Probabilistic Methods in Linguistics
Lecture 2

Roger Levy

UC San Diego
Department of Linguistics

October 2, 2012
A bit of review & terminology

- A Bernoulli distribution was defined as

\[
P(X = x) = \begin{cases} 
\pi & \text{if } x = 1 \\
1 - \pi & \text{if } x = 0 \\
0 & \text{otherwise}
\end{cases}
\]
A bit of review & terminology

- A Bernoulli distribution was defined as

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

- The probability of success, \( \pi \), is the Bernoulli parameter
A bit of review & terminology

- A **Bernoulli distribution** was defined as

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

- The probability of success, \( \pi \), is the **Bernoulli parameter**
- Each different choice of \( \pi \) gives us a different specific probability mass function
A bit of review & terminology

- A **Bernoulli distribution** was defined as

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

- The probability of success, \(\pi\), is the **Bernoulli parameter**
- Each different choice of \(\pi\) gives us a different specific probability mass function
- All these specific probability mass functions are members of the same **parametric family of probability distributions**
A bit of review & terminology

- A **Bernoulli distribution** was defined as

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

- The probability of success, \( \pi \), is the **Bernoulli parameter**
- Each different choice of \( \pi \) gives us a different specific probability mass function
- All these specific probability mass functions are members of the same **parametric family of probability distributions**
- This—the parametric family—is actually what we mean when we use the shorthand terminology “Bernoulli distribution”
A bit of review & terminology

- A **Bernoulli distribution** was defined as

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

- The probability of success, \( \pi \), is the **Bernoulli parameter**
- Each different choice of \( \pi \) gives us a different specific probability mass function
- All these specific probability mass functions are members of the same **parametric family of probability distributions**
- This—the parametric family—is actually what we mean when we use the shorthand terminology “Bernoulli distribution”
- You can think of a parametric family of probability distributions as a “recipe”: give it one or more parameters, and you have a probability mass function
More on Discrete Random Variables

- The probability function for a discrete random variable is called a PROBABILITY MASS FUNCTION

...(this is known as the EXPONENTIAL distribution)
More on Discrete Random Variables

- The probability function for a discrete random variable is called a PROBABILITY MASS FUNCTION
- Probability mass functions are bounded above by 1, to ensure properness

(this is known as the EXPONENTIAL distribution)
More on Discrete Random Variables

- The probability function for a discrete random variable is called a **probability mass function**
- Probability mass functions are bounded above by 1, to ensure properness
- This means that there are some real numbers which have a non-zero probability of occurring

(this is known as the **exponential distribution**)
More on Discrete Random Variables

- The probability function for a discrete random variable is called a **probability mass function**
- Probability mass functions are bounded above by 1, to ensure properness
- This means that there are some real numbers which have a non-zero probability of occurring
- For properness, the set of such numbers must be **countable**

(this is known as the **exponential distribution**)
More on Discrete Random Variables

- The probability function for a discrete random variable is called a **PROBABILITY MASS FUNCTION**
- Probability mass functions are bounded above by 1, to ensure properness
- This means that there are some real numbers which have a non-zero probability of occurring
- For properness, the set of such numbers must be **COUNTABLE**
- It could still be infinite!—e.g., a simple probabilistic model of sentence length:

\[
P(\text{sentence has } n \text{ words}) = \frac{1}{2^n} \quad n \in 1, 2, \ldots
\]

(this is known as the **EXPONENTIAL** distribution)
Normalized and unnormalized probability distributions

- I’ve made a big deal out of PROPENSITY thus far
Normalized and unnormalized probability distributions

- I’ve made a big deal out of **properness** thus far
- In practice, however, we’ll sometimes encounter a function $F$ that satisfies all the properties of a probability function except for properness
Normalized and unnormalized probability distributions

- I’ve made a big deal out of **properness** thus far
- In practice, however, we’ll sometimes encounter a function $F$ that satisfies all the properties of a probability function except for properness
- In these cases, let us define the **partition function** $Z$ as

$$Z \overset{\text{def}}{=} \sum_x F(x)$$
I’ve made a big deal out of **properness** thus far.

In practice, however, we’ll sometimes encounter a function $F$ that satisfies all the properties of a probability function except for properness.

