Probabilistic Methods in Linguistics
Day 1

Roger Levy

UC San Diego
Department of Linguistics

September 27, 2012
What are probabilities?

You and your friend meet at the park for a game of tennis. In order to determine who will serve first, you jointly decide to flip a coin. Your friend produces a quarter and tells you that it is a fair coin. What exactly does your friend mean by this?
What are probabilities?

You and your friend meet at the park for a game of tennis. In order to determine who will serve first, you jointly decide to flip a coin. Your friend produces a quarter and tells you that it is a fair coin. What exactly does your friend mean by this?

- **Frequentist** answer:
  
  If you were to toss the coin many, many times, the proportion of Heads outcomes would be guaranteed to eventually approach 50%
What are probabilities?

You and your friend meet at the park for a game of tennis. In order to determine who will serve first, you jointly decide to flip a coin. Your friend produces a quarter and tells you that it is a fair coin. What exactly does your friend mean by this?

- **Frequentist** answer:
  
  If you were to toss the coin many, many times, the proportion of Heads outcomes would be guaranteed to eventually approach 50%.

- **Bayesian** answer:
  
  Your friend believes that Heads and Tails are equally likely outcomes if you flip the coin (and would be willing to take even odds on a bet regarding the outcome).
Huge warning!
Huge warning!

PROBABILITIES ARE NOT THE SAME THING AS OBSERVED FREQUENCIES!
Huge warning!

PROBABILITIES ARE NOT THE SAME THING AS OBSERVED FREQUENCIES!

- Most of this course is about *inferring* probabilities from frequencies. *Doing this is totally non-trivial!*
Huge warning!

PROBABILITIES ARE NOT THE SAME THING AS OBSERVED FREQUENCIES!

- Most of this course is about *inferring* probabilities from frequencies. Doing this is totally non-trivial!
- Usually I will be very clear in stating when I am dealing with frequencies, and when I am talking about probabilities
Huge warning!

PROBABILITIES ARE NOT THE SAME THING AS OBSERVED FREQUENCIES!

- Most of this course is about *inferring* probabilities from frequencies. *Doing this is totally non-trivial!*
- Usually I will be very clear in stating when I am dealing with frequencies, and when I am talking about probabilities.
- Occasionally I will explicitly blur the boundary, saying “we have so-and-so frequencies, *and let’s treat them as the probabilities too*”.
Huge warning!

PROBABILITIES ARE NOT THE SAME THING AS OBSERVED FREQUENCIES!

▸ Most of this course is about *inferring* probabilities from frequencies. *Doing this is totally non-trivial!*

▸ Usually I will be very clear in stating when I am dealing with frequencies, and when I am talking about probabilities.

▸ Occasionally I will explicitly blur the boundary, saying “we have so-and-so frequencies, *and let’s treat them as the probabilities too*.”

▸ But this is the exception, not the rule. If at any point you’re not sure whether I’m talking about probabilities or frequencies, *ask me*
An example of how frequencies are not probabilities

- You play seven tennis matches with your friend, and he flips the same coin each time to determine who serves first.
An example of how frequencies are not probabilities

- You play seven tennis matches with your friend, and he flips the same coin each time to determine who serves first.
- Of these seven flips, four land heads, three land tails.
An example of how frequencies are not probabilities

- You play seven tennis matches with your friend, and he flips the same coin each time to determine who serves first
- Of these seven flips, four land heads, three land tails
- These are the frequencies of heads and tails
You play seven tennis matches with your friend, and he flips the same coin each time to determine who serves first.

Of these seven flips, four land heads, three land tails.

These are the frequencies of heads and tails.

What are the probabilities of heads and tails for your friend’s coin?
An example of how frequencies are not probabilities

➤ You play seven tennis matches with your friend, and he flips the same coin each time to determine who serves first
➤ Of these seven flips, four land heads, three land tails
➤ These are the frequencies of heads and tails
➤ What are the probabilities of heads and tails for your friend’s coin?
➤ [They are unknown, but you have some idea of what they are]
Probability spaces

Traditionally, probability spaces are defined in terms of sets. An event $E$ is a subset of a sample space $\Omega$: $E \subset \Omega$. 
Probability spaces

Traditionally, probability spaces are defined in terms of sets. An event $E$ is a subset of a sample space $\Omega$: $E \subseteq \Omega$.
A probability space $P$ on a sample space $\Omega$ is a function from events $E$ in $\Omega$ to real numbers such that the following three axioms hold:

