Basic Probability Theory

Adrian Brasoveanu

[based primarily on slides by Sharon Goldwater & Frank Keller]

Winter 2012 · Seminar in Semantics · UCSC Linguistics

1. Sample Spaces and Events
    - Sample Spaces
    - Events
    - Axioms and Rules of Probability

2. Conditional Probability and Bayes' Theorem
    - Conditional Probability
    - Total Probability
    - Bayes' Theorem

3. Random Variables and Distributions
    - Random Variables
    - Distributions
    - Expectation

Terminology

Terminology for probability theory:

- *experiment:* process of observation or measurement; e.g., coin flip;
- *outcome:* result obtained through an experiment; e.g., coin shows tails;
- *sample space:* set of all possible outcomes of an experiment; e.g., sample space for coin flip: $S = \{H, T\}$.

Sample spaces can be finite or infinite.

Example: Finite Sample Space

Roll two dice, each with numbers 1–6. Sample space:

$$S_1 = \{(x, y) : x \in \{1, 2, \ldots, 6\} \wedge \{y = 1, 2, \ldots, 6\}\}$$

Alternative sample space for this experiment – sum of the dice:

$$S_2 = \{x + y : x \in \{1, 2, \ldots, 6\} \wedge \{y = 1, 2, \ldots, 6\}\}$$

$$S_2 = \{z : z = 2, 3, \ldots, 12\}$$

Example: Infinite Sample Space

Flip a coin until heads appears for the first time:

$$S_3 = \{H, TH, TTH, TTTH, TTTTH, \ldots\}$$

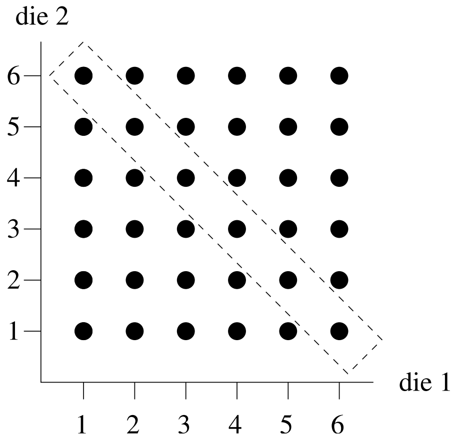
Often we are not interested in individual outcomes, but in events. An *event* is a subset of a sample space.

Example

With respect to S_1, describe the event B of rolling a total of 7 with the two dice.

$$B = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$$

The event B can be represented graphically:

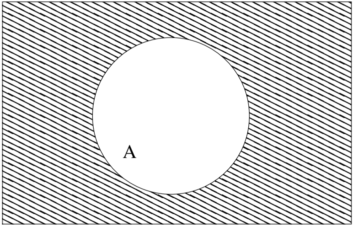


Events

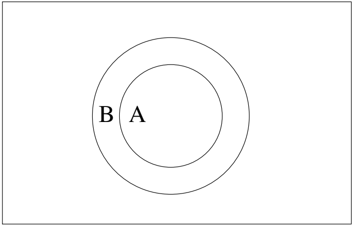
Often we are interested in combinations of two or more events. This can be represented using set theoretic operations. Assume a sample space S and two events A and B:

- *complement $\bar{A}$ (also A'):* all elements of S that are not in A;
- *subset $A \subseteq B$:* all elements of A are also elements of B;
- *union $A \cup B$:* all elements of S that are in A or B;
- *intersection $A \cap B$:* all elements of S that are in A and B.

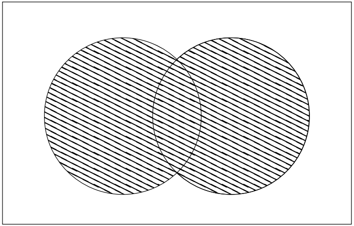
These operations can be represented graphically using *Venn diagrams.*



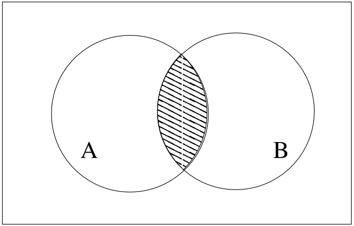
$\bar{A}$



$A \subseteq B$



$A \cup B$



$A \cap B$

Events are denoted by capital letters A, B, C, etc. The *probability* of and event A is denoted by $P(A)$.

