

PROBABILITY

David Kauchak  
CS159 – Spring 2011

## Admin

- Posted some links in Monday's lecture for regular expressions
- Logging in remotely
  - ▣ ssh to vpn.cs.pomona.edu (though don't run anything here!) or project.cs.pomona.edu
  - ▣ use your Pomona CS login
  - ▣ ssh to one of the lab machines
    - cs227-33.cs.pomona.edu (cs227-43, cs227-44)
    - In regex terms: cs227-[1-4][1-4]\.cs\.pomona\.edu

## Next Monday

## Corpus statistics

<p>Number of times the word <b>'government'</b> was used in Obama's State of the Union addresses.</p> <table border="0" style="margin-left: 20px;"> <tr> <td style="text-align: center;">7</td> <td style="text-align: center;">13</td> <td style="text-align: center;">18</td> </tr> <tr> <td style="text-align: center;">2009</td> <td style="text-align: center;">2010</td> <td style="text-align: center;">2011</td> </tr> </table> <p>Number of times the word <b>'dream'</b> was used in Obama's State of the Union addresses.</p> <table border="0" style="margin-left: 20px;"> <tr> <td style="text-align: center;">2</td> <td style="text-align: center;">3</td> <td style="text-align: center;">12</td> </tr> <tr> <td style="text-align: center;">2009</td> <td style="text-align: center;">2010</td> <td style="text-align: center;">2011</td> </tr> </table> <p>Number of times the word <b>'education'</b> was used in Obama's State of the Union addresses.</p> <table border="0" style="margin-left: 20px;"> <tr> <td style="text-align: center;">14</td> <td style="text-align: center;">5</td> <td style="text-align: center;">11</td> </tr> <tr> <td style="text-align: center;">2009</td> <td style="text-align: center;">2010</td> <td style="text-align: center;">2011</td> </tr> </table> <p>Number of times the word <b>'industry'</b> was used in Obama's State of the Union addresses.</p> <table border="0" style="margin-left: 20px;"> <tr> <td style="text-align: center;">5</td> <td style="text-align: center;">1</td> <td style="text-align: center;">9</td> </tr> <tr> <td style="text-align: center;">2009</td> <td style="text-align: center;">2010</td> <td style="text-align: center;">2011</td> </tr> </table>	7	13	18	2009	2010	2011	2	3	12	2009	2010	2011	14	5	11	2009	2010	2011	5	1	9	2009	2010	2011	<p>Number of times the word <b>'innovation'</b> was used in Obama's State of the Union addresses.</p> <table border="0" style="margin-left: 20px;"> <tr> <td style="text-align: center;">2</td> <td style="text-align: center;">3</td> <td style="text-align: center;">9</td> </tr> <tr> <td style="text-align: center;">2009</td> <td style="text-align: center;">2010</td> <td style="text-align: center;">2011</td> </tr> </table> <p>Number of times the word <b>'win'</b> was used in Obama's State of the Union addresses.</p> <table border="0" style="margin-left: 20px;"> <tr> <td style="text-align: center;">1</td> <td style="text-align: center;">2</td> <td style="text-align: center;">9</td> </tr> <tr> <td style="text-align: center;">2009</td> <td style="text-align: center;">2010</td> <td style="text-align: center;">2011</td> </tr> </table> <p>Number of times the word <b>'Afghan(istan)'</b> was used in Obama's State of the Union addresses.</p> <table border="0" style="margin-left: 20px;"> <tr> <td style="text-align: center;">2</td> <td style="text-align: center;">5</td> <td style="text-align: center;">8</td> </tr> <tr> <td style="text-align: center;">2009</td> <td style="text-align: center;">2010</td> <td style="text-align: center;">2011</td> </tr> </table> <p>Number of times the word <b>'Internet'</b> was used in Obama's State of the Union addresses.</p> <table border="0" style="margin-left: 20px;"> <tr> <td style="text-align: center;">0</td> <td style="text-align: center;">0</td> <td style="text-align: center;">6</td> </tr> <tr> <td style="text-align: center;">2009</td> <td style="text-align: center;">2010</td> <td style="text-align: center;">2011</td> </tr> </table>	2	3	9	2009	2010	2011	1	2	9	2009	2010	2011	2	5	8	2009	2010	2011	0	0	6	2009	2010	2011
7	13	18																																															
2009	2010	2011																																															
2	3	12																																															
2009	2010	2011																																															
14	5	11																																															
2009	2010	2011																																															
5	1	9																																															
2009	2010	2011																																															
2	3	9																																															
2009	2010	2011																																															
1	2	9																																															
2009	2010	2011																																															
2	5	8																																															
2009	2010	2011																																															
0	0	6																																															
2009	2010	2011																																															