In these cases, let us define the **partition function** $Z$ as

$$Z \overset{\text{def}}{=} \sum_x F(x)$$

It must be the case that $Z > 0$ (why?)
Normalized and unnormalized probability distributions

- I’ve made a big deal out of **properness** thus far
- In practice, however, we’ll sometimes encounter a function $F$ that satisfies all the properties of a probability function except for properness
- In these cases, let us define the **partition function** $Z$ as

$$Z \overset{\text{def}}{=} \sum_x F(x)$$

- It must be the case that $Z > 0$ (why?)
- If $Z$ is finite, then we can define a probability distribution based on $F$:

$$P(X = x) = \frac{1}{Z} F(x)$$

$Z$ is sometimes called the **normalizing constant** (as well as “partition function”)
Normalized and unnormalized probability distributions

- I’ve made a big deal out of **properness** thus far
- In practice, however, we’ll sometimes encounter a function $F$ that satisfies all the properties of a probability function except for properness
- In these cases, let us define the **partition function** $Z$ as

$$Z \overset{\text{def}}{=} \sum_x F(x)$$

- It must be the case that $Z > 0$ (why?)
- If $Z$ is finite, then we can define a probability distribution based on $F$:

$$P(X = x) = \frac{1}{Z} F(x)$$

$Z$ is sometimes called the **normalizing constant** (as well as “partition function”)
- In these situations we also sometimes write

$$P(x) \propto F(x)$$
Example: constituent-order model

- Suppose that we wanted to construct a probability distribution over total orderings of three constituents Subject, Object, Verb of simple transitive clauses
Example: constituent-order model

- Suppose that we wanted to construct a probability distribution over total orderings of three constituents **Subject**, **Object**, **Verb** of simple transitive clauses
- There are $6 (=3!)$ logically possible orderings of $\{S,O,V\}$
  - **SOV** Japanese, Hindi
  - **SVO** English, Mandarin
  - **VSO** Irish, Classical Arabic
  - **OSV** Nias (Austronesian; Indonesia)
  - **OVS** Hixkaryana (Carib; Brazil)
  - **OSV** Nadëb (Nadahup; Brazil)

Hence a five-parameter multinomial distribution could capture any logically possible probability distributions over constituent orders
Constituent-order model

- **General principle:** it is often of interest to look for *more constrained* probability distributions that nevertheless closely match reality
Constituent-order model

- **General principle**: it is often of interest to look for *more constrained* probability distributions that nevertheless closely match reality
  - Finding such distributions often indicates we have extracted some generalization about the world
Constituent-order model

- **General principle:** it is often of interest to look for *more constrained* probability distributions that nevertheless closely match reality
  - Finding such distributions often indicates we have extracted some generalization about the world
- In our case, a widespread idea in typological circles is that word-order patterns can be reduced down to pairwise preferences about relative constituent orders
Constituent-order model

- **General principle:** it is often of interest to look for *more constrained* probability distributions that nevertheless closely match reality
  - Finding such distributions often indicates we have extracted some generalization about the world
- In our case, a widespread idea in typological circles is that word-order patterns can be reduced down to pairwise preferences about relative constituent orders
- There are three constituent pairs: \{S,O\}, \{V,O\}, \{S,V\}
Constituent-order model

- **General principle:** it is often of interest to look for *more constrained* probability distributions that nevertheless closely match reality
  - Finding such distributions often indicates we have extracted some generalization about the world
- In our case, a widespread idea in typological circles is that word-order patterns can be reduced down to pairwise preferences about relative constituent orders
- There are three constituent pairs: \{S,O\}, \{V,O\}, \{S,V\}
- Imagine an analogy between a Bernoulli trial for each constituent-pair ordering

\[
F(S \prec O) \propto \gamma_1 \\
F(V \prec O) \propto \gamma_2 \\
F(S \prec V) \propto \gamma_3
\]

where \(X \prec Y\) means that “\(X\) linearly precedes \(Y\)”
Constituent-order model

\[ F(S \prec O) \propto \gamma_1 \]
\[ F(V \prec O) \propto \gamma_2 \]
\[ F(S \prec V) \propto \gamma_3 \]

However, the three precedence events are not totally independent: there are two outcomes that are impossible:

\( (1) S \prec O, V \prec S, O \prec V \) \hspace{1cm} \( (2) O \prec S, S \prec V, V \prec O \)
Constituent-order model

\begin{align*}
F(S \prec O) & \propto \gamma_1 \\
F(V \prec O) & \propto \gamma_2 \\
F(S \prec V) & \propto \gamma_3
\end{align*}

- However, the three precedence events are not totally independent: there are two outcomes that are impossible:

  (1) $S \prec O, V \prec S, O \prec V$  
  (2) $O \prec S, S \prec V, V \prec O$

- Hence $F$ is not technically a probability function: it is not proper
Constituent-order model

\[ F(S \preceq O) \propto \gamma_1 \]
\[ F(V \preceq O) \propto \gamma_2 \]
\[ F(S \preceq V) \propto \gamma_3 \]

However, the three precedence events are not totally independent: there are two outcomes that are impossible:

(1) \( S \preceq O, V \preceq S, O \preceq V \)  \hspace{1cm} (2) \( O \preceq S, S \preceq V, V \preceq O \)

Hence \( F \) is not technically a probability function: it is not proper.

