Introduction to Probability Theory

L645

Dept. of Linguistics, Indiana University
Fall 2015
To start out the course, we need to know something about statistics and probability

- This is only an introduction; for a fuller understanding, you would need to take a statistics course

**Probability theory** = theory to determine how likely it is that some outcome will occur
Probability spaces

We state things in terms of an **experiment** (or trial)—e.g., flipping three coins

- **outcome**: one particular possible result
  - e.g., coin 1 = heads, coin 2 = tails, coin 3 = tails (HTT)
- **event**: one particular possible set of results, i.e., a more abstract idea
  - e.g., two tails and one head (\{HTT, THT, TTH\})

The set of basic outcomes makes up the **sample space** (Ω)

- Discrete sample space: countably infinite outcomes (1, 2, 3, ...), e.g., heads or tails
- Continuous sample space: uncountably infinite outcomes (1.1293..., 8.765..., ...), e.g., height

We will use \( \mathcal{F} \) to refer to the set of events, or **event space**
Sample space

Die rolling

If we have a 6-sided die

- Sample space $\Omega = \{\text{One, Two, Three, Four, Five, Six}\}$
- Event space $\mathcal{F} = \{\{\text{One}\}, \{\text{One, Two}\}, \{\text{One, Three, Five}\} \ldots \}$
  - With 6 options, there are $2^6 = 64$ distinct events
Principles of counting

- **Multiplication Principle**: if there are two independent events, $P$ and $Q$, and $P$ can happen in $p$ different ways and $Q$ in $q$ different ways, then $P$ and $Q$ can happen in $p \cdot q$ ways.

- **Addition Principle**: if there are two independent events, $P$ and $Q$, and $P$ can happen in $p$ different ways and $Q$ in $q$ different ways, then $P$ or $Q$ can happen in $p + q$ ways.
Principles of counting

Examples

▶ example 1: If there are 3 roads leading from Bloomington to Indianapolis and 5 roads from Indianapolis to Chicago, how many ways are there to get from Bloomington to Chicago?
Principles of counting

Examples

- example 1: If there are 3 roads leading from Bloomington to Indianapolis and 5 roads from Indianapolis to Chicago, how many ways are there to get from Bloomington to Chicago?

  answer: \( 3 \cdot 5 = 15 \)
Principles of counting

Examples

- example 1: If there are 3 roads leading from Bloomington to Indianapolis and 5 roads from Indianapolis to Chicago, how many ways are there to get from Bloomington to Chicago?
  
  answer: $3 \cdot 5 = 15$

- example 2: If there are 2 roads going south from Bloomington and 6 roads going north. How many roads are there going south or north?
Introduction to Probability Theory

**Counting**

**Examples**

- **Example 1:** If there are 3 roads leading from Bloomington to Indianapolis and 5 roads from Indianapolis to Chicago, how many ways are there to get from Bloomington to Chicago?
  
  answer: \(3 \cdot 5 = 15\)

- **Example 2:** If there are 2 roads going south from Bloomington and 6 roads going north. How many roads are there going south or north?
  
  answer: \(2 + 6 = 8\)
Exercises

- How many different 7-place license plates are possible if the first two places are for letters and the other 5 for numbers?
Exercises

- How many different 7-place license plates are possible if the first two places are for letters and the other 5 for numbers?
- John, Jim, Jack and Jay have formed a band consisting of 4 instruments.
  - If each of the boys can play all 4 instruments, how many different arrangements are possible?
  - What if John and Jim can play all 4 instruments, but Jay and Jack can each play only piano and drums?
Probability functions

Probabilities: if $A$ is an event, $P(A)$ is its probability

- $0 \leq P(A) \leq 1$

A **probability function (distribution)** distributes a probability mass of 1 over the sample space $\Omega$

- Probability function: any function $P : \mathcal{F} \rightarrow [0, 1]$, where:
  - $P(\Omega) = 1$
  - $A_j \in \mathcal{F}$: $P(\bigcup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} P(A_j)$ (countable additivity)

... for *disjoint* sets: $A_j \cap A_k = \emptyset$ for $j \neq k$

i.e., the probability of any event $A_j$ happening = sum of the probabilities of any individual event happening

- e.g., $P(\text{roll} = 1 \cup \text{roll} = 2) = P(\text{roll} = 1) + P(\text{roll} = 2)$
Example

Toss a fair coin three times. What is the chance of exactly 2 heads coming up?

