Lecture 5

August 4, 2016

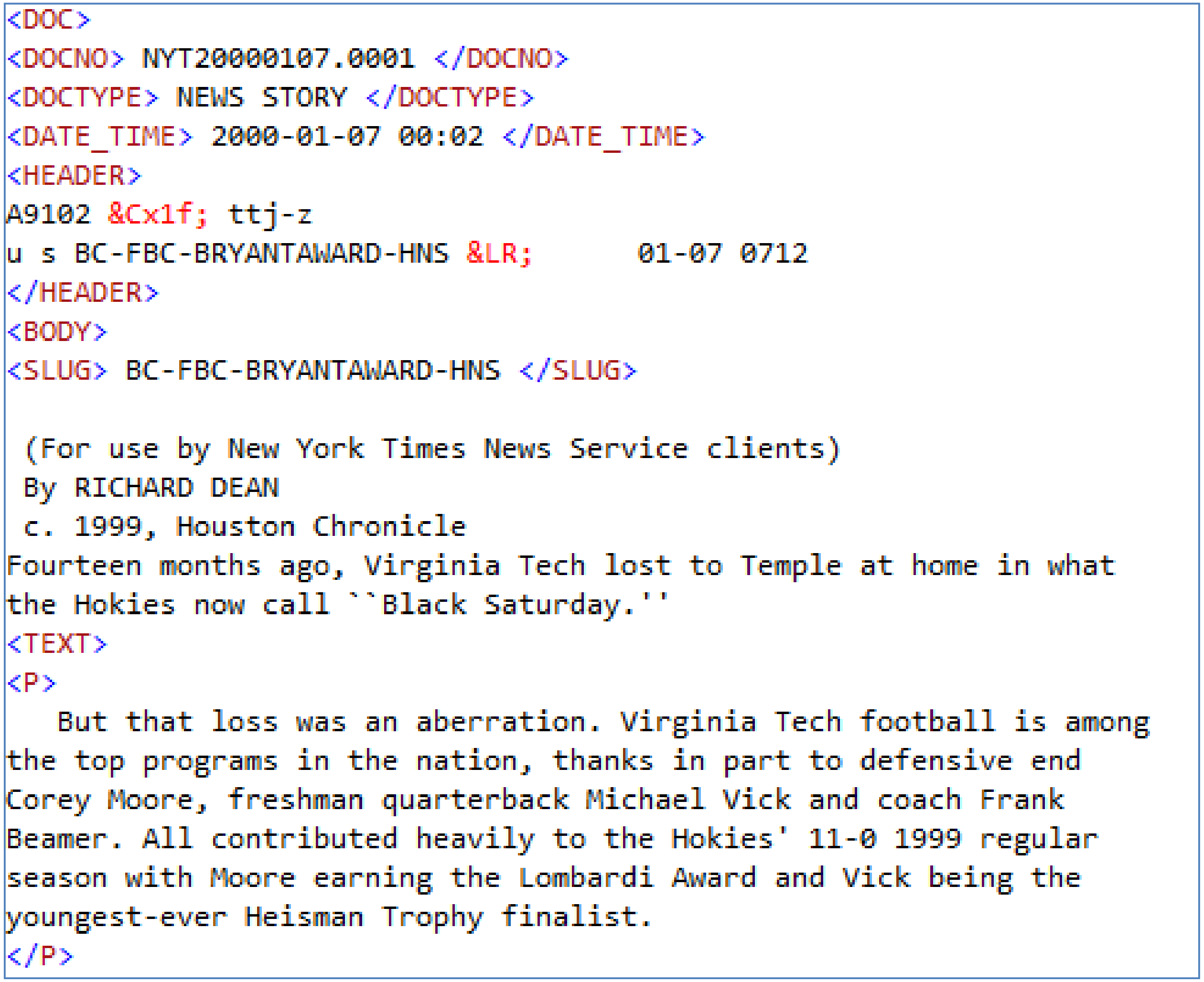
Bayes Theorem

Reminder: start the recording

Project 2

- http://courses.washington.edu/ling473/Project2.pdf
- SGML format
- Steps
  - clean unwanted characters
  - tokenize into words
  - tally each distinct word
  - sort by tally (descending)
  - output each word, with its tally

SGML



```
<DOC>
<DOCNO> NYT20000107.0001 </DOCNO>
<DOCTYPE> NEWS STORY </DOCTYPE>
<DATE_TIME> 2000-01-07 00:02 </DATE_TIME>
<HEADER>
A9102 &Cx1f; ttj-z
u s BC-FBC-BRYANTAWARD-HNS &LR;      01-07 0712
</HEADER>
<BODY>
<SLUG> BC-FBC-BRYANTAWARD-HNS </SLUG>

 (For use by New York Times News Service clients)
 By RICHARD DEAN
 c. 1999, Houston Chronicle
Fourteen months ago, Virginia Tech lost to Temple at home in what
the Hokies now call ``Black Saturday.''
<TEXT>
<P>
   But that loss was an aberration. Virginia Tech football is among
the top programs in the nation, thanks in part to defensive end
Corey Moore, freshman quarterback Michael Vick and coach Frank
Beamer. All contributed heavily to the Hokies' 11-0 1999 regular
season with Moore earning the Lombardi Award and Vick being the
youngest-ever Heisman Trophy finalist.
</P>
```

Writing assignment

- Due September 6th , 2016
  http://courses.washington.edu/ling473/writing-assignment.html
- Short Critical review of a paper from the computational linguistics literature
- Formatted according to ACL-2015 guidelines
  – http://acl2015.org/call_for_papers.html
- Any published journal or peer-reviewed paper on a comp. ling. topic is acceptable

independent random variables

Random variables AA and BB are independent *iff*

$$PP(AA \cap BB) = PP(AA)PP(BB)$$

Recall conditional probability:

$$PP(AA \cap BB) = PP(AA|BB)PP(BB)$$

This means that, if AA and BB are independent,

$$PP(AA|BB) = PP(AA)$$

$PP(AA \cap BB)$ $PP(AA, BB)$ $PP(AABB)$

Reminder: these are three notations for the same thing:
the joint probability of *A* and *B*. That is, that both events occur in a single trial

Conditional independence

A and *B* are independent *iff*

$$PP(AA \cap BB) = PP(AA)PP(BB)$$

A and *B* are conditionally independent given *K* *iff*

$$PP(AA \cap BB|KK) = PP(AA|KK)PP(BB|KK)$$

Just as with conditional probability, *K* constrains the sample space. Conditional independence means that *A* and *B* are independent if we know that *K* has occurred.

Conditional independence

A and *B* are conditionally independent given *K* *iff*

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

Given that *K* has occurred, knowing that *B* has occurred gives us no additional information about the probability of *A* (and vice-versa)

Q: Does this imply that *A* and *B* are independent?

A: No. *A* and *B* could be either independent or dependent in the absence of knowledge about *K*

Conditional independence

$$PP(AA \cap BB|KK) = PP(AA|KK)PP(BB|KK)$$

Two events (*A* and *B*) are conditionally independent given a third event (*K*) if their probabilities conditioned on K are independent. The following will also be true:

$$PP(AA|BB \cap KK) = PP(AA|KK)$$
$$PP(BB|AA \cap KK) = PP(BB|KK)$$

To show this derivation, let's go to the next slide.

Conditional independence

1. $PP(EE \cap KK) = PP(EE|KK)PP(KK)$ definition of conditional probability

2. $PP(AA \cap BB \cap KK) = PP(AA \cap BB|KK)PP(KK)$ substitute $AA \cap BB$ for EE in (1)

3. $PP(AA \cap BB \cap KK) = PP(AA|BB \cap KK)PP(BB \cap KK)$ substitute $BB \cap KK$ for KK in (1)

4. $PP(AA \cap BB|KK)PP(KK) = PP(AA|BB \cap KK)PP(BB \cap KK)$ equate (2) and (3)

5. $\boxed{PP(AA \cap BB|KK) = PP(AA|KK)PP(BB|KK)}$ the def. of conditional independence

6. $\boxed{PP(AA|KK)PP(BB|KK)}\,PP(KK) = PP(AA|BB \cap KK)PP(BB \cap KK)$ substitute (5) into (4)

7. $PP(AA|KK)\cancel{PP(BB|KK)PP(KK)} = PP(AA|BB \cap KK)\cancel{PP(BB|KK)PP(KK)}$