But we can use \( F \) as the basis of a probability function by computing its normalizing constant!
Constituent Ordering Model

<table>
<thead>
<tr>
<th></th>
<th>S__O</th>
<th>S__V</th>
<th>O__V</th>
<th>Outcome X</th>
<th>F(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>≺</td>
<td>≺</td>
<td>≺</td>
<td>SOV</td>
<td>( \gamma_1 \gamma_2 \gamma_3 )</td>
</tr>
<tr>
<td>2</td>
<td>≺</td>
<td>≺</td>
<td>≺</td>
<td>SVO</td>
<td>( \gamma_1 \gamma_2 (1 - \gamma_3) )</td>
</tr>
<tr>
<td>3</td>
<td>≺</td>
<td>≺</td>
<td>≺</td>
<td>impossible</td>
<td>( \gamma_1 (1 - \gamma_2) \gamma_3 )</td>
</tr>
<tr>
<td>4</td>
<td>≺</td>
<td>≺</td>
<td>≺</td>
<td>VSO</td>
<td>( \gamma_1 (1 - \gamma_2) (1 - \gamma_3) )</td>
</tr>
<tr>
<td>5</td>
<td>≺</td>
<td>≺</td>
<td>≺</td>
<td>OSV</td>
<td>( (1 - \gamma_1) \gamma_2 \gamma_3 )</td>
</tr>
<tr>
<td>6</td>
<td>≺</td>
<td>≺</td>
<td>≺</td>
<td>impossible</td>
<td>( (1 - \gamma_1) \gamma_2 (1 - \gamma_3) )</td>
</tr>
<tr>
<td>7</td>
<td>≺</td>
<td>≺</td>
<td>≺</td>
<td>OVS</td>
<td>( (1 - \gamma_1) (1 - \gamma_2) \gamma_3 )</td>
</tr>
<tr>
<td>8</td>
<td>≺</td>
<td>≺</td>
<td>≺</td>
<td>VOS</td>
<td>( (1 - \gamma_1) (1 - \gamma_2) (1 - \gamma_3) )</td>
</tr>
</tbody>
</table>

- So the normalizing constant is \( Z = 1 - F(X_3) - F(X_6) \)
Constituent Ordering Model


<table>
<thead>
<tr>
<th>Order</th>
<th>SOV</th>
<th>SVO</th>
<th>VSO</th>
<th>OSV</th>
<th>OVS</th>
<th>VOS</th>
</tr>
</thead>
<tbody>
<tr>
<td># Langs</td>
<td>566</td>
<td>488</td>
<td>95</td>
<td>25</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>Rel. Freq.</td>
<td>0.476</td>
<td>0.410</td>
<td>0.080</td>
<td>0.021</td>
<td>0.009</td>
<td>0.003</td>
</tr>
</tbody>
</table>
Constituent Ordering Model


<table>
<thead>
<tr>
<th>Order</th>
<th>SOV</th>
<th>SVO</th>
<th>VSO</th>
<th>OSV</th>
<th>OVS</th>
<th>VOS</th>
</tr>
</thead>
<tbody>
<tr>
<td># Langs</td>
<td>566</td>
<td>488</td>
<td>95</td>
<td>25</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>Rel. Freq.</td>
<td>0.476</td>
<td>0.410</td>
<td>0.080</td>
<td>0.021</td>
<td>0.009</td>
<td>0.003</td>
</tr>
</tbody>
</table>

... and constituent-order probabilities for
\( \gamma_1 = 0.9, \gamma_2 = 0.8, \gamma_3 = 0.5 \)
Constituent Ordering Model


<table>
<thead>
<tr>
<th>Order</th>
<th>SOV</th>
<th>SVO</th>
<th>VSO</th>
<th>OSV</th>
<th>OVS</th>
<th>VOS</th>
</tr>
</thead>
<tbody>
<tr>
<td># Langs</td>
<td>566</td>
<td>488</td>
<td>95</td>
<td>25</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>Rel. Freq.</td>
<td>0.476</td>
<td>0.410</td>
<td>0.080</td>
<td>0.021</td>
<td>0.009</td>
<td>0.003</td>
</tr>
<tr>
<td>Probs</td>
<td>0.414</td>
<td>0.414</td>
<td>0.103</td>
<td>0.046</td>
<td>0.011</td>
<td>0.011</td>
</tr>
</tbody>
</table>