- Sample space $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
- Event of interest $A = \{HHT, HTH, THH\}$

Since the coin is fair, we have a uniform distribution, i.e., each outcome is equally likely (1/8)

- $P(A) = \frac{|A|}{|\Omega|} = \frac{3}{8}$
Additivity for non-disjoint sets

\[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]

- The probability of unioning \( A \) and \( B \) requires adding up their individual probabilities
- ... then subtracting out their intersection, so as not to double count that portion
Conditional probability

The **conditional probability** of an event $A$ occurring given that event $B$ has already occurred is notated as $P(A|B)$

- Prior probability of $A$: $P(A)$
- Posterior probability of $A$ (after additional knowledge $B$): $P(A|B)$

\[
P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A, B)}{P(B)}
\]

- $P(A, B)$ (or $P(AB)$) is the **joint probability**
- In some sense, $B$ has become the sample space
Conditional probability

Example

- A coin is flipped twice. If we assume that all four points in the sample space
  \( \Omega = \{(H, H), (H, T), (T, H), (T, T)\} \)
  are equally likely, what is the conditional probability that both flips result in heads given that the first flip does?
Conditional probability

Example

- A coin is flipped twice. If we assume that all four points in the sample space
  \( \Omega = \{ (H, H), (H, T), (T, H), (T, T) \} \)
  are equally likely, what is the conditional probability that both flips result in heads given that the first flip does?
- \( A = \{ (H, H) \}, \ B = \{ (H, H), (H, T) \} \)
- \( P(A) = \frac{1}{4} \)
  \( P(B) = \frac{2}{4} \)
  \( P(A, B) = \frac{1}{4} \)
Conditional probability

Example

- A coin is flipped twice. If we assume that all four points in the sample space
  \( \Omega = \{(H, H), (H, T), (T, H), (T, T)\} \)
  are equally likely, what is the conditional probability that both flips result in heads given that the first flip does?

- \( A = \{(H, H)\}, B = \{(H, H), (H, T)\} \)

- \( P(A) = \frac{1}{4} \)
  \( P(B) = \frac{2}{4} \)
  \( P(A, B) = \frac{1}{4} \)

- \( P(A|B) = \frac{P(A, B)}{P(B)} = \frac{\frac{1}{4}}{\frac{2}{4}} = \frac{1}{2} \)
The chain rule

The multiplication rule restates $P(A|B) = \frac{P(A \cap B)}{P(B)}$:

(2) $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$

The chain rule (used in Markov models):

(3) $P(A_1 \cap ... \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)...P(A_n|\bigcap_{i=1}^{n-1} A_i)$

i.e., to obtain the probability of events occurring:

- select the first event
- select the second event, given the first
- ...
- select the $n^{th}$ event, given all the previous ones
Independence

Two events are independent if knowing one does not affect the probability of the other.