[www.nytimes.com](http://www.nytimes.com) 1/25/2011

## Regular expression by language

- **grep**
  - command-line tool for regular expressions (general regular expression print/parser)
  - returns all lines that match a regular expression
  - `grep "@" twitter.posts`
  - `grep "http:" twitter.posts`
  - can't use metacharacters (`\d`, `\w`), use `[]` instead

## Regular expression by language

- **sed**
  - another command-line tool using that regexs to print and manipulate strings
  - very powerful, though we'll just play with it
  - Most common is substitution:
    - `sed "s/ is a / is not a/" twitter.posts`
    - `sed "s/+ / /" twitter.posts`
    - `grep "#" twitter.posts | sed -E "s/#([a-zA-Z]+)/REF: \1/g"`
      - Use `-E` if you want to use character classes
      - `\1` references the first match (i.e. whatever is matched in `...`)
  - Can also do things like delete all that match, etc.

## Basic Probability Theory: terminology

- An **experiment** has a set of potential outcomes, e.g., throw a dice, "look at" another sentence
- The **sample space** of an experiment is the set of all possible outcomes, e.g.,  $\{1, 2, 3, 4, 5, 6\}$
- In NLP our sample spaces tend to be **very large**
  - All words, bigrams, 5-grams
  - All sentences of length 20 (given a finite vocabulary)
  - All sentences
  - All parse trees over a given sentence

## Basic Probability Theory: terminology

- An **event** is a subset of the sample space
- Dice rolls
  - $\{2\}$
  - $\{3, 6\}$
  - $\text{even} = \{2, 4, 6\}$
  - $\text{odd} = \{1, 3, 5\}$
- NLP
  - a particular word/part of speech occurring in a sentence
  - a particular topic discussed (politics, sports)
  - sentence with a parasitic gap
  - pick your favorite phenomena...

## Events

- We're interested in probabilities of events
  - $p(\{2\})$
  - $p(\text{even})$
  - $p(\text{odd})$
  - $p(\text{parasitic gap})$
  - $p(\text{word})$

## Random variables

- A random variable is a mapping from the sample space to a number (think events)
- It represents all the possible values of something we want to measure in an experiment
- For example, random variable,  $X$ , could be the number of heads for a coin

space	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
$X$	3	2	2	1	2	1	1	0

- Really for notational convenience, since the event space can sometimes be irregular

## Random variables

- We can then talk about the probability of the different values of a random variable
- The definition of probabilities over *all* of the possible values of a random variable defines a **probability distribution**

space	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
$X$	3	2	2	1	2	1	1	0

$X$	$P(X)$
3	$P(X=3) = 1/8$
2	$P(X=2) = 3/8$
1	$P(X=1) = 3/8$
0	$P(X=0) = 1/8$

## Probability distribution

- To be explicit
  - A probability distribution assigns probability values to all possible values of a random variable
  - These values must be  $\geq 0$  and  $\leq 1$
  - These values must sum to 1 for all possible values of the random variable

$X$	$P(X)$
3	$P(X=3) = 1/2$
2	$P(X=2) = 1/2$
1	$P(X=1) = 1/2$
0	$P(X=0) = 1/2$

$X$	$P(X)$
3	$P(X=3) = -1$
2	$P(X=2) = 2$
1	$P(X=1) = 0$
0	$P(X=0) = 0$

## Unconditional/prior probability

- Simplest form of probability is
  - $P(X)$
- Prior probability: without any additional information, what is the probability
  - What is the probability of a heads?
  - What is the probability of a sentence containing a pronoun?
  - What is the probability of a sentence containing the word "banana"?
  - What is the probability of a document discussing politics?
  - ...