- ... and constituent-order probabilities for \( \gamma_1 = 0.9, \gamma_2 = 0.8, \gamma_3 = 0.5 \)
- So this model can produce probabilities that are in the ballpark of the empirical frequencies!
Constituent Ordering Model


<table>
<thead>
<tr>
<th>Order</th>
<th>SOV</th>
<th>SVO</th>
<th>VSO</th>
<th>OSV</th>
<th>OVS</th>
<th>VOS</th>
</tr>
</thead>
<tbody>
<tr>
<td># Langs</td>
<td>566</td>
<td>488</td>
<td>95</td>
<td>25</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>Rel. Freq.</td>
<td>0.476</td>
<td>0.410</td>
<td>0.080</td>
<td>0.021</td>
<td>0.009</td>
<td>0.003</td>
</tr>
<tr>
<td>Probs</td>
<td>0.414</td>
<td>0.414</td>
<td>0.103</td>
<td>0.046</td>
<td>0.011</td>
<td>0.011</td>
</tr>
</tbody>
</table>

- ...and constituent-order probabilities for
  - $\gamma_1 = 0.9$, $\gamma_2 = 0.8$, $\gamma_3 = 0.5$
- So this model can produce probabilities that are in the ballpark of the empirical frequencies!
- Actually, this model can do even better than this fit; we’ll cover the principles of how to find the best fit in Chapter 4
Continuous random variables and probability density functions

- Sometimes we want to model distributions on a CONTINUUM of possible outcomes:
Continuous random variables and probability density functions

- Sometimes we want to model distributions on a CONTINUUM of possible outcomes:
  - The amount of time an infant lives before it hears its first parasitic gap construction
Continuous random variables and probability density functions

- Sometimes we want to model distributions on a \text{CONTINUUM} of possible outcomes:
  - The amount of time an infant lives before it hears its first parasitic gap construction
  - Formant frequencies for different vowel productions
Continuous random variables and probability density functions

- Sometimes we want to model distributions on a continuum of possible outcomes:
  - The amount of time an infant lives before it hears its first parasitic gap construction
  - Formant frequencies for different vowel productions
- We use continuous random variables for this
Continuous random variables and probability density functions

- Sometimes we want to model distributions on a continuum of possible outcomes:
  - The amount of time an infant lives before it hears its first parasitic gap construction
  - Formant frequencies for different vowel productions
- We use continuous random variables for this
- Because there are uncountably many possible outcomes, we cannot use a probability mass function
Continuous random variables and probability density functions

- Sometimes we want to model distributions on a continuum of possible outcomes:
  - The amount of time an infant lives before it hears its first parasitic gap construction
  - Formant frequencies for different vowel productions
- We use continuous random variables for this
- Because there are uncountably many possible outcomes, we cannot use a probability mass function
- Instead, continuous random variables have a probability density function $p(x)$ assigning non-negative density to every real number
Continuous random variables and probability density functions

- Sometimes we want to model distributions on a **continuum** of possible outcomes:
  - The amount of time an infant lives before it hears its first parasitic gap construction
  - Formant frequencies for different vowel productions
- We use **continuous random variables** for this
- Because there are **uncountably many** possible outcomes, we cannot use a probability mass function
- Instead, continuous random variables have a **probability density function** $p(x)$ assigning non-negative density to every real number
- For continuous random variables, properness requires that
  \[ \int_{-\infty}^{\infty} p(x) \, dx = 1 \]
Simplest possible continuous probability distribution

- The **uniform distribution** is a two-parameter family defined as

\[ P(x|a, b) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \]
Simplest possible continuous probability distribution

- The **uniform distribution** is a two-parameter family defined as

\[
P(x|a, b) = \begin{cases} 
\dfrac{1}{b-a} & a \leq x \leq b \\
0 & \text{otherwise}
\end{cases}
\]

- Its probability density function looks pretty simple:
Simplest possible continuous probability distribution

- The **UNIFORM DISTRIBUTION** is a two-parameter family defined as

\[
P(x|a, b) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}
\]

- Its probability density function looks pretty simple:

- Note that the density function need not be bounded above by 1!
A terminological note

- A lot of the time I will have things to say that apply both to discrete and continuous probability distributions
A terminological note

- A lot of the time I will have things to say that apply both to discrete and continuous probability distributions
- In these cases I will sometimes use the term *density* to refer to either a probability density function (continuous) or a probability mass function (discrete)
A terminological note