Axioms of Probability

1. The probability of an event is a nonnegative real number: $P(A) \geq 0$ for any $A \subseteq S$.
2. $P(S) = 1$.
3. If $A_1, A_2, A_3, \ldots,$ is a sequence of mutually exclusive events of S, then:

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots$$

Theorem: Probability of an Event

If A is an event in a sample space S and $O_1, O_2, \ldots, O_n$, are the individual outcomes comprising A, then $P(A) = \sum_{i=1}^{n} P(O_i)$

Example

Assume all strings of three lowercase letters are equally probable. Then what's the probability of a string of three vowels?

There are 26 letters, of which 5 are vowels. So there are $N = 26^3$ three letter strings, and $n = 5^3$ consisting only of vowels. Each outcome (string) is equally likely, with probability $\frac{1}{N}$, so event A (a string of three vowels) has probability $P(A) = \frac{n}{N} = \frac{5^3}{26^3} = 0.00711$.

Theorems: Rules of Probability

1. If A and $\bar{A}$ are complementary events in the sample space S, then $P(\bar{A}) = 1 - P(A)$.
2. $P(\emptyset) = 0$ for any sample space S.
3. If A and B are events in a sample space S and $A \subseteq B$, then $P(A) \leq P(B)$.
4. $0 \leq P(A) \leq 1$ for any event A.

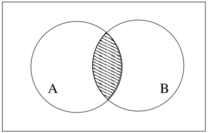
Axiom 3 allows us to add the probabilities of mutually exclusive events. What about events that are not mutually exclusive?

Theorem: General Addition Rule

If A and B are two events in a sample space S, then:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Ex: $A =$ "has glasses", $B =$ "is blond". $P(A) + P(B)$ counts blondes with glasses twice, need to subtract once.



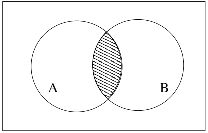
Definition: Conditional Probability, Joint Probability

If A and B are two events in a sample space S, and $P(A) \neq 0$ then the *conditional probability* of B given A is:

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

$P(A \cap B)$ is the *joint probability* of A and B, also written $P(A, B)$.

Intuitively, $P(B|A)$ is the probability that B will occur given that A has occurred.
Ex: The probability of being blond given that one wears glasses: $P(\text{blond}|\text{glasses})$.



Example

A manufacturer knows that the probability f an order being ready on time is 0.80, and the probability of an order being ready on time and being delivered on time is 0.72.

What is the probability of an order being delivered on time, given that it is ready on time?

R: order is ready on time; D: order is delivered on time.
$P(R) = 0.80$, $P(R, D) = 0.72$. Therefore:

$$P(D|R) = \frac{P(R, D)}{P(R)} = \frac{0.72}{0.80} = 0.90$$

Example

Consider sampling an adjacent pair of words (bigram) from a large text T. Let $\mathcal{BI}$ = the set of bigrams in T (this is our sample space), A = "first word is *run*" $= \{(run, w_2) : w_2 \in T\} \subseteq \mathcal{BI}$ and B = "second word is *amok*" $= \{(w_1, amok) : w_1 \in T\} \subseteq \mathcal{BI}$.

If $P(A) = 10^{-3.5}$, $P(B) = 10^{-5.6}$, and $P(A, B) = 10^{-6.5}$, what is the probability of seeing *amok* following *run*? *Run* preceding *amok*?

$$P(\text{"}run \text{ before } amok\text{"}) = P(A|B) = \frac{P(A, B)}{P(B)} = \frac{10^{-6.5}}{10^{-5.6}} = .126$$

$$P(\text{"}amok \text{ after } run\text{"}) = P(B|A) = \frac{P(A, B)}{P(A)} = \frac{10^{-6.5}}{10^{-3.5}} = .001$$

[How do we determine $P(A)$, $P(B)$, $P(A, B)$ in the first place?]