## Joint distributions

- We can also talk about probability distributions over multiple variables
- $P(X,Y)$ 
  - probability of  $X$  and  $Y$
  - a distribution over the cross product of possible values

NLPPass	P(NLPPass)	NLPPass AND EngPass	P(NLPPass, EngPass)
true	0.89	true, true	.88
false	0.11	true, false	.01
EngPass	P(EngPass)	false, true	.04
true	0.92	false, false	.07
false	0.08		

## Joint distribution

- Still a probability distribution
  - all values between 0 and 1, inclusive
  - all values sum to 1
- All questions/probabilities of the two variables can be calculate from the joint distribution

NLPPass AND EngPass	P(NLPPass, EngPass)
true, true	.88
true, false	.01
false, true	.04
false, false	.07

What is P(ENGPass)?

## Joint distribution

- Still a probability distribution
  - all values between 0 and 1, inclusive
  - all values sum to 1
- All questions/probabilities of the two variables can be calculate from the joint distribution

NLPPass AND EngPass	P(NLPPass, EngPass)
true, true	.88
true, false	.01
false, true	.04
false, false	.07

0.92

How did you figure that out?

## Joint distribution

$$P(x) = \sum_{y \in Y} p(x, y)$$

NLPPass AND EngPass	P(NLPPass, EngPass)
true, true	.88
true, false	.01
false, true	.04
false, false	.07

## Conditional probability

- As we learn more information, we can update our probability distribution
- $P(X|Y)$  models this (read "probability of X given Y")
  - What is the probability of a heads given that both sides of the coin are heads?
  - What is the probability the document is about politics, given that it contains the word "Clinton"?
  - What is the probability of the word "banana" given that the sentence also contains the word "split"?
- Notice that it is still a distribution over the values of X

## Conditional probability

$$p(X|Y) = ?$$



In terms of prior and joint distributions, what is the conditional probability distribution?


## Conditional probability

$$p(X|Y) = \frac{P(X, Y)}{P(Y)}$$



Given that y has happened, what proportion of those events does x also happen

### Conditional probability

$$p(X|Y) = \frac{P(X,Y)}{P(Y)}$$


Given that y has happened, what proportion of those events does x also happen

NLPPass AND EngPass	P(NLPPass, EngPass)
true, true	.88
true, false	.01
false, true	.04
false, false	.07

What is:  $p(\text{NLPPass}=\text{true} \mid \text{EngPass}=\text{false})?$

### Conditional probability

$$p(X|Y) = \frac{P(X,Y)}{P(Y)}$$

NLPPass AND EngPass	P(NLPPass, EngPass)
true, true	.88
true, false	.01
false, true	.04
false, false	.07

What is:  $p(\text{NLPPass}=\text{true} \mid \text{EngPass}=\text{false})?$

$$\frac{P(\text{true}, \text{false}) = 0.01}{P(\text{EngPass} = \text{false}) = 0.01 + 0.07 = 0.08} = 0.125$$

Notice this is very different than  $p(\text{NLPPass}=\text{true}) = 0.89$

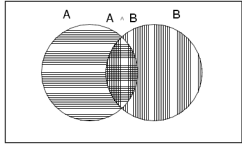
### A note about notation

- When talking about a particular assignment, you should technically write  $p(X=x)$ , etc.
- However, when it's clear, we'll often shorten it
- Also, we may also say  $P(X)$  or  $p(x)$  to generically mean any particular value, i.e.  $P(X=x)$

$$\frac{P(\text{true}, \text{false}) = 0.01}{P(\text{EngPass} = \text{false}) = 0.01 + 0.07 = 0.08} = 0.125$$

### Properties of probabilities

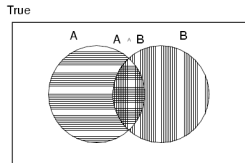
- $P(A \text{ or } B) = ?$



True

## Properties of probabilities

$$\square P(A \text{ or } B) = P(A) + P(B) - P(A, B)$$



## Properties of probabilities

$$\square P(\neg E) = 1 - P(E)$$

□ More generally:

□ Given events  $E = e_1, e_2, \dots, e_n$

$$p(e_i) = 1 - \sum_{j=1, j \neq i}^n p(e_j)$$

$$\square P(E1, E2) \leq P(E1)$$

## Chain rule (aka product rule)

$$p(X|Y) = \frac{P(X,Y)}{P(Y)} \quad \Rightarrow \quad p(X,Y) = P(X|Y)P(Y)$$

We can view calculating the probability of X AND Y occurring as two steps:

1. Y occurs with some probability  $P(Y)$
2. Then, X occurs, given that Y has occurred

or you can just trust the math... 😊

## Chain rule

$$p(X,Y,Z) = P(X|Y,Z)P(Y,Z)$$

$$p(X,Y,Z) = P(X,Y|Z)P(Z)$$

$$p(X,Y,Z) = P(X|Y,Z)P(Y|Z)P(Z)$$

$$p(X,Y,Z) = P(Y,Z|X)P(X)$$

$$p(X_1, X_2, \dots, X_n) = ?$$

## Applications of the chain rule

- We saw that we could calculate the individual prior probabilities using the joint distribution

$$p(x) = \sum_{y \in Y} p(x, y)$$

- What if we don't have the joint distribution, but do have conditional probability information:

- $P(Y)$
- $P(X|Y)$

$$p(x) = \sum_{y \in Y} p(y)p(x|y)$$

## Bayes' rule (theorem)

$$p(X|Y) = \frac{P(X,Y)}{P(Y)} \implies p(X,Y) = P(X|Y)P(Y)$$

$$p(Y|X) = \frac{P(X,Y)}{P(X)} \implies p(X,Y) = P(Y|X)P(X)$$

$$p(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

## Bayes rule

- Allows us to talk about  $P(Y|X)$  rather than  $P(X|Y)$
- Sometimes this can be more intuitive
- Why?

$$p(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

## Bayes rule

- $p(\text{disease} | \text{symptoms})$ 
  - For everyone who had those symptoms, how many had the disease?
  - $p(\text{symptoms} | \text{disease})$ 
    - For everyone that had the disease, how many had this symptom?
- $p(\text{linguistic phenomena} | \text{features})$ 
  - For all examples that had those features, how many had that phenomena?
  - $p(\text{features} | \text{linguistic phenomena})$ 
    - For all the examples with that phenomena, how many had this feature
- $p(\text{cause} | \text{effect})$  vs.  $p(\text{effect} | \text{cause})$



### Gaps

I just won't put these away.

↓  
direct object

These, I just won't put away.

↙  
filler

I just won't put \_\_\_\_\_ away.

gap

### Gaps

What did you put \_\_\_\_\_ away.

gap

The socks that I put \_\_\_\_\_ away.

gap

### Gaps

Whose socks did you fold \_\_\_\_\_ and put \_\_\_\_\_ away?

gap                  gap

↓

Whose socks did you fold \_\_\_\_\_ ?

gap

Whose socks did you put \_\_\_\_\_ away?

gap

### Parasitic gaps

These I'll put \_\_\_\_\_ away without folding \_\_\_\_\_ .

gap                  gap

↓

These I'll put \_\_\_\_\_ away.

gap

These without folding \_\_\_\_\_ .

gap

## Parasitic gap

These I'll put gap away without folding gap .

### 1. Cannot exist by themselves (parasitic)

These I'll put my pants away without folding gap .

### 2. They're optional

These I'll put gap away without folding them.

## Parasitic gaps

- <http://literal-minded.wordpress.com/2009/02/10/dougs-parasitic-gap/>

## Frequency of parasitic gaps

- Parasitic gaps occur on average in 1/100,000 sentences
- Problem:
  - Joe Linguist has developed a complicated set of regular expressions to try and identify parasitic gaps. If a sentence has a parasitic gap, it correctly identifies it 95% of the time. If it doesn't, it will incorrectly say it does with probability 0.005. Suppose we run it on a sentence and the algorithm says it is a parasitic gap, what is the probability it actually is?