8. $PP(AA|BB \cap KK) = PP(AA|KK)$
   $PP(BB|AA \cap KK) = PP(BB|KK)$ cancel

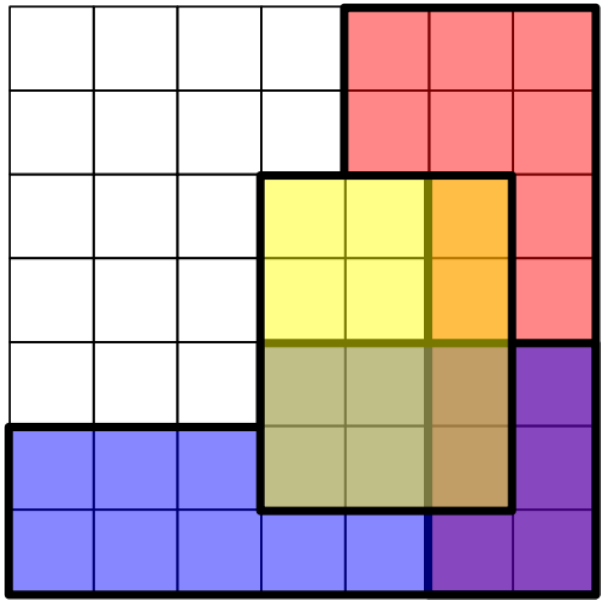
reminder: conditional probability

$$PP(AA \cap BB) = PP(AA|BB)PP(BB)$$

$$PP(AA|BB) = \frac{PP(AA \cap BB)}{PP(BB)}$$

The "given" notation '|' has lowest precedence

Conditional independence



$$P(R) = \frac{16}{49}$$

$$P(B) = \frac{18}{49}$$

$$P(Y) = \frac{12}{49}$$

none of these are independent

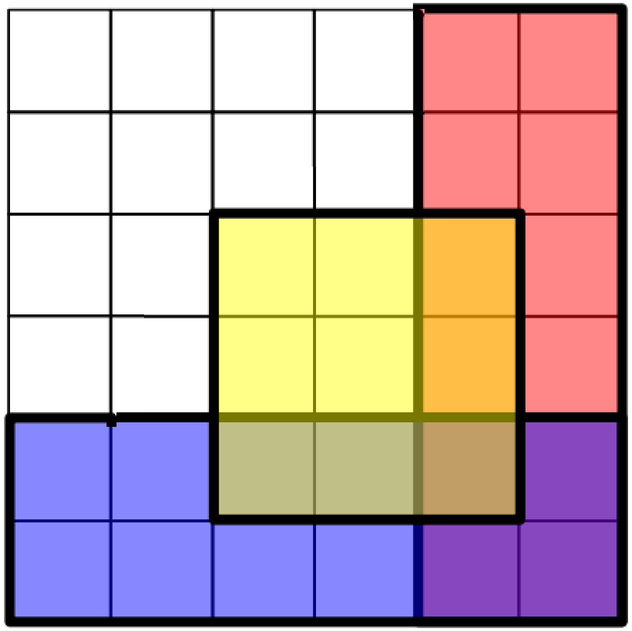
$$P(R|Y) = \frac{4}{12} = \frac{1}{3}$$

$$P(B|Y) = \frac{6}{18} = \frac{1}{3}$$

$$P(R \cap B|Y) = \frac{2}{12} = \frac{1}{6}$$

but R and B are conditionally independent given Y

Conditional independence



$$P(R) = \frac{12}{36} = \frac{1}{3}$$

$$P(B) = \frac{12}{36} = \frac{1}{3}$$

$$P(R \cap B) = \frac{4}{36} = \frac{1}{9} = P(R)P(B)$$

R and B are independent

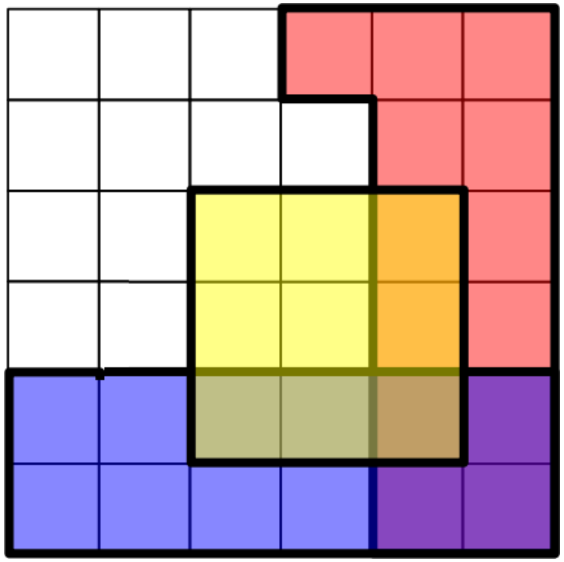
$$P(R|Y) = \frac{3}{9} = \frac{1}{3}$$

$$P(B|Y) = \frac{3}{9} = \frac{1}{3}$$

$$P(R \cap B|Y) = \frac{1}{9} = P(R|Y)P(B|Y)$$

R and B are {also, still} conditionally independent given Y

Conditional independence



$$P(R) = \frac{13}{36}$$

$$P(B) = \frac{12}{36} = \frac{1}{3}$$

$$P(R \cap B) = \frac{4}{36} = .1111$$

$$P(R)P(B) = \frac{13}{108} = .1214$$

…these are not equal, so R and B are dependent. But…

$$P(R|Y) = \frac{3}{9} = \frac{1}{3}$$

$$P(B|Y) = \frac{3}{9} = \frac{1}{3}$$

$$P(R \cap B|Y) = \frac{1}{9}$$

R and B can be *conditionally* independent given Y, even if they are dependent in the absence of information about Y

Review: Derivation of Bayes Theorem

$$PP(AA|BB) = \frac{PP(AA \cap BB)}{PP(BB)}$$

$$PP(BB|AA) = \frac{PP(BB \cap AA)}{PP(AA)}$$

$$PP(AA|BB)PP(BB) = PP(AA \cap BB)$$

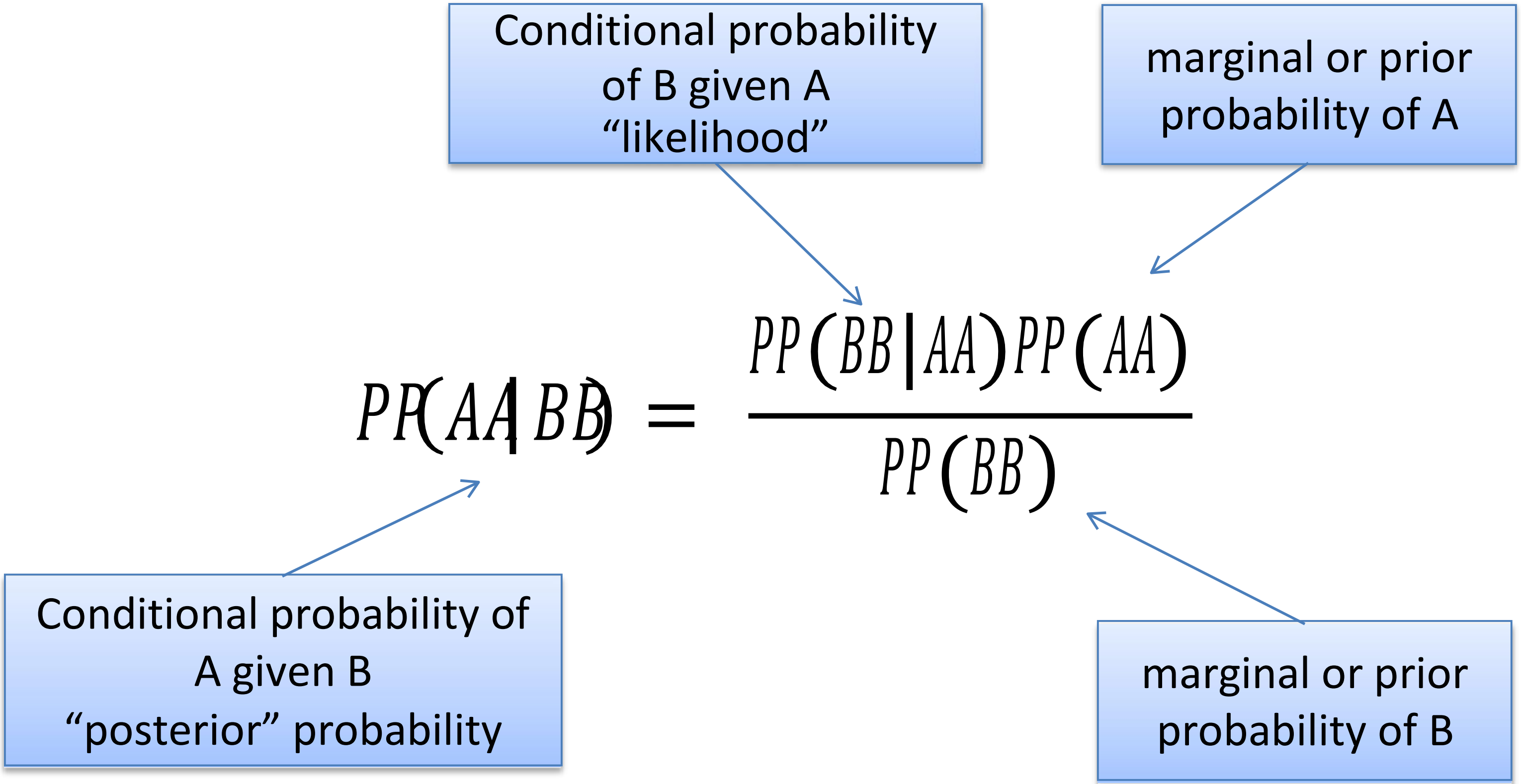
$$PP(BB|AA)PP(AA) = PP(BB \cap AA)$$

$$PP(AA|BB)PP(BB) = PP(BB|AA)PP(AA)$$

$$PP(AA|BB) = \frac{PP(BB|AA)PP(AA)}{PP(BB)}$$

Bayes Theorem

Rev. Thomas Bayes (1701-1761)