From the definition of conditional probability, we obtain:

Theorem: Multiplication Rule

If A and B are two events in a sample space S and $P(A) \neq 0$, then:

$$P(A, B) = P(A)P(B|A)$$

Since $A \cap B = B \cap A$, we also have that:

$$P(A, B) = P(B)P(A|B)$$

Definition: Independent Events

Two events A and B are independent iff:

$$P(A, B) = P(A)P(B)$$

Intuition: two events are independent if knowing whether one event occurred does not change the probability of the other.

Note that the following are equivalent:

$$\begin{aligned} P(A, B) &= P(A)P(B) \quad (1) \\ P(A|B) &= P(A) \quad (2) \\ P(B|A) &= P(B) \quad (3) \end{aligned}$$

Example

A coin is flipped three times. Each of the eight outcomes is equally likely. A: heads occurs on each of the first two flips, B: tails occurs on the third flip, C: exactly two tails occur in the three flips. Show that A and B are independent, B and C dependent.

$$
\begin{array}{ll}
A = \{HHH, HHT\} & P(A) = \frac{1}{4} \\
B = \{HHT, HTT, THT, TTT\} & P(A) = \frac{1}{2} \\
C = \{HTT, THT, TTH\} & P(C) = \frac{3}{8} \\
A \cap B = \{HHT\} & P(A \cap B) = \frac{1}{8} \\
B \cap C = \{HTT, THT\} & P(B \cap C) = \frac{1}{4}
\end{array}
$$

$P(A)P(B) = \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8} = P(A \cap B)$, hence A and B are independent.
$P(B)P(C) = \frac{1}{2} \cdot \frac{3}{8} = \frac{3}{16} \neq P(B \cap C)$, hence B and C are dependent.

Definition: Conditionally Independent Events

Two events A and B are conditionally independent given event C iff:

$$P(A, B|C) = P(A|C)P(B|C)$$

Intuition: Once we know whether C occurred, knowing about A or B doesn't change the probability of the other.

Show that the following are equivalent:

$$\begin{aligned} P(A, B|C) &= P(A|C)P(B|C) \quad (4) \\ P(A|B, C) &= P(A|C) \quad (5) \\ P(B|A, C) &= P(B|C) \quad (6) \end{aligned}$$

Example

In a noisy room, I whisper the same number $n \in \{1, \ldots, 10\}$ to two people A and B on two separate occasions. A and B imperfectly (and independently) draw a conclusion about what number I whispered. Let the numbers A and B think they heard be n_a and n_b, respectively.

Are n_a and n_b independent (a.k.a. marginally independent)? No. E.g., we'd expect $P(n_a = 1 | n_b = 1) > P(n_a = 1)$.

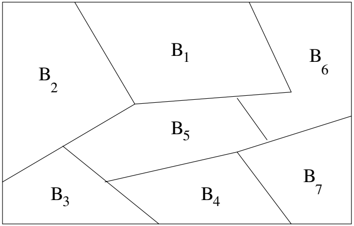
Are n_a and n_b conditionally independent given n? Yes: if you know the number that I actually whispered, the two variables are no longer correlated.
E.g., $P(n_a = 1 | n_b = 1, n = 2) = P(n_a = 1 | n = 2)$

Theorem: Rule of Total Probability

If events $B_1, B_2, \ldots, B_k$ constitute a partition of the sample space S and $P(B_i) \neq 0$ for $i = 1, 2, \ldots, k$, then for any event A in S:

$$P(A) = \sum_{i=1}^{k} P(B_i)P(A|B_i)$$

$B_1, B_2, \ldots, B_k$ form a *partition* of S if they are pairwise mutually exclusive and if $B_1 \cup B_2 \cup \ldots \cup B_k = S$.