## Prob of parasitic gaps

Joe Linguist has developed a complicated set of regular expressions to try and identify parasitic gaps. If a sentence has a parasitic gap, it correctly identifies it 95% of the time. If it doesn't, it will incorrectly say it does with probability 0.005. Suppose we run it on a sentence and the algorithm says it is a parasitic gap, what is the probability it actually is?

G = gap  
T = test positive

What question do we want to ask?

### Prob of parasitic gaps

Joe Linguist has developed a complicated set of regular expressions to try and identify parasitic gaps. If a sentence has a parasitic gap, it correctly identifies it 95% of the time. If it doesn't, it will incorrectly say it does with probability 0.005. Suppose we run it on a sentence and the algorithm says it is a parasitic gap, what is the probability it actually is?

G = gap  
T = test positive

$$p(g | t) = \frac{p(t | g)p(g)}{p(t)}$$

$$= \frac{p(t | g)p(g)}{\sum_{g \in G} p(g)p(t | g)} = \frac{p(t | g)p(g)}{p(g)p(t | g) + p(\bar{g})p(t | \bar{g})}$$

### Prob of parasitic gaps

Joe Linguist has developed a complicated set of regular expressions to try and identify parasitic gaps. If a sentence has a parasitic gap, it correctly identifies it 95% of the time. If it doesn't, it will incorrectly say it does with probability 0.005. Suppose we run it on a sentence and the algorithm says it is a parasitic gap, what is the probability it actually is?

G = gap  
T = test positive

$$p(g | t) = \frac{p(t | g)p(g)}{p(g)p(t | g) + p(\bar{g})p(t | \bar{g})}$$

$$= \frac{0.95 * 0.00001}{0.00001 * 0.95 + 0.99999 * 0.005} \approx 0.002$$

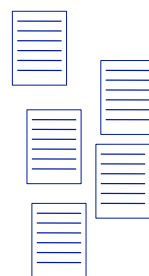
### Obtaining probabilities



- We've talked a lot about probabilities, but not where they come from
  - What is the probability of "the" occurring in a sentence?
  - What is the probability of "Pomona" occurring in a sentence?
  - What is the probability of "I think today is a good day to be me" as a sentence?

### Obtaining probabilities

training data



## Estimating probabilities

- What is the probability of “the” occurring in a sentence?

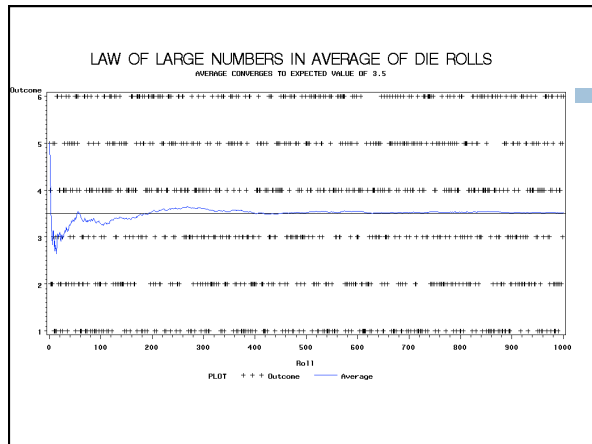
We don't know!

We can *estimate* that based on data, though:

$$\frac{\text{number of sentences with "the"}}{\text{total number of sentences}}$$

## Maximum likelihood estimation

- Intuitive
- Sets the probabilities so as to maximize the probability of the training data
- Problems?
  - Amount of data
    - particularly problematic for rare events
  - Is our training data representative



## Back to parasitic gaps

- Say the actual probability is 1/100,000
- We don't know this, though, so we're estimating it from a small data set of 10K sentences
- What is the probability that, by chance, we have a parasitic gap sentence in our sample?

## Back to parasitic gaps

- $p(\text{not\_parasitic}) = 0.99999$
- $p(\text{not\_parasitic})^{10000} \approx 0.905$  is the probability of us NOT finding one
- So, probability of us finding one is  $\sim 10\%$ , in which case we would incorrectly assume that the probability is  $1/10,000$  (10 times too large)
- Remember Zipf's law from last time... NLP is all about rare events!