Conditional probability of B given A
“likelihood”

marginal or prior probability of A

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

Conditional probability of A given B
“posterior” probability

marginal or prior probability of B

Bayes Theorem

- Expresses one conditional probability in terms of its inverse
- $PP(AA|BB)$ depends not only on BB, but also on $PP(AA)$ and $PP(BB)$ in the general population

		actual condition AA	
		yes $PP(AA)$	no
test result BB	positive $PP(BB)$	true positive	false positive
	negative	false negative	true negative

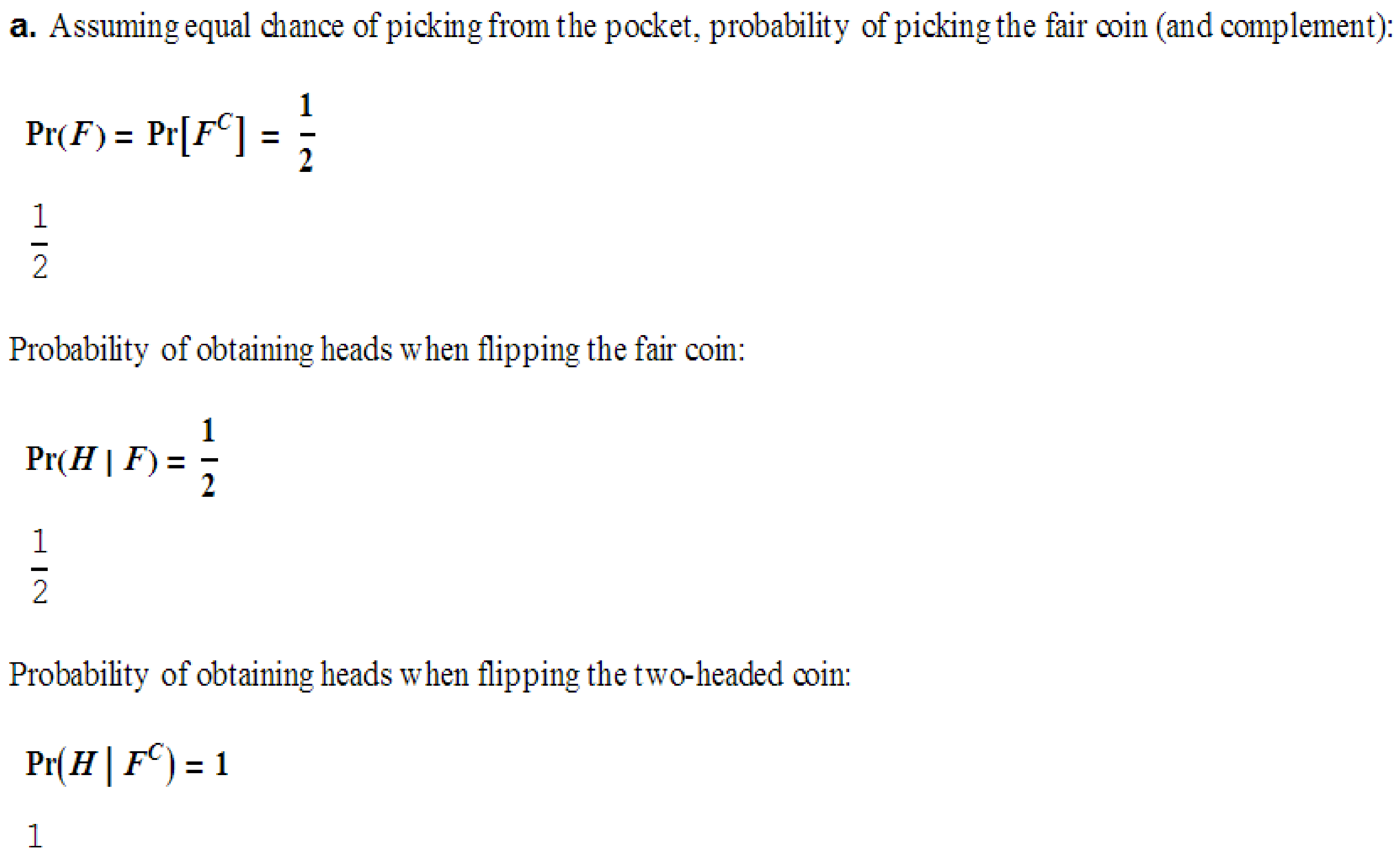
Recipe for Bayes Theorem

- What you need:
  1. The probability of actually satisfying the criteria (regardless of 2)
  2. The probability of testing positive for the criteria (regardless of 1)
  3. And either:
     a. the probability of testing positive given the criteria is satisfied
     b. the probability of satisfying the criteria given the test is positive

example #1

A gambler has two coins in his pocket, one fair coin and one two-headed one.

a. He selects one at random and flips it. It comes up heads. What is the probability that is the fair coin?



a. Assuming equal chance of picking from the pocket, probability of picking the fair coin (and complement):

$$\Pr(F) = \Pr\left[F^C\right] = \frac{1}{2}$$

$$\frac{1}{2}$$

Probability of obtaining heads when flipping the fair coin:

$$\Pr(H \mid F) = \frac{1}{2}$$

$$\frac{1}{2}$$

Probability of obtaining heads when flipping the two-headed coin:

$$\Pr\left(H \mid F^C\right) = 1$$

$$1$$

Overall prior probability of flipping heads

$$\Pr(H) = \Pr(F)\Pr(H \mid F) + \Pr\left(F^C\right)\Pr\left(H \mid F^C\right)$$

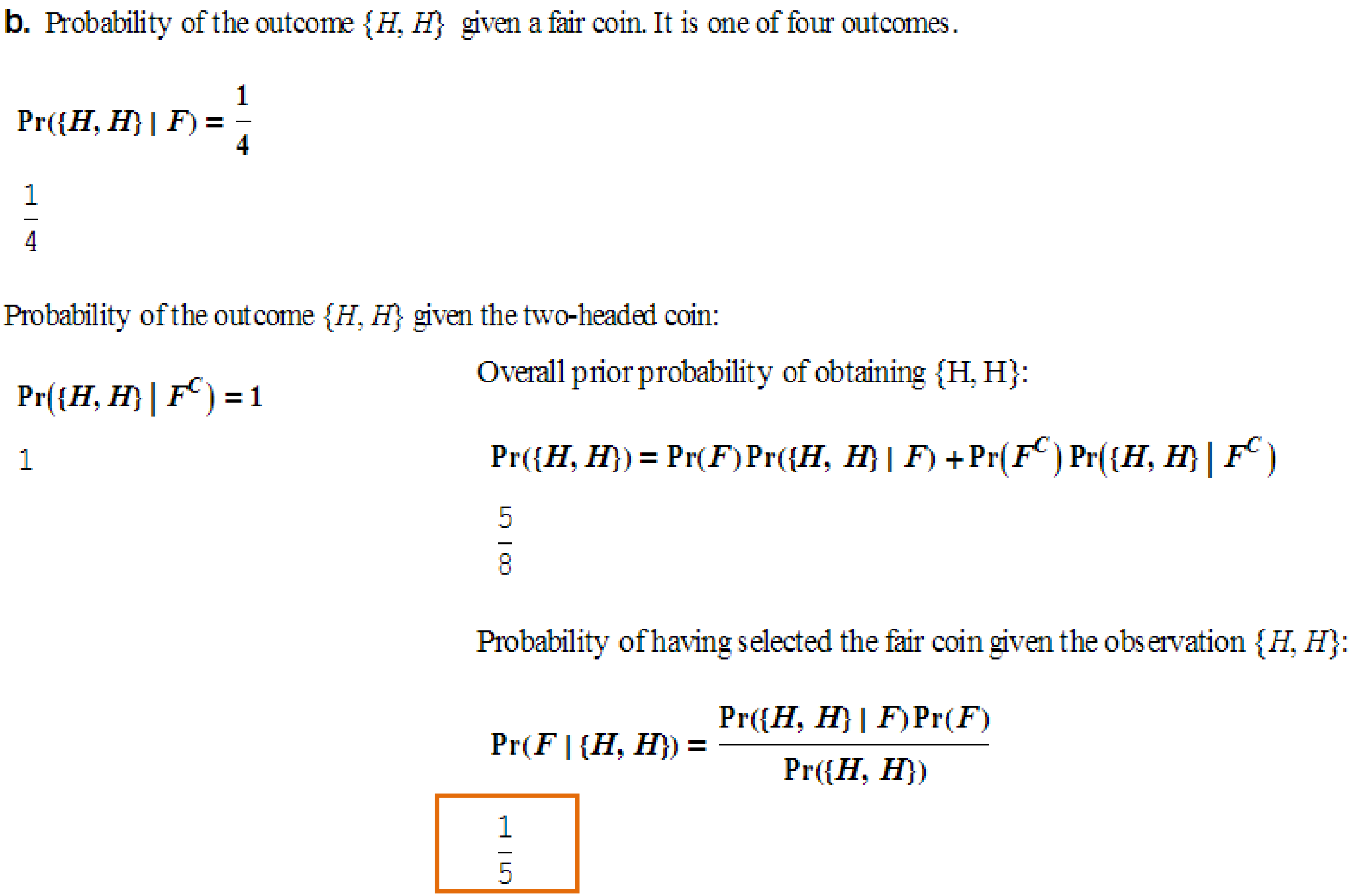
$$\frac{3}{4}$$

Probability of having picked the fair coin given that we flipped heads:

$$\Pr(F \mid H) = \frac{\Pr(H \mid F)\Pr(F)}{\Pr(H)}$$

$$\boxed{\frac{1}{3}}$$

b. He now flips the same coin a second time, and it again comes up heads. What is the probability that it is the fair coin?