Example

In an experiment on human memory, participants have to memorize a set of words (B_1), numbers (B_2), and pictures (B_3). These occur in the experiment with the probabilities $P(B_1) = 0.5$, $P(B_2) = 0.4$, $P(B_3) = 0.1$.

Then participants have to recall the items (where A is the recall event). The results show that $P(A|B_1) = 0.4$, $P(A|B_2) = 0.2$, $P(A|B_3) = 0.1$. Compute $P(A)$, the probability of recalling an item.

By the theorem of total probability:

$$\begin{aligned} P(A) &= \textstyle\sum_{i=1}^{k} P(B_i)P(A|B_i) \\ &= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3) \\ &= 0.5 \cdot 0.4 + 0.4 \cdot 0.2 + 0.1 \cdot 0.1 = 0.29 \end{aligned}$$

Bayes' Theorem

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

(Derived using mult. rule: $P(A,B) = P(A|B)P(B) = P(B|A)P(A)$)

- Denominator can be computed using theorem of total probability: $P(A) = \sum_{i=1}^{k} P(B_i)P(A|B_i)$.
- Denominator is a normalizing constant (ensures $P(B|A)$ sums to one). If we only care about relative sizes of probabilities, we can ignore it: $P(B|A) \propto P(A|B)P(B)$.

Example

Consider the memory example again. What is the probability that an item that is correctly recalled (A) is a picture (B_3)?

By Bayes' theorem:

$$\begin{aligned} P(B_3|A) &= \frac{P(B_3)P(A|B_3)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)} \\ &= \frac{0.1 \cdot 0.1}{0.29} = 0.0345 \end{aligned}$$

The process of computing $P(B|A)$ from $P(A|B)$ is sometimes called *Bayesian inversion.*

Definition: Random Variable

If S is a sample space with a probability measure and X is a real-valued function defined over the elements of S, then X is called a random variable.

We symbolize random variables (r.v.s) by capital letters (e.g., X), and their values by lower-case letters (e.g., x).

Example

Given an experiment in which we roll a pair of 4-sided dice, let the random variable X be the total number of points rolled with the two dice.

E.g. $X = 5$ 'picks out' the set $\{(1,4),(2,3),(3,2),(4,1)\}$.

Specify the full function denoted by X and determine the probabilities associated with each value of X.

Example

Assume a balanced coin is flipped three times. Let X be the random variable denoting the total number of heads obtained.

Outcome	Probability	x
HHH	$\frac{1}{8}$	3
HHT	$\frac{1}{8}$	2
HTH	$\frac{1}{8}$	2
THH	$\frac{1}{8}$	2

Outcome	Probability	x
TTH	$\frac{1}{8}$	1
THT	$\frac{1}{8}$	1
HTT	$\frac{1}{8}$	1
TTT	$\frac{1}{8}$	0

Hence, $P(X = 0) = \frac{1}{8}$, $P(X = 1) = P(X = 2) = \frac{3}{8}$, $P(X = 3) = \frac{1}{8}$.

Definition: Probability Distribution

If X is a random variable, the function $f(x)$ whose value is $P(X = x)$ for each value x in the range of X is called the probability distribution of X.

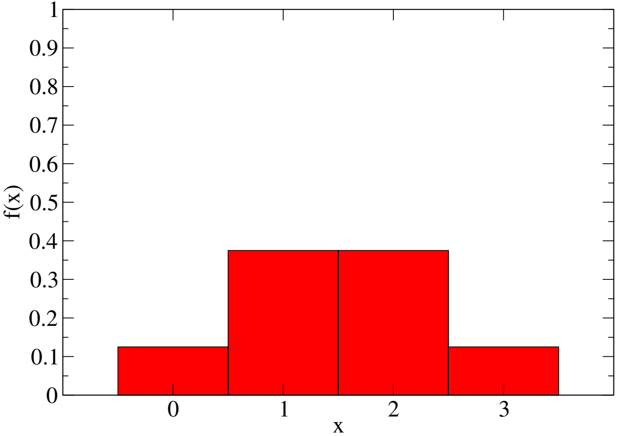
Note: the set of values x ('the support') = the domain of f = the range of X.

