and the probabilities associated with those values. This description of a random variable is independent of any experiment.

Definition 11 *The* probability density function *(pdf) for a random variable X is the function $f_X : (R) \to [0, 1]$ defined by:*

$$f_X(t) = \mathbf{Pr}(X = t).$$

For a value t not in the range of X, $f_X(t) = 0$. Note that $\sum_{t \in \mathbb{R}} f_X(t) = 1$,

Definition 12 *The* cumulative distribution function *(cdf) for a random variable X is the function $F_X : \mathbb{R} \to [0,1]$ defined by:*

$$F_X(t) = \mathbf{Pr}(X \leq t) = \sum_{s \leq t} f_X(s).$$

Two common distributions enountered are the *uniform distribution* and the *binomial distribution.*

Uniform Distribution

Let U be a random variable that takes values in the range $\{1, \ldots, N\}$, such that each value is equally likely. Such a variable is said to be uniformly distributed. The pdf and cdf for this distribution are:

$$f_U(t) = \frac{1}{N}, \qquad\qquad F_U(t) = \frac{t}{N}, \text{ for} 1 \leq t \leq N.$$

Binomial Distribution

Let H be the number of heads in n inpdendent tosses of a biased coin. Each toss of the coin has probability p of being heads and probability $1-p$ of being tails. Such a variable is said to have a *binomial* distribution. The pdf of this distribution is given by

$$f_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

As a sanity check, we can verify that

$$\sum_{k=0}^{n} f_{n,p}(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1$$

Expected Value

Definition 13 *The* expectation $\mathbf{E}[X]$ *of a random variable X on a sample space S is defined as:*

$$\mathbf{E}[X] = \sum_{s \in S} X(s) \cdot \mathbf{Pr}(\{s\}).$$

An equivalent definition is:

Definition 14 *The* expectation *of a random variable X is*

$$\mathbf{E}[X] = \sum_{t \in range(X)} t \cdot \mathbf{Pr}(X = t).$$

If the range of a random variable is non-negative integers, there is an another way to compute the expectation.

Theorem 3 *If X is a random variable which takes values in the non-negative integers, then*

$$\mathbf{E}[X] = \sum_{t=0}^{\infty} \mathbf{Pr}(X > t).$$

Proof: Note that

$$\mathbf{Pr}(X > t) = \mathbf{Pr}(X = t+1) + \mathbf{Pr}(X = t+2) + \mathbf{Pr}(X = t+3) + \cdots$$

Thus,

$$\begin{aligned}
\sum_{t=0}^{\infty} \mathbf{Pr}(X > t) &= \mathbf{Pr}(X > 0) + \mathbf{Pr}(X > 1) + \mathbf{Pr}(X > 2) + \cdots \\
&= \mathbf{Pr}(X = 1) + \mathbf{Pr}(X = 2) + \mathbf{Pr}(X = 3) + \cdots \\
& \mathbf{Pr}(X = 2) + \mathbf{Pr}(X = 3) + \mathbf{Pr}(X = 4) + \cdots \\
& \mathbf{Pr}(X = 3) + \mathbf{Pr}(X = 4) + \mathbf{Pr}(X = 5) + \cdots \\
&= 1 \cdot \mathbf{Pr}(X = 1) + 2 \cdot \mathbf{Pr}(X = 2) + 3 \cdot \mathbf{Pr}(X = 3) + \cdots \\
&= \sum_{t=0}^{\infty} t \cdot \mathbf{Pr}(X = t) \\
&= \mathbf{E}[X].
\end{aligned}$$

■

Linearity of Expectation

Theorem 4 (Linearity of Expectation) *For any random variables X_1 and X_2, and constants $c_1, c_2 \in \mathbb{R}$,*

$$\mathbf{E}[c_1 X_1 + c_2 X_2] = c_1\, \mathbf{E}[X_1] + c_2\, \mathbf{E}[X_2]$$

Note that the above theorem holds irrespective of the dependence between X_1 and X_2.

Corollary 2 *For any random variables $X_1, \ldots, X_k$, and constants $c_1, \ldots, c_k \in \mathbb{R}$,*

$$\mathbf{E}\left[\sum_{i=1}^{k} c_i X_i\right] = \sum_{i=1}^{k} c_i\, \mathbf{E}[X_i].$$

Conditional Expectation

Definition 15 *We define* conditional expectation, $\mathbf{E}[X|A]$, *of a random variable, given event A, to be*

$$\mathbf{E}[X|A] = \sum_k k \cdot \mathbf{Pr}(X = k|A).$$

The rules for expectation also apply to conditional expectation:

Theorem 5

$$\mathbf{E}[c_1 X_1 + c_2 X_2 | A] = c_1\, \mathbf{E}[X_1|A] + c_2\, \mathbf{E}[X_2|A].$$

The following theorem shows how conditional expectation allows us to compute the expectation by case analysis.

Theorem 6 (Law of Total Expectation) *If the sample space is the disjoint union of events $A_1, A_2, \ldots$, then*

$$\mathbf{E}[X] = \sum_i \mathbf{E}[X|A_i]\, \mathbf{Pr}(A_i).$$

Expected value of a product

In general, the expected value of the product of two random variables need not be equal to the product of their expectations. However, this holds when the random variables are independent:

Theorem 7 *For any two* independent *random variables,* X_1 *and* X_2*,*

$$\mathbf{E}[X_1 \cdot X_2] = \mathbf{E}[X_1] \cdot \mathbf{E}[X_2].$$

Corollary 3 *If random variables* $X_1, X_2, \ldots, X_k$ *are mutually independent, then*

$$\mathbf{E}\left[\prod_{i=1}^{k} X_i\right] = \prod_{i=1}^{k} \mathbf{E}[X_i].$$

Note that in general,

$$\mathbf{E}\left[\frac{1}{T}\right] \neq \frac{1}{\mathbf{E}[T]}.$$

Linearity of expectation also holds for infinite sums, provided the summations considered are absolutely convergent:

Theorem 8 (Infinite Linearity of Expectation) *Let* $X_1, X_2, \ldots$ *be random variables such that* $\sum_{i=1}^{\infty} \mathbf{E}[|X_i|]$ *converges. Then*

$$\mathbf{E}\left[\sum_{i=1}^{\infty} X_i\right] = \sum_{i-0}^{\infty} \mathbf{E}[X_i].$$