b. Probability of the outcome $\{H, H\}$ given a fair coin. It is one of four outcomes.

$$\Pr(\{H, H\} \mid F) = \frac{1}{4}$$

$$\frac{1}{4}$$

Probability of the outcome $\{H, H\}$ given the two-headed coin:

$$\Pr(\{H, H\} \mid F^C) = 1$$

$$1$$

Overall prior probability of obtaining {H, H}:

$$\Pr(\{H, H\}) = \Pr(F)\Pr(\{H, H\} \mid F) + \Pr(F^C)\Pr(\{H, H\} \mid F^C)$$

$$\frac{5}{8}$$

Probability of having selected the fair coin given the observation $\{H, H\}$:

$$\Pr(F \mid \{H, H\}) = \frac{\Pr(\{H, H\} \mid F)\Pr(F)}{\Pr(\{H, H\})}$$

$$\boxed{\frac{1}{5}}$$

c. Suppose he flips the coin a third time, and it comes up tails. What is the probability that it is the fair coin?

- 1.0
- The two-headed coin cannot come up ‘tails’

example #2

- There is a prize behind either door ‘*a*’, door ‘*b*’, or door ‘*c*’
- You choose door ‘*a*.’ The host, who knows where the prize is, reveals that door ‘*b*’ does not have the prize, and asks if you want to switch
- Should you switch your choice?

- Original chance of choosing the prize: $\frac{1}{3}$
- Event $BB = \{\text{ door 'b' is revealed }\}$
- Random variable $ZZ = \{\text{ door with the prize }\}$

- $PP(ZZ = aa|BB) = \dfrac{PP(BB|ZZ = aa)PP(ZZ=aa)}{PP(BB)} = \dfrac{\frac{1}{2}\times\frac{1}{3}}{\frac{1}{2}} = \dfrac{1}{3}$

- $PP(ZZ = bb|BB) = \dfrac{PP(BB|ZZ = bb)PP(ZZ=bb)}{PP(BB)} = \dfrac{0\times\frac{1}{3}}{\frac{1}{2}} = 0$

- $PP(ZZ = cc|BB) = \dfrac{PP(BB|ZZ = cc)PP(ZZ=cc)}{PP(BB)} = \dfrac{1\times\frac{1}{3}}{\frac{1}{2}} = \dfrac{2}{3}$

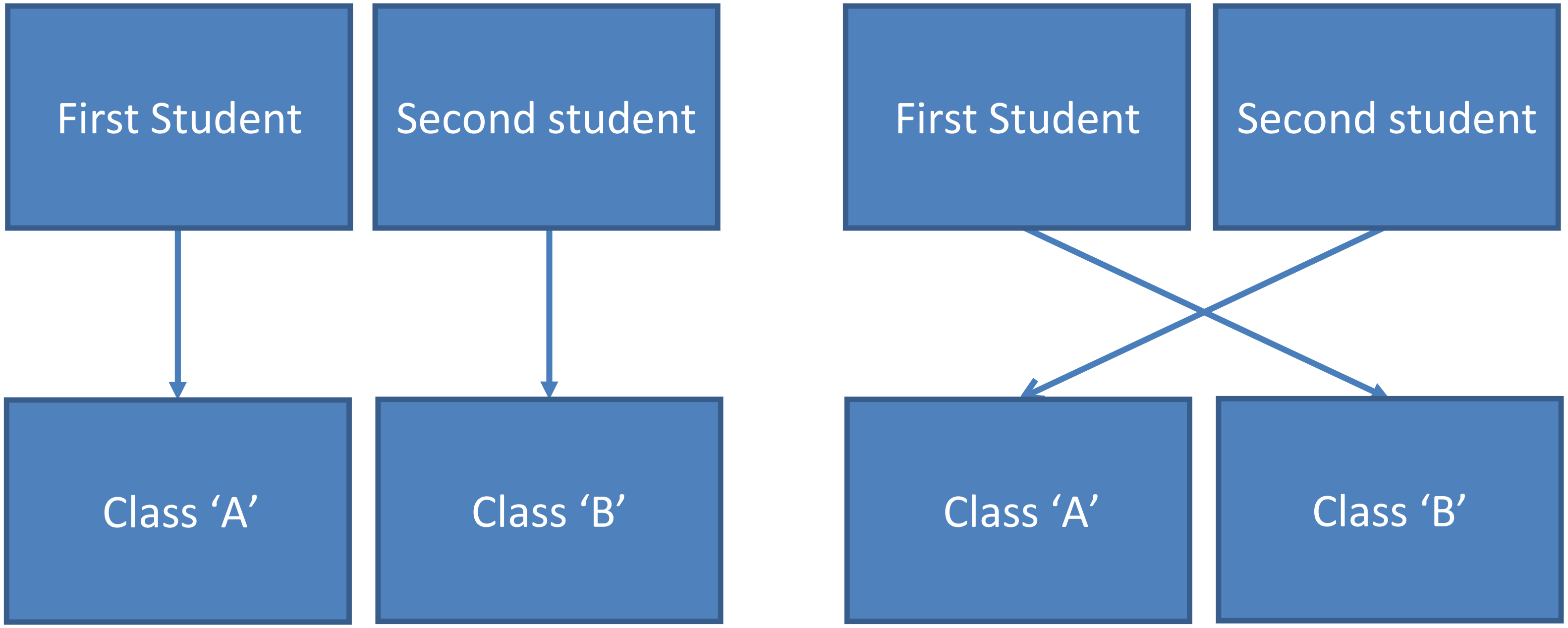
- Assuming the host always selects at random when he can (i.e. he has a choice), it is always better to switch your choice

example #3

- Class ‘*A*’ has 15 good students, 10 fair students, and 5 poor students. Class ‘*B*’ has 5 good students, 10 fair students, and 15 poor students.
- One student from each class is selected at random. They arrive for testing in random order. The first is fair and the second is poor. What is the probability that the first student examined is from class ‘*A*’?

2 possibilities for the actual matchup

MM M�



- Prior probabilities:

$$PP(MM) = 0.5$$

$$PP(M\quad) = 1 - PP(MM) = 0.5$$

�

(either match-up is equally likely)

We observe the sequence: (f, p)

Probability of seeing this given MM:

# fair in ‘A’

poor in ‘B’

$$PP((\mathrm{f},\mathrm{p})\,|MM) = \frac{10}{30} \times \frac{15}{30} = \frac{1}{6}$$

Probability of seeing this given M�:

# fair in ‘B’

poor in ‘A’

$$PP((\mathrm{f},\mathrm{p})\,|M\ ) = \frac{10}{30} \times \frac{5}{30} = \frac{1}{18}$$

�

- Overall prior probability of seeing (f, p):

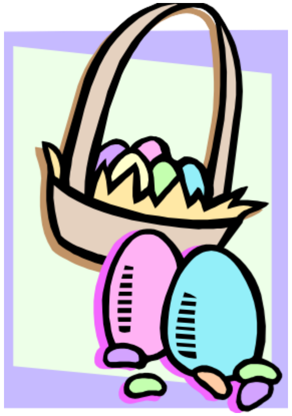
$$PP(MM) \times PP(\,(\text{f, p})\ |MM) + PP(\overset{M}{\text{�}}\ ) \times PP(\,(\text{f, p})\ |M\text{�})$$

$$= \frac{1}{2} \times \frac{1}{6} + \frac{1}{2} \times \frac{1}{18}$$

$$\text{P}((\text{f, p})) = \frac{1}{9}$$

example #4

A basket contains many small plastic eggs, some painted red and some are painted blue.



40% of the eggs in the bin contain pearls

30% of eggs containing pearls are painted blue, and 10% of eggs containing nothing are painted blue.