Example

For the probability function defined in the previous example:

x	$f(x)$
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$

A probability distribution is often represented as a *probability histogram.* For the previous example:



Any probability distribution function (or simply: probability distribution) f of a random variable X is such that:

1. $f(x) \geq 0, \forall x \in \textbf{Domain}(f)$
2. $\sum_{x \in \textbf{Domain}(f)} f(x) = 1.$

Example: geometric distribution

Let X be the number of coin flips needed before getting heads, where p_h is the probability of heads on a single flip. What is the distribution of X?

Assume flips are independent, so:

$$P(T^{n-1}H) = P(T)^{n-1}P(H)$$

Therefore:

$$P(X = n) = (1 - p_h)^{n-1}p_h$$

The notion of mathematical expectation derives from games of chance. It's the product of the amount a player can win and the probability of wining.

Example

In a raffle, there are 10,000 tickets. The probability of winning is therefore $\frac{1}{10{,}000}$ for each ticket. The prize is worth \$4,800. Hence the expectation per ticket is $\frac{\$4{,}800}{10{,}000} = \$0.48$.

In this example, the expectation can be thought of as the average win per ticket.

This intuition can be formalized as the *expected value* (or *mean*) of a random variable:

Definition: Expected Value

If X is a random variable and $f(x)$ is the value of its probability distribution at x, then the expected value of X is:

$$E(X) = \sum_x x \cdot f(x)$$

Example

A balanced coin is flipped three times. Let X be the number of heads. Then the probability distribution of X is:

$$f(x) = \begin{cases} \frac{1}{8} & \text{for } x = 0 \\ \frac{3}{8} & \text{for } x = 1 \\ \frac{3}{8} & \text{for } x = 2 \\ \frac{1}{8} & \text{for } x = 3 \end{cases}$$

The expected value of X is:

$$E(X) = \sum_x x \cdot f(x) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{3}{2}$$

The notion of expectation can be generalized to cases in which a function $g(X)$ is applied to a random variable X.

Theorem: Expected Value of a Function

If X is a random variable and $f(x)$ is the value of its probability distribution at x, then the expected value of $g(X)$ is:

$$E[g(X)] = \sum_{x} g(x)f(x)$$

Example

Let X be the number of points rolled with a balanced (6-sided) die. Find the expected value of X and of $g(X) = 2X^2 + 1$.

The probability distribution for X is $f(x) = \frac{1}{6}$. Therefore:

$$E(X) = \sum_x x \cdot f(x) = \sum_{x=1}^{6} x \cdot \frac{1}{6} = \frac{21}{6}$$

$$E[g(X)] = \sum_x g(x)f(x) = \sum_{x=1}^{6} (2x^2 + 1)\frac{1}{6} = \frac{94}{6}$$

Summary

- Sample space S contains all possible outcomes of an experiment; events A and B are subsets of S.
- rules of probability: $P(\bar{A}) = 1 - P(A)$.
  if $A \subseteq B$, then $P(A) \leq P(B)$.
  $0 \leq P(B) \leq 1$.
- addition rule: $P(A \cup B) = P(A) + P(B) - P(A, B)$.
- conditional probability: $P(B|A) = \frac{P(A,B)}{P(A)}$.
- independence: $P(A, B) = P(A)P(B)$.
- total probability: $P(A) = \sum_{B_i} P(B_i)P(A|B_i)$.
- Bayes' theorem: $P(B|A) = \frac{P(B)P(A|B)}{P(A)}$.
- any value of an r.v. 'picks out' a subset of the sample space.
- for any value of an r.v., a distribution returns a probability.
- the expectation of an r.v. is its average value over a distribution.