Events A and B are independent if

- \( P(A) = P(A|B) \)
- i.e., \( P(A \cap B) = P(A)P(B) \)

i.e., probability of seeing A and B together = product of seeing each one individually, as one does not affect other.
Indepedence

Example

- fair die: event A = divisible by two; event B = divisible by three
  - $P(AB) = P(\{\text{six}\}) = \frac{1}{6}$
  - $P(A) \times P(B) = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$

- event C = divisible by four
  - $P(C) = P(\{\text{four}\}) = \frac{1}{6}$
  - $P(AC) = P(\{\text{four}\}) = \frac{1}{6}$
  - $P(A) \times P(C) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$
Bayes’ Theorem allows one to calculate $P(B|A)$ in terms of $P(A|B)$

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

Getting the most likely event

Bayes’ Theorem takes into account the normalizing constant $P(A)$

\[(5) \quad P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}\]

If $P(A)$ is the same for every event of interest, and we want to find the value of $B$ which maximizes the function:

\[(6) \quad \arg \max_B \frac{P(A|B)P(B)}{P(A)} = \arg \max_B P(A|B)P(B)\]

... so, in these cases, we can ignore the denominator
Let $E$ and $F$ be events.

$$E = EF \cup EF^c$$

where $EF$ and $EF^c$ are mutually exclusive.

$$P(E) = P(EF) + P(EF^c)$$

$$= P(E|F)P(F) + P(E|F^c)P(F^c)$$

$$= P(E|F)P(F) + P(E|F^c)P(1 - P(F))$$
Example

- An insurance company groups people into two classes: Those who are accident-prone and those who are not.
- Their statistics show that an accident-prone person will have an accident within a fixed 1-year period with prob. 0.4. Whereas the prob. decreases to 0.2 for a non-accident-prone person.
- Let us assume that 30 percent of the population are accident prone.
- What is the prob. that a new policyholder will have an accident within a year of purchasing a policy?
Example (2)

- $Y = \text{policyholder will have an accident within one year}$
- $A = \text{policyholder is accident prone}$
- look for $P(Y)$

\[
P(Y) = P(Y|A)P(A) + P(Y|A^c)P(A^c) = 0.4 \times 0.3 + 0.2 \times 0.7 = 0.26
\]
Example (2)

- Y = policyholder will have an accident within one year
- A = policyholder is accident prone
- look for $P(Y)$

\[
P(Y) = P(Y|A)P(A) + P(Y|A^c)P(A^c)
\]
\[
= 0.4 \cdot 0.3 + 0.2 \cdot 0.7 = 0.26
\]

= 0.4 \cdot 0.3 + 0.2 \cdot 0.7 = 0.26
Let’s say we have $i$ different, disjoint sets $B_i$, and these sets partition $A$ (i.e., $A \subseteq \bigcup_i B_i$)

Then, the following is true:

\[
(7) \quad P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)
\]

This gives us a more general form of Bayes’ Theorem:

\[
(8) \quad P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}
\]
Example of Bayes’ Theorem

Assume the following:
- Bowl $B_1$ ($P(B_1) = \frac{1}{3}$) has 2 red and 4 white chips
- Bowl $B_2$ ($P(B_2) = \frac{1}{6}$) has 1 red and 2 white chips
- Bowl $B_3$ ($P(B_3) = \frac{1}{2}$) has 5 red and 4 white chips

Given that we have pulled a red chip, what is the probability that it came from bowl $B_1$? In other words, what is $P(B_1|R)$?
Example of Bayes’ Theorem

Assume the following:
- Bowl $B_1$ ($P(B_1) = \frac{1}{3}$) has 2 red and 4 white chips
- Bowl $B_2$ ($P(B_2) = \frac{1}{6}$) has 1 red and 2 white chips
- Bowl $B_3$ ($P(B_3) = \frac{1}{2}$) has 5 red and 4 white chips

Given that we have pulled a red chip, what is the probability that it came from bowl $B_1$? In other words, what is $P(B_1|R)$?