What is the probability that a blue egg contains a pearl?

$$PP(bbbbbbbb) = PP\ bb(bbbbbbb\ |ppbbaappbb) \times PP\ pp(bbaappbb) + PP\ bb(bbbbbbb\ |\overline{ppbbaappbb}) \times PP(\overline{ppbbaappbb})bb$$

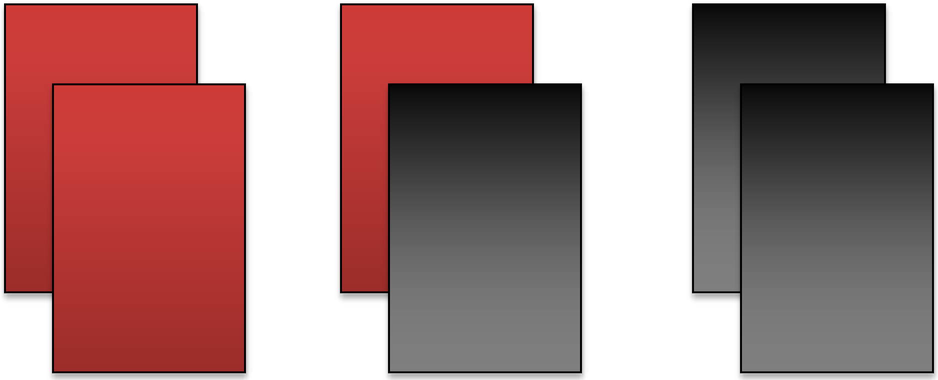
$$= .3 \times .4 + .1 \times .6$$

$$= .18$$

$$PP(ppbbaappbb|bbbbbbbb) = \frac{PP(bbbbbbbb|ppbbaappbb)PP(ppbbaappbb)}{PP(bbbbbbbb)}$$

Bayes theorem

- 3 cards:



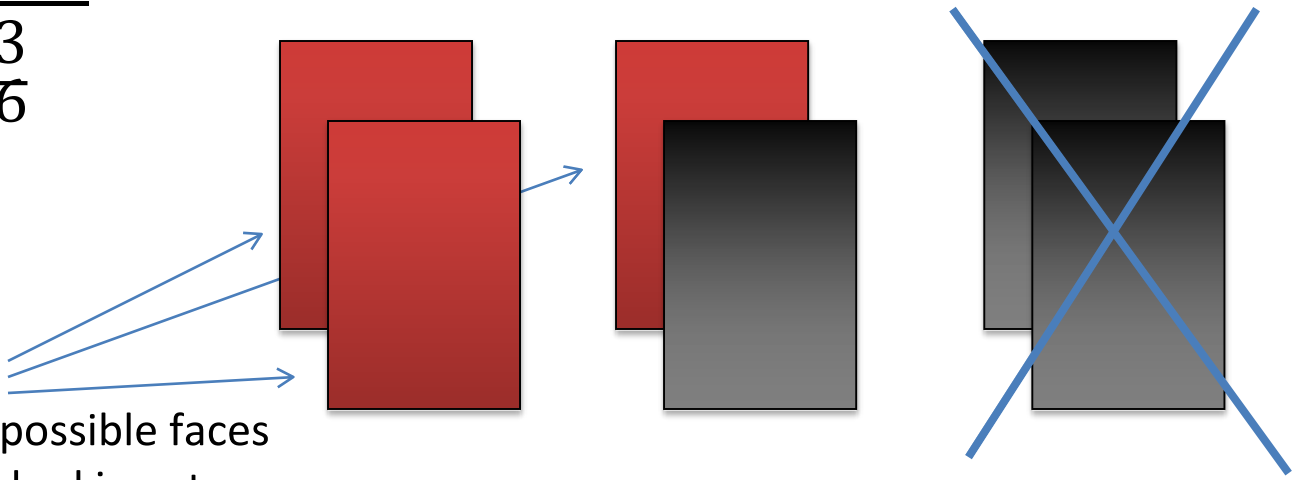
- We select a card at random and note that one side is red. What is the chance that it's the red-red card?

Bayes theorem

$$PP(\,(ppbbrr, ppbbrr)|RR) = \frac{PP(RR|(ppbbrr, ppbbrr)PP(ppbbrr, ppbbrr)}{PP(RR)}$$

$$= \frac{1 \times \frac{1}{3}}{\frac{3}{6}}$$

$$= \frac{2}{3}$$



For each of the 3 possible faces that you could be looking at, how many of them have red on its *other* side?

From Theory to Practice

- If we find a theoretical probability space that seems to correspond with some real-world phenomenon, this might help us predict something about the future occurrence of that phenomenon

  …or better understand past occurrences

- We want to do this carefully, so that we preserve as much of probability's mathematical soundness as we can
- It so happens that probability spaces *are* useful for characterizing real-world phenomena

Stochastic Trials

- Real-world events tend to have some degree of indeterminacy
  - In fact, no measuring apparatus can guarantee completely deterministic results (Heisenberg 1927, but disputed by Karl Popper)
  - Let’s proceed anyway
- A “measurable” real-world processes is called a stochastic process
- A measurement, or observational trial, of a stochastic process is a stochastic trial

You may also encounter the alternate terms random process, or random trial, but I avoid these terms because they seem to cloud the idea that we’re ultimately seeking to *characterize* the process, and that it is *characterizable*

Random Variables

Random Variables are the bridge between probability and statistics

Probability:

- Theoretically precise
- All outcomes are accounted for
- All outcomes are considered "at once"
- There is no 'trial'

random variable

X

event

A

Statistics:

- Empirical application
- We assume that a well-formed probability space applies to a real-world phenomenon
- We predict future outcomes

They help us equate a theoretical sample space with the measurable space of a stochastic process

Random Variables

- In order to:
  - generalize events;
  - allow for the variability of stochastic trials; and
  - map outcomes to empirical measurement values

  …we use random variables

A random variable is a function that maps a probability space Ω to the set of real numbers ℝ

$$X : \Omega \rightarrow$$

This is the formal definition. Like with Events, it's often clearer to define random variables using textual descriptions of outcomes

Random Variables

- Random variables map every possible outcome in a sample space to a scalar (1-dimensional) value
  - Like events, they will also be notated with upper case, italic, capital letters
  - For the following sample space:

$$\Omega = \{ aa, bb, cc, rr, bb, \ldots zz\}$$

we can define the random variable

$$WW = tt\boldsymbol{t}bb\ nnbbnnbbbbpp\ oooo\ ttttnnbbtt\ tt\boldsymbol{t}bb\ bbbbttttbbpp\ aappppbbaapptt\ ttnn\ tt\boldsymbol{t}bb\ rrooccbbnnbbnntt$$

We'll see more about defining random variables in a moment

- Do not confuse them with events. Remember, an event partitions the sample space into subsets *E* and E^C

$$"tt\boldsymbol{t}bb\ bbbbttttbbpp\ tttt\ aa\ vvoovvbbbb"$$

$$EE = \{ aa, bb, tt, oo, bb \}$$

"yes" or "no"

Discrete v. Continuous

- Like events, random variables can becontinuous or discrete
  - Discrete random variables assume a finite set of values
  - If the random variable models a count that isn't necessarily bounded ("countably infinite"), it is still considered discrete
- In Lecture 3, we mentioned continuous sample spaces but didn't discuss events (or their probabilities) in such spaces
  - That's because the probability of any particular event in a continuous sample space is zero.

    Ω = { *the mass of an orange* }

    E = { *the mass of the orange is 500g* }

    $P(E) = 0.0$

  - Continuous random variables will help here

Discrete v. continuous

X = { *the number of miles (to the nearest mile) a commuter drives to work* }

X is a discrete random variable

X = { *the distance a commuter drives to work* }

X is a continuous random variable

X = { *the commuter drives to work* }

X is an event

Discrete random variable examples

- It’s a good thing we studied counting and combinatorics last week, because discrete random variables in computational linguistics are often count data
  - the number of times a noun follows a determiner in a document
  - the number of bytes downloaded from a URL
  - the number of language informants in a study whose idiolect exhibits a certain phonological feature
  - the number of times a certain pair of words occur together in a corpus (number of collocations)

ⓘ For Project 2, you will tally (count) words in a corpus

Continuous random variable examples

- In computational linguistics, continuous random variables often have to do with timing and pitch frequency values in the speech-related subfields
  - duration of a phonological segment in a speech corpus
  - discourse particle usage interval timing in a sample of recorded discourse
  - F1 (1st formant) value in phonetic analysis
  - average frequency in voice recognition

Defining discrete random variables

- Sometimes, outcomes in the sample space naturally suggest certain numeric values

    example: a random variable for the result of rolling a six-sided die

defining a
discret
e
random

$$XX = \begin{cases} 1, \mathit{ttoo\ tt}\boldsymbol{t}\mathit{bb\ rrttbb\ tt}\boldsymbol{t}\mathit{oovvtt}\ \mathrm{Q}\ , \\ 2, \mathit{ttoo\ tt}\boldsymbol{t}\mathit{bb\ rrttbb\ tt}\boldsymbol{t}\mathit{oovvtt}\ \mathrm{R}\ , \\ 3, \mathit{ttoo\ tt}\boldsymbol{t}\mathit{bb\ rrttbb\ tt}\boldsymbol{t}\mathit{oovvtt}\ \mathrm{S}\ , \\ 4, \mathit{ttoo\ tt}\boldsymbol{t}\mathit{bb\ rrttbb\ tt}\boldsymbol{t}\mathit{oovvtt}\ \mathrm{T}\ , \\ 5, \mathit{ttoo\ tt}\boldsymbol{t}\mathit{bb\ rrttbb\ tt}\boldsymbol{t}\mathit{oovvtt}\ \mathrm{U}\ , \\ 6, \mathit{ttoo\ tt}\boldsymbol{t}\mathit{bb\ rrttbb\ tt}\boldsymbol{t}\mathit{oovvtt}\ \mathrm{V}\ . \end{cases}$$

Using text and pictures here (as opposed to numbers) emphasizes that the random variable is defined in terms of the actual real-world occurrence

Defining a discrete random variable

- The easiest way to define a discrete random variable is to use the numeric value that it measures:

X = { *the value of the roll of a single die* }

This establishes a mapping between an event and its conventional interpretation:

Q →1, R →2, S →3, T →4, U →5, V →6

- This can also be written in a more detailed notation…

Defining discrete random variables

$$XX = \begin{cases} 1, ttoo\ tt\boldsymbol{t}bb\ rrttbb\ tttoovvtt\ \mathrm{Q}\ , \\ 2, ttoo\ tt\boldsymbol{t}bb\ rrttbb\ tttoovvtt\ \mathrm{R}\ \ , \\ 3, ttoo\ tt\boldsymbol{t}bb\ rrttbb\ tttoovvtt\ \mathrm{S}\ \ , \\ 4, ttoo\ tt\boldsymbol{t}bb\ rrttbb\ tttoovvtt\ \mathrm{T}\ \ , \\ 5, ttoo\ tt\boldsymbol{t}bb\ rrttbb\ tttoovvtt\ \mathrm{U}\ , \\ 6, ttoo\ tt\boldsymbol{t}bb\ rrttbb\ tttoovvtt\ \mathrm{V}\ \ . \end{cases}$$

- Here, the random variable is defined in terms of six mutually exclusive events

Q: do they have to be collectively exhaustive?