1. $P(E) \geq 0$ for all $E \subseteq \Omega$ (non-negativity).
2. If $E_1$ and $E_2$ are disjoint, then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ (disjoint union).
3. $P(\Omega) = 1$ (properness).
Probability spaces

Traditionally, probability spaces are defined in terms of sets. An event $E$ is a subset of a sample space $\Omega$: $E \subset \Omega$. A probability space $P$ on a sample space $\Omega$ is a function from events $E$ in $\Omega$ to real numbers such that the following three axioms hold:

1. $P(E) \geq 0$ for all $E \subset \Omega$ (non-negativity).
2. If $E_1$ and $E_2$ are disjoint, then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ (disjoint union).
3. $P(\Omega) = 1$ (properness).

We can also think of these things as involving logical rather than set relations:

- **Subset**: $A \subset B$  
  $B \rightarrow A$
- **Disjointness**: $E_1 \cap E_2 = \emptyset$  
  $\neg(E_1 \land E_2)$
- **Union**: $E_1 \cup E_2$  
  $E_1 \lor E_2$
Probability spaces: an example

Here are some relative frequencies of the first word in a sentence of English being each of the four open-class parts of speech:
Here are some relative frequencies of the first word in a sentence of English being each of the four open-class parts of speech:

N
Probability spaces: an example

Here are some relative frequencies of the first word in a sentence of English being each of the four open-class parts of speech:

N  12.6%
Probability spaces: an example

Here are some relative frequencies of the first word in a sentence of English being each of the four open-class parts of speech:

N 12.6%
V
Here are some relative frequencies of the first word in a sentence of English being each of the four open-class parts of speech:

N 12.6%
V 4.4%
Probability spaces: an example

Here are some relative frequencies of the first word in a sentence of English being each of the four open-class parts of speech:

N  12.6%
V  4.4%
Adj
Probability spaces: an example

Here are some relative frequencies of the first word in a sentence of English being each of the four open-class parts of speech:

N 12.6%
V 4.4%
Adj 2.1%
Probability spaces: an example

Here are some relative frequencies of the first word in a sentence of English being each of the four open-class parts of speech:

- **N** 12.6%
- **V** 4.4%
- **Adj** 2.1%
- **Adv**
Here are some relative frequencies of the first word in a sentence of English being each of the four open-class parts of speech:

- **N** 12.6%
- **V** 4.4%
- **Adj** 2.1%
- **Adv** 8.9%
Probability spaces: an example

Here are some relative frequencies of the first word in a sentence of English being each of the four open-class parts of speech:

- N 12.6%
- V 4.4%
- Adj 2.1%
- Adv 8.9%
Probability spaces: an example

Here are some relative frequencies of the first word in a sentence of English being each of the four open-class parts of speech:

- N 12.6%
- V 4.4%
- Adj 2.1%
- Adv 8.9%

Let’s consider these probabilities as well.
Here are some relative frequencies of the first word in a sentence of English being each of the four open-class parts of speech:

- N 12.6%
- V 4.4%
- Adj 2.1%
- Adv 8.9%

Let’s consider these probabilities as well.
What else can we derive about the probabilities of various events in the sample space of possible first words of a sentence of English?
The conditional probability of event $B$ given that $A$ has occurred/is known is defined as follows:

$$P(B|A) \equiv \frac{P(A \cap B)}{P(A)}$$
Conditional Probability: an example

Hypothetical probabilities from historical English:

<table>
<thead>
<tr>
<th>Object</th>
<th>Pronoun</th>
<th>Not Pronoun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preverbal</td>
<td>0.224</td>
<td>0.655</td>
</tr>
<tr>
<td>Postverbal</td>
<td>0.014</td>
<td>0.107</td>
</tr>
</tbody>
</table>
Conditional Probability: an example

Hypothetical probabilities from historical English:

<table>
<thead>
<tr>
<th></th>
<th>Pronoun</th>
<th>Not Pronoun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object Preverbal</td>
<td>0.224</td>
<td>0.655</td>
</tr>
<tr>
<td>Object Postverbal</td>
<td>0.014</td>
<td>0.107</td>
</tr>
</tbody>
</table>

How do we calculate the following?

\[ P(\text{Pronoun}|\text{Postverbal}) \]
Conditional Probability: an example

Hypothetical probabilities from historical English:

<table>
<thead>
<tr>
<th></th>
<th>Pronoun</th>
<th>Not Pronoun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object <strong>Preverbal</strong></td>
<td>0.224</td>
<td>0.655</td>
</tr>
<tr>
<td>Object <strong>Postverbal</strong></td>
<td>0.014</td>
<td>0.107</td>
</tr>
</tbody>
</table>

How do we calculate the following?

\[ P(\text{Pronoun}|\text{Postverbal}) \]
Conditional Probability: an example

Hypothetical probabilities from historical English:

<table>
<thead>
<tr>
<th></th>
<th>Pronoun</th>
<th>Not Pronoun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object Preverbal</td>
<td>0.224</td>
<td>0.655</td>
</tr>
<tr>
<td>Object Postverbal</td>
<td>0.014</td>
<td>0.107</td>
</tr>
</tbody>
</table>

How do we calculate the following?

\[
P(\text{Pronoun}|\text{Postverbal}) = \frac{P(\text{Postverbal} \cap \text{Pronoun})}{P(\text{Postverbal})}
\]

\[
= \frac{0.014}{0.014 + 0.107} = 0.116
\]
The chain rule

Conditional probability generalizes to more than two variables:
The chain rule

Conditional probability generalizes to more than two variables:

\[ P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_n | E_1 \cap E_2 \cap \cdots \cap E_{n-1}) \cdots P(E_2 | E_1)P(E_1) \]
Bayes’ Rule

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

With extra "background" random variables \( I \):

\[ P(A|B, I) = \frac{P(B|A, I)P(A|I)}{P(B|I)} \]
Bayes’ Rule, more closely inspected

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

Likelihood \quad Prior

Normalizing constant
Bayes’ Rule in action

Returning to hypothetical old English

\[
P(\text{Object } \text{Animate}) = 0.4
\]
\[
P(\text{Object } \text{Postverbal}|\text{Object } \text{Animate}) = 0.7
\]
\[
P(\text{Object } \text{Postverbal}|\text{Object } \text{Inanimate}) = 0.8
\]

Imagine you’re an incremental sentence processor. You encounter a transitive verb but haven’t encountered the object yet. How likely is it that the object is animate? I.e., compute

\[
P(\text{Anim}|\text{PostV})
\]
(Conditional) Independence

Events $A$ and $B$ are said to be Conditionally Independent given information $C$ if

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

(1)
A discrete random variable $X$ is literally a function from the sample space $\Omega$ of a probability space to a finite, or countably infinite, set of real numbers ($\mathbb{R}$). This is the probability mass function.
Discrete random variables and probability mass functions

A discrete random variable $X$ is literally a function from the sample space $\Omega$ of a probability space to a finite, or countably infinite, set of real numbers ($\mathbb{R}$). This is the probability mass function.

Simplest example: a Bernoulli trial has two outcomes, “success” and “failure”, which are associated with the integers 1 and 0. If $X$ is a Bernoulli-distributed random variable, then

$$P(X = x) = \begin{cases} 
\pi & \text{if } x = 1 \\
1 - \pi & \text{if } x = 0 \\
0 & \text{otherwise}
\end{cases}$$
Multinomial trials

Generalizing the Bernoulli trial to cases with multiple outcomes $c_1, \ldots, c_r$, we get a multinomial trial with $r - 1$ parameters $\pi_1, \ldots, \pi_{r-1}$.

$$P(X = x) = \begin{cases} 
\pi_1 & \text{if } x = c_1 \\
\pi_2 & \text{if } x = c_2 \\
\vdots & \\
\pi_{r-1} & \text{if } x = c_{r-1} \\
1 - \sum_{i=1}^{r-1} \pi_i & \text{if } x = c_r \\
0 & \text{otherwise}
\end{cases}$$
Example of multinomial trials

You decide to pull *Alice in Wonderland* off your bookshelf, open to a random page, put your finger down randomly on that page, and record the letter that your finger is resting on.
Example of multinomial trials

You decide to pull *Alice in Wonderland* off your bookshelf, open to a random page, put your finger down randomly on that page, and record the letter that your finger is resting on.

In *Alice in Wonderland*, 12.6% of the letters are e, 9.9% are t, 8.2% are a, and so forth.
Example of multinomial trials

You decide to pull *Alice in Wonderland* off your bookshelf, open to a random page, put your finger down randomly on that page, and record the letter that your finger is resting on.

In *Alice in Wonderland*, 12.6% of the letters are e, 9.9% are t, 8.2% are a, and so forth.

We could write the parameters of our model as $\pi_e = 0.126$, $\pi_t = 0.099$, $\pi_a = 0.082$, and so forth.