- $P(B_1) = \frac{1}{3}$
- $P(R|B_1) = \frac{2}{2+4} = \frac{1}{3}$
- $P(R) = P(B_1 \cap R) + P(B_2 \cap R) + P(B_3 \cap R)$
  $= P(R|B_1)P(B_1) + P(R|B_2)P(B_2) + P(R|B_3)P(B_3) = \frac{4}{9}$

So, we have:

$$P(B_1|R) = \frac{P(R|B_1)P(B_1)}{P(R)} = \frac{\frac{1}{3} \cdot \frac{1}{3}}{\frac{4}{9}} = \frac{1}{4}$$
Example of Bayes’ Theorem

Assume the following:
- Bowl $B_1$ ($P(B_1) = \frac{1}{3}$) has 2 red and 4 white chips
- Bowl $B_2$ ($P(B_2) = \frac{1}{6}$) has 1 red and 2 white chips
- Bowl $B_3$ ($P(B_3) = \frac{1}{2}$) has 5 red and 4 white chips

Given that we have pulled a red chip, what is the probability that it came from bowl $B_1$? In other words, what is $P(B_1|R)$?

- $P(B_1) = \frac{1}{3}$
- $P(R|B_1) = \frac{2}{2+4} = \frac{1}{3}$

\[
P(R) = P(B_1 \cap R) + P(B_2 \cap R) + P(B_3 \cap R) = P(R|B_1)P(B_1) + P(R|B_2)P(B_2) + P(R|B_3)P(B_3) = \frac{4}{9}
\]

So, we have:
\[
P(B_1|R) = \frac{P(R|B_1)P(B_1)}{P(R)} = \frac{(1/3)(1/3)}{4/9} = \frac{1}{4}
\]
## Random Variables
### Mapping Outcomes to Numerical Values

If we roll 2 dice:

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Numerical Value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>{(1,1)}</td>
<td>2</td>
<td>1/36</td>
</tr>
<tr>
<td>{(1,2),(2,1)}</td>
<td>3</td>
<td>1/18</td>
</tr>
<tr>
<td>{(1,3),(2,2),(3,1)}</td>
<td>4</td>
<td>1/12</td>
</tr>
<tr>
<td>{(1,4),(2,3),(3,2),(4,1)}</td>
<td>5</td>
<td>1/9</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>{(4,6),(5,5),(6,4)}</td>
<td>10</td>
<td>1/12</td>
</tr>
<tr>
<td>{(5,6),(6,5)}</td>
<td>11</td>
<td>1/18</td>
</tr>
<tr>
<td>{(6,6)}</td>
<td>12</td>
<td>1/36</td>
</tr>
</tbody>
</table>
Random Variables

Motivation

A random variable maps the set of outcomes $\Omega$ into the set of real numbers.

- Formally, a random variable is a function $X \rightarrow \mathbb{R}^n$

Motivation for random variables:

- They abstract away from outcomes by putting outcomes into equivalence classes
- Mapping to numerical values makes calculations easier.
- They facilitate numerical manipulations, such as the definition of mean and standard deviation.
Example

Suppose that our experiment consists of tossing 3 fair coins. If we let \( Y \) denote the number of heads appearing, then \( Y \) is a random variable taking on one of the values 0, 1, 2, 3 with respective probabilities.

\[
\begin{align*}
P\{ Y = 0 \} &= P\{(T, T, T)\} = \frac{1}{8} \\
P\{ Y = 1 \} &= P\{(T, T, H), (T, H, T), (H, T, T)\} = \frac{3}{8} \\
P\{ Y = 2 \} &= P\{(H, H, T), (H, T, H), (T, H, H)\} = \frac{3}{8} \\
P\{ Y = 3 \} &= P\{(H, H, H)\} = \frac{1}{8}
\end{align*}
\]
Example (2)

\[ P\{ Y = 0\} = P((T, T, T)) = \frac{1}{8} \]
\[ P\{ Y = 1\} = P((T, T, H), (T, H, T), (H, T, T)) = \frac{3}{8} \]
\[ P\{ Y = 2\} = P((H, H, T), (H, T, H), (T, H, H)) = \frac{3}{8} \]
\[ P\{ Y = 3\} = P((H, H, H)) = \frac{1}{8} \]