A: Yes, a discrete random variable must define a value for every possible outcome

…and there’s no obvious way to stop a die from coming up with some value(s), unless we define our experiment as discarding certain trials

- Q: What about this: $XX = \begin{cases} 88.8, \text{ttoo tt}\boldsymbol{t}\text{bb rrttbb tt}\boldsymbol{t}\text{oovvtt Q} , \\ -2, \text{ttoo tt}\boldsymbol{t}\text{bb rrttbb tt}\boldsymbol{t}\text{oovvtt R} , \\ 123, \text{ttoo tt}\boldsymbol{t}\text{bb rrttbb tt}\boldsymbol{t}\text{oovvtt S} , \\ 0, \text{ttoo tt}\boldsymbol{t}\text{bb rrttbb tt}\boldsymbol{t}\text{oovvtt T} \quad \text{oopp U} , \\ 6.02 \times 10^{23}, \text{oott}\boldsymbol{t}\text{bbppvvttttbb}. \end{cases}$?

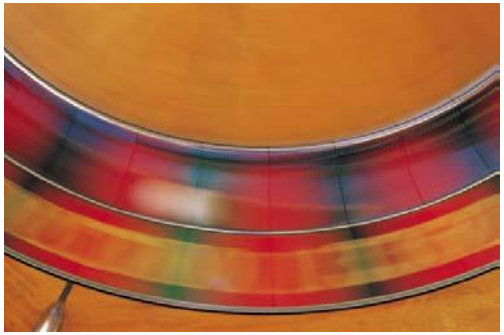
- A: Ok. The values of a discrete random variable are arbitrary.

As we'll see, we are really only interested in the random variable's probabilities, which are defined *in terms of* its values. So the values a random variable takes on only serve to establish *correspondence* between some event and the probability of that event

Defining discrete random variables

- We can use text labels
  - This random variable captures whether a roulette wheel comes up red or black

$$WW = \text{�} \begin{matrix} ppbbrr \\ bbbbaaccbb \end{matrix}$$



  - This random variable will be used to model the traditional gender categories

$$XX = \text{�} \begin{matrix} nnaabbbb \\ oobbnnaabbbb \end{matrix}$$

Defining continuous random variables

- For continuous random variables, we obviously cannot list a value for every point in the range
- Usually, the variable is defined as the real-valued data observation

X = { *rime duration* /bay/ *time (ms.) in the test population* }

Continuous random variables can also be defined according to a continuous function, but this is less useful for modeling *measurements*

This is the reason why it often makes sense to define the RV with a text (prose) description.

Events v. Random Variables

- An event is a single outcome, or some subset of outcomes, from Ω

$$"tt\boldsymbol{t}bb\ ttoottaabb\ tt\boldsymbol{t}oovvttnnss\ oonn\ tt\boldsymbol{t}bb\ ttvvoo\ rrttccbb\ tttt\ ttbbvvbbnn"$$

$$EE = \{(\mathrm{Q}\ ,\mathrm{V}\ ), (\mathrm{R}\ \ ,\mathrm{U}), (\mathrm{S}\ \ ,\mathrm{T}\ \ ), (\mathrm{T}\ \ ,\mathrm{S}\ \ ), (\mathrm{U}\ \ ,\mathrm{R}\ ), (\mathrm{V}\ \ ,\mathrm{Q})\}$$

- A random variable is a function that maps any possible outcome to a real number (or quantifiable label)

$$XX = FF\left(tt\boldsymbol{t}bb\ ttoottaabb\ tt\boldsymbol{t}oovvttnnss\ oonn\ tt\boldsymbol{t}bb\ ttvvoo\ rrttccbb\ ttnn\ aa\ ttppttaabb\ \right)$$

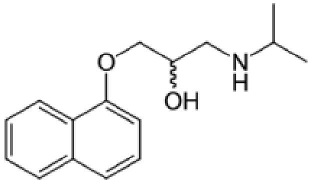
$$= FF\big(\big(\ \mathrm{P}\ \ + \mathrm{P}\ \ \big)\big)$$

$$xx \in \{\ 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\ \}$$

will be explained shortly. It represents the value of one trial for the

Event or R.V.?

- Average F1 formant value in the recording
- The F1 value is above 600MHz
- The coin shows ‘heads’ ten times in a row
- “Call me Ishmael” are the first three words in the book
- The number of words before the word “Ishmael” in the book
- The number of clinical trial studies not discussing $\beta\beta$-adrenergic blocking agents



A random variable is just a mapping function

- You may have noticed that, by itself, a random variable is not too useful
- A random variable is just a function that maps an outcome to some real number:

  It does not say anything about the likelihood of getting a particular value

- In Lecture 3, we used counting to get the probability for each outcome from the sample space, according to

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

- For discrete spaces, we could calculate the probability of a random variable in the same way

Probability and random variables

- However, random variables give us a much more powerful ways of working
- Like we did with events, we introduce the notion of a probability distribution

A probability distribution is a function that maps all possible values (for discrete random variables) or ranges of values (for continuous random variables) of a random variable into a well-formed ("proper") probability space

A probability distribution represents a generalization over multiple trials

Probability distributions

The probability that the discrete random variable X will have the value x is notated by

$$PP(XX = xx)$$

Alternate notation: $\rho\rho_{XX}(xx)$

This is the probability mass function (pmf) of *X*

The probability that the continuous random variable X will have a value between aa and bb is notated by

$$PP(aa \le XX \le bb) = \underset{aa}{\overset{bb}{\text{�}}}\, oo_{XX}(xx)\, rrxx$$

The function $oo_{XX}(xx)$ is the probability density function (pdf) of X

Notating discrete random variable probabilities

- For discrete random variables, we explicitly list the probability for each possible value

use the lower case letter of the random variable to signify a single trial when defining discrete random variable

$$PP(XX = xx) = \begin{cases} 0.1667, ttoo\ xx = 1; \\ 0.1667, ttoo\ xx = 2; \\ 0.1667, ttoo\ xx = 3; \\ 0.1667, ttoo\ xx = 4; \\ 0.1667, ttoo\ xx = 5; \\ 0.1667, ttoo\ xx = 6; \\ 0, ootttbbppvvttttbb. \end{cases}$$

the last line can be left off

- This is the correct form for expressing the probability distribution of a discrete random variable—its probability mass function
- This shows why the *values* for x are arbitrary. They can be chosen to most conveniently match empirical measurement

Defining a random variable ⚠ Defining the probability of a random variable

$$XX = \begin{cases} 1, \mathit{ttoo\ tt}\boldsymbol{t}\mathit{bb\ rrttbb\ tt}\boldsymbol{t}\mathit{oovvtt}\ \text{Q}\ , \\ 2, \mathit{ttoo\ tt}\boldsymbol{t}\mathit{bb\ rrttbb\ tt}\boldsymbol{t}\mathit{oovvtt}\ \text{R} \\ , \\ 3, \mathit{ttoo\ tt}\boldsymbol{t}\mathit{bb\ rrttbb\ tt}\boldsymbol{t}\mathit{oovvtt}\ \text{S}\ , \\ 4, \mathit{ttoo\ tt}\boldsymbol{t}\mathit{bb\ rrttbb\ tt}\boldsymbol{t}\mathit{oovvtt}\ \text{T}\ , \\ 5, \mathit{ttoo\ tt}\boldsymbol{t}\mathit{bb\ rrttbb\ tt}\boldsymbol{t}\mathit{oovvtt}\ \text{U}\ , \end{cases}$$

$$6, \mathit{ttoo\ tt}\boldsymbol{t}\mathit{bb\ rrttbb\ tt}\boldsymbol{t}\mathit{oovvtt}\ \text{V}$$

.

$$PP(XX = xx) = \begin{cases} 0.1667, \mathit{ttoo}\ xx = 1; \\ 0.1667, \mathit{ttoo}\ xx = 2; \\ 0.1667, \mathit{ttoo}\ xx = 3; \\ 0.1667, \mathit{ttoo}\ xx = 4; \\ 0.1667, \mathit{ttoo}\ xx = 5; \\ 0.1667, \mathit{ttoo}\ xx = 6; \\ 0, \mathit{oott}\boldsymbol{t}\mathit{bbppvvttttbb}. \end{cases}$$

Each value of a discrete random variable basically acts as an index for matching an *event* to a *probability value*.