- A lot of the time I will have things to say that apply both to discrete and continuous probability distributions.
- In these cases I will sometimes use the term *density* to refer to either a probability density function (continuous) or a probability mass function (discrete).
- Sometimes I will also use the term *discrete density* to mean probability mass function.
A terminological note

- A lot of the time I will have things to say that apply both to discrete and continuous probability distributions.
- In these cases I will sometimes use the term *density* to refer to either a probability density function (continuous) or a probability mass function (discrete).
- Sometimes I will also use the term *discrete density* to mean probability mass function.
- If you’re unsure about my usage, ask!
Expectation and Variance

- You’re probably pretty familiar with the idea of a **mean**, and perhaps with **variance** as well
Expectation and Variance

- You’re probably pretty familiar with the idea of a **mean**, and perhaps with **variance** as well.
- These terms are ambiguous between being properties of a **probability distribution** or of a **sample** of data.
You’re probably pretty familiar with the idea of a \textit{mean}, and perhaps with \textit{variance} as well.

These terms are ambiguous between being properties of a \textit{probability distribution} or of a \textit{sample} of data.

Here we’ll cover their definitions with respect to probability distributions.
The expected value of a random variable $X$ is defined as

$$E(X) = \sum_{i} x_i P(X = x_i)$$

for discrete random variables, and as

$$E(X) = \int_{-\infty}^{\infty} x \, p(x) \, dx$$

for continuous random variables.
Expectation and Variance

▶ The **expected value** of a random variable $X$ is defined as

\[ E(X) = \sum_i x_i P(X = x_i) \]

for discrete random variables, and as

\[ E(X) = \int_{-\infty}^{\infty} x \, p(x) \, dx \]

for continuous random variables

▶ It is variously written as $E(X)$, $E[X]$, $\langle X \rangle$, $\mu_X$, or just $\mu$
Expectation and Variance

- The **expected value** of a random variable $X$ is defined as

\[ E(X) = \sum_{i} x_i P(X = x_i) \]

for discrete random variables, and as

\[ E(X) = \int_{-\infty}^{\infty} x \cdot p(x) \, dx \]

for continuous random variables

- It is variously written as $E(X)$, $E[X]$, $\langle X \rangle$, $\mu_X$, or just $\mu$.
- It is sometimes also called the **mean**.
Expectation and Variance

- The **expected value** of a random variable $X$ is defined as

$$E(X) = \sum_i x_i P(X = x_i)$$

for discrete random variables, and as

$$E(X) = \int_{-\infty}^{\infty} x \, p(x) \, dx$$

for continuous random variables.

- It is variously written as $E(X)$, $E[X]$, $\langle X \rangle$, $\mu_X$, or just $\mu$.

- It is sometimes also called the **mean**.

- What is the expected value of a Bernoulli-distributed random variable? of a uniform-distributed random variable?
Variance

- Variance is a “second-order” mean: it quantifies how broadly dispersed the outcomes of the r.v. are.
Variance

- Variance is a “second-order” mean: it quantifies how broadly dispersed the outcomes of the r.v. are.

- Definition:

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

or equivalently,

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

- Variance is a “second-order” mean: it quantifies how broadly dispersed the outcomes of the r.v. are.
- Definition:

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

or equivalently,

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

- What is the variance of a Bernoulli random variable? When is its variance smallest? largest?
The normal distribution

- The **NORMAL DISTRIBUTION** is perhaps the continuous distribution you’re most likely to encounter.
The normal distribution

- The NORMAL DISTRIBUTION is perhaps the continuous distribution you’re most likely to encounter.
- It’s a two-parameter distribution: the mean $\mu$ and the variance $\sigma^2$. 
The normal distribution

- The **NORMAL DISTRIBUTION** is perhaps the continuous distribution you’re most likely to encounter.
- It’s a two-parameter distribution: the mean \( \mu \) and the variance \( \sigma^2 \).
- Its probability density function is:

\[
p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right]
\]
The normal distribution

- The **normal distribution** is perhaps the continuous distribution you’re most likely to encounter.
- It’s a two-parameter distribution: the mean $\mu$ and the variance $\sigma^2$.
- Its probability density function is:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right]$$

- We’ll spend some time deconstructing this scary-looking function... soon you will come to know and love it!
Normal distributions with different means and variances

\[ p(x) \]

- \( \mu=0, \sigma^2=1 \)
- \( \mu=0, \sigma^2=2 \)
- \( \mu=0, \sigma^2=0.5 \)
- \( \mu=2, \sigma^2=1 \)