Since \( Y \) must take on one of the values 0 through 3,
\[ 1 = P(\bigcup_{i=0}^{3}\{ Y = i \}) = \sum_{i=0}^{3} P\{ Y = i \} \]
If $X$ is a discrete random variable with probability mass function $p(a)$, the *expectation* or *expected value* of $X$, denoted by $E[X]$, is defined by

$$E[X] = \sum_{x:p(x)>0} x \cdot p(x)$$
Expectation of a Random Variable

If $X$ is a discrete random variable with probability mass function $p(a)$, the expectation or expected value of $X$, denoted by $E[X]$, is defined by

$$E[X] = \sum_{x: p(x) > 0} x \cdot p(x)$$

In other words: $E[X]$ is the weighted average of the possible values of $X$, each value being weighted by the probability that $X$ has that particular value.
Example

Find $E[X]$ where $X$ is the outcome when we roll a fair die.

Solution

Since $p(1) = \cdots = p(6) = \frac{1}{6},$

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2}$$
Example (2)

Assume we have the following weighted die:

- $p(X = 1) = p(X = 2) = p(X = 5) = \frac{1}{6}$
- $p(X = 3) = p(X = 4) = \frac{1}{12}$
- $p(X = 6) = \frac{1}{3}$

What is the expectation here?
Example (3)

A school class of 120 students are driven in three buses to a symphonic performance. There are 36 students in one of the buses, 40 in another, and 44 in the third bus. When the buses arrive, one of the 120 students is randomly chosen. Let $X$ be a random variable denoting the students on the bus of that randomly chosen student.

Task: Find $E[X]$
Variance

Expectation alone ignores the question:

- Do the values of a random variable tend to be consistent over many trials or do they tend to vary a lot?

If \( X \) is a discrete random variable with mean \( \mu \), then the variance of \( X \), denoted by \( \text{Var}(X) \), is defined by

\[
\text{Var}(X) = E[(X - \mu)^2]
\]

\[
= E(X^2) - (E(X))^2
\]

The standard deviation of \( X \), denoted by \( \sigma(X) \), is defined as

\[
\sigma(X) = \sqrt{\text{Var}(X)}
\]
Example

Calculate $\text{Var}(X)$ if $X$ represents the outcome when a fair die is rolled.
Example

Calculate $\text{Var}(X)$ if $X$ represents the outcome when a fair die is rolled.

Solution

$$E[X^2] = 1^2 \left(\frac{1}{6}\right) + 2^2 \left(\frac{1}{6}\right) + 3^2 \left(\frac{1}{6}\right) + 4^2 \left(\frac{1}{6}\right) + 5^2 \left(\frac{1}{6}\right) + 6^2 \left(\frac{1}{6}\right)$$

$$= \left(\frac{1}{6}\right)(91)$$

Hence

$$\text{Var}(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$
Example for expectation and variance

When we roll two dice, what is the expectation and the variance for the sum of the numbers on the two dice?

\[
E(X) = E(Y + Y) \\
= E(Y) + E(Y) \\
= 3.5 + 3.5 = 7
\]

\[
(9)
\]

\[
\text{Var}(X) = E((X - E(X))^2) \\
= \sum_x p(x)(x - E(X))^2 \\
= \sum_x p(x)(x - 7)^2 = 5\frac{5}{6}
\]

\[
(10)
\]
Joint distributions

With (discrete) random variables, we can define:

- **Joint pmf**: The probability of both $x$ and $y$ happening

  \[ p(x, y) = P(X = x, Y = y) \]  

- **Marginal pmfs**: The probability of $x$ happening is the sum of the occurrences of $x$ with all the different $y$'s

  \[ p_X(x) = \sum_y p(x, y) \]  
  \[ p_Y(y) = \sum_x p(x, y) \]  

If $X$ and $Y$ are independent, then $p(x, y) = p_X(x)p_Y(y)$, so, e.g., the probability of rolling two sixes is:

- $p(X = 6, Y = 6) = p(X = 6)p(Y = 6) = \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{1}{36}$