Probability mass function

- The probability $PP(XX = xx)$ of a discrete random variable is called its probability mass function (pmf)

- We can plot it:

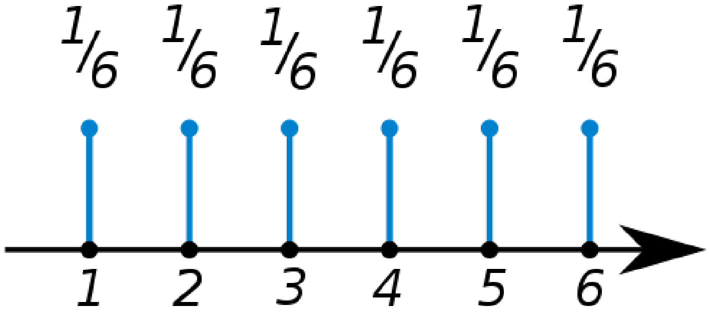


image: wikipedia

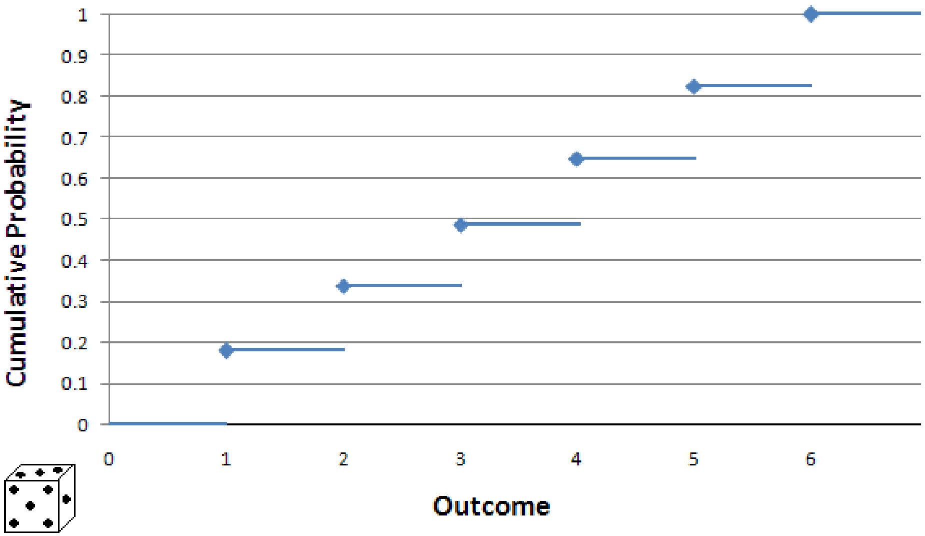
- This doesn’t work for continuous random variable YY because $\forall\, xx, PP\,(XX = xx) = 0$ (see later slides)

Cumulative distribution function

- The cumulative distribution function (cdf) shows the accumulated mass of the pmf

$$PP(XX \leq xx)$$

- For a discrete random variable, this will be a step function



Probability density function

- We need to use some calculus to describe continuous random variable XX. The first step is to find a cumulative distribution function

$$PP(XX \leq xx) = \overset{xx}{\underset{-\infty}{\text{�}}}\, 00_{XX}(bb\,)rrbb$$

Then, the derivative of this, $00_{XX}(xx)$ is the pdf of the continuous random variable

$$\text{pdf}_{XX} = 00_{XX}(xx) = \frac{rr \int_{-\infty}^{xx} 00_{XX}(bb)\,rrbb}{rrxx}$$

pdf of a continuous random variable

$$\mathrm{pdf}_{XX} = oo_{XX}(xx) = \frac{rr \int_{-\infty}^{xx} oo_{XX}(bb)\, rrbb}{rrxx}$$

We can not use $PP(XX = xx)$ for the left side of this, because that is zero

Following the convention of most texts, we use $oo_{XX}(x)$ for the pdf*

We will walk through an example in a moment

*Usually, the subscript which indicates the r. v. that the function applies to are *dropped*, even working with many different random variables. This can be confusing. For a discussion of these and other notation issues, see http://lingpipe-blog.com/2009/10/13/whats-wrong-with-probability-notation/

So where do probabilities come from?

- They have meaning only by construction
- For discrete event *E*, we conjured probability $PP(EE)$
- For discrete random variable *X*, we conjured probability

  $$PP(XX = xx)$$

- For continuous random variable *X*, we conjured cumulative distribution

  $$PP(XX \leq xx) = \overset{xx}{\underset{-\infty}{\text{�}}} oo_{XX}(bb) rrbb$$

So where do probabilities come from?

Because a random variable—like an event—encapsulates all possible outcomes in a sample space, its probability function *meaningfully* characterizes that sample space

This licenses us to use the random variable as an estimate—or proxy—for the entire sample space

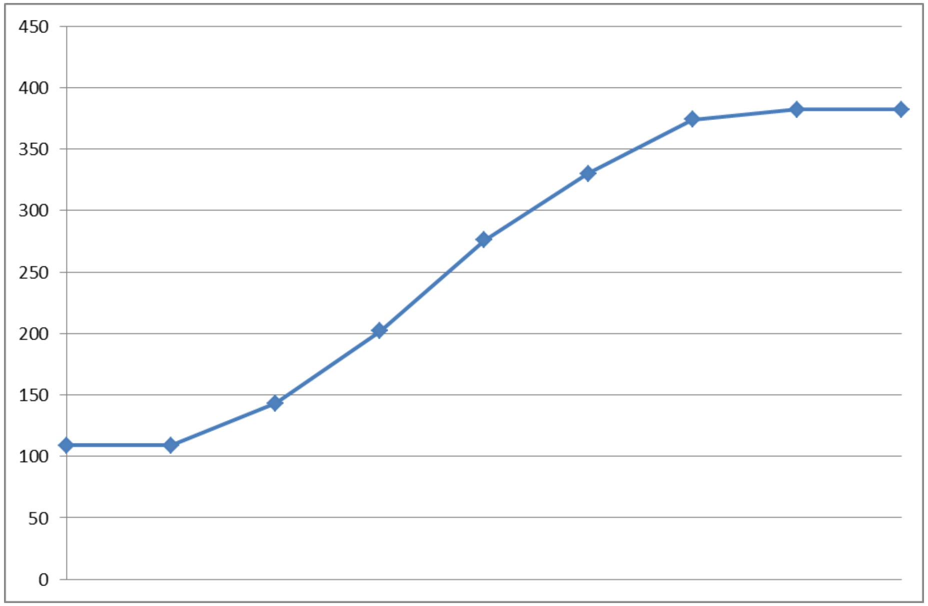
And, conversely, *use observed probabilities* to define the random variable. Let’s study a continuous random variable as an example.

Continuous random variable: example

X = { *rime duration* /bay/ *time (ms.) in the test population* }

raw data: {202, 374, 279, 330, 382, 140, 109 }

sorted: { 109, 140, 202, 279, 330, 374, 382 }



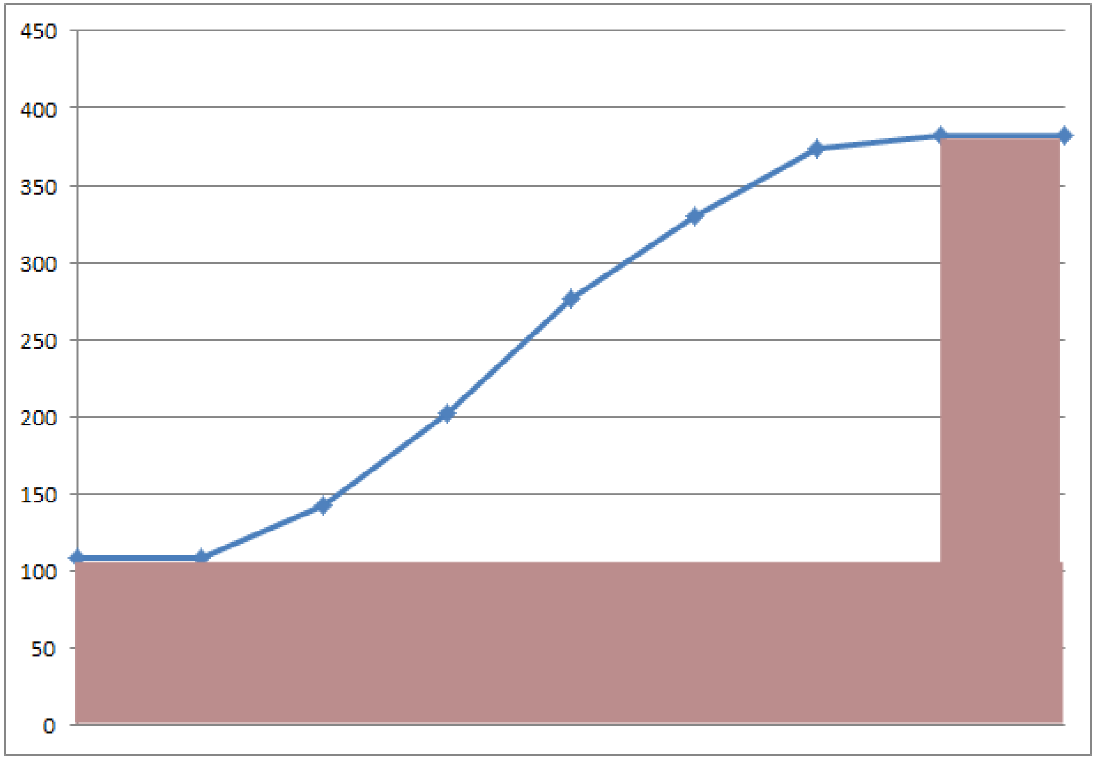
Here’s our raw data

They are equally spaced because each observation is equally important

We extend the first and last points horizontally to show that there are no further observations

Continuous random variable: example

Normalize the data by first throwing away the shaded part…



We are working towards making our data into a cdf

We must normalize the data to meet the following cdf criteria:

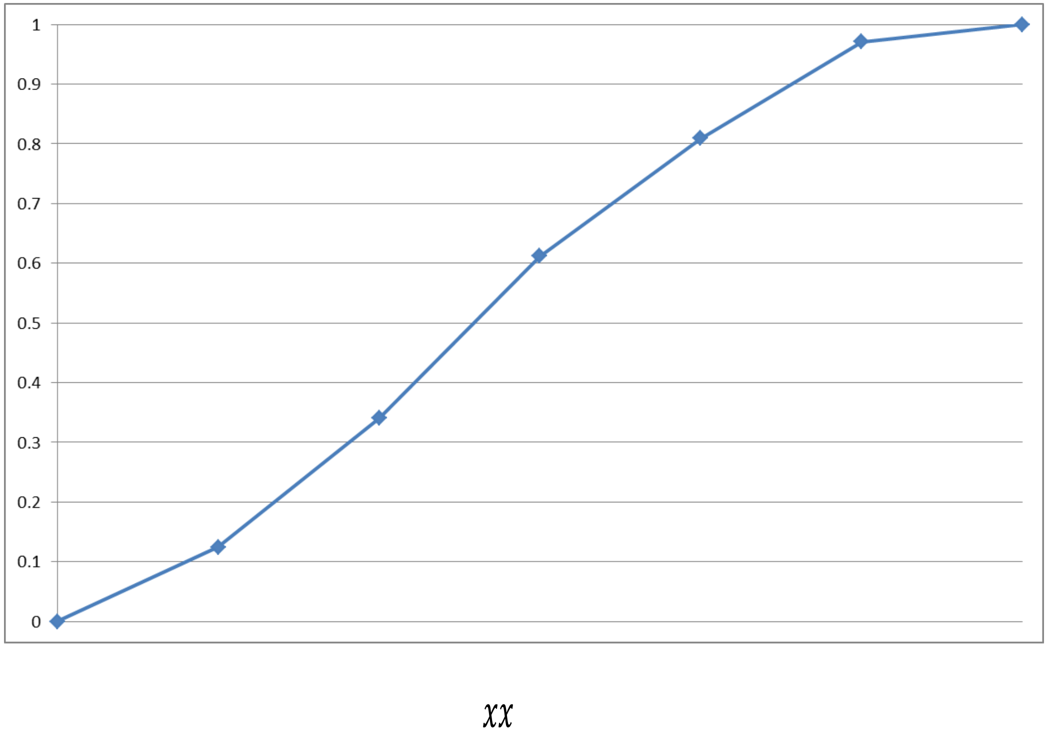
- It has the value 0 at $-\infty$
- It has the value 1 at $+\infty$
- It is monotonically non-decreasing

We also remember the normalization factors we used so that we can reverse this process

Continuous random variable: example

…and scaling the top value to 1

(this is the “by construction” part)

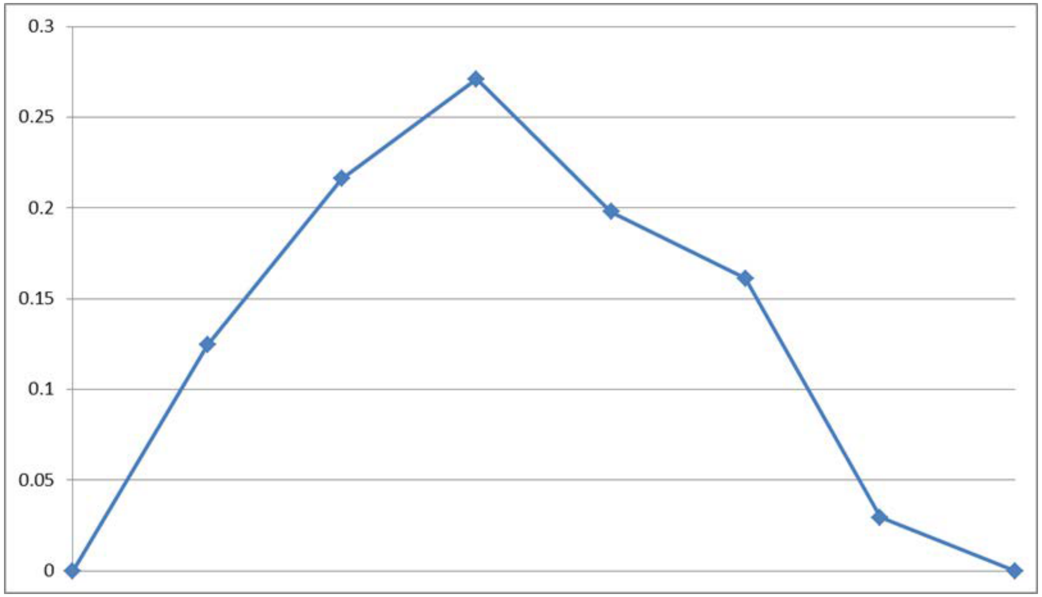


This now meets the definition of a cumulative density function

$$PP(XX \leq xx) = \underset{-\infty}{\overset{xx}{�}}\, oo_{XX}(bb)rrbb$$

Continuous random variable: example

Finally, take the derivative. This is the set of differences between adjacent points in the cdf



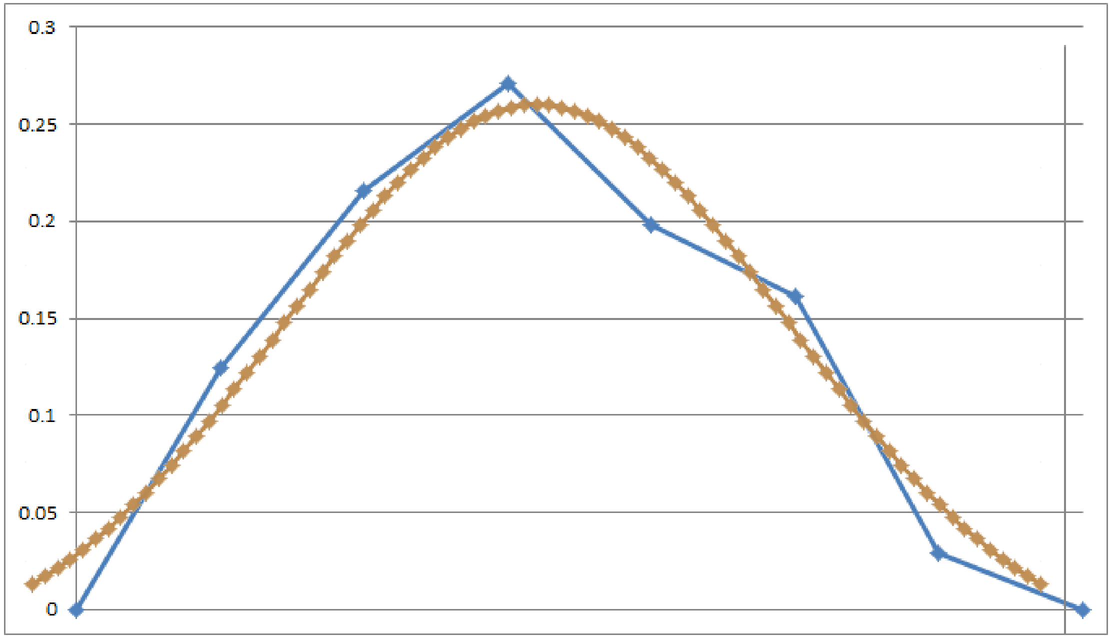
This is a probability density function (pdf) for our random variable

$$00_{XX}(xx) = \frac{rr\int_{-\infty}^{xx} 00_{XX}\left(bb\right)rrbb}{rrxx}$$

Because of the way we constructed it, the total area under this curve will equal 1.0.

Continuous random variable example

Even though we only had a few data points, it looks like our phonetics data fit the shape of a normal curve pretty closely



The normal curve is a type of probability distribution we'll be studying later

Summary of the example

- The example shows that we can take a few raw data points make a general statement about our data:

"Measurement of rime duration of the syllable /bay/ in the test population for speakers in our study is approximately normal."

- If we assume that this distribution *characterizes* our continuous random variable, then we can use this distribution to predict the studied feature in other speakers, or in the general population

Summary: pmf / pdf / cdf

The probability mass function (pmf) of a discrete random variable X is notated by

$PP_{XX}(XX = xx)$ alternate notation: $\rho\rho_{XX}(xx)$

The probability density function (pdf) of a continuous random variable X is notated by

00_{XX} ( ) xx

For either type, the cumulative distribution function (cdf) is notated by

$PP_{XX}(XX \leq xx)$ alternate notation: $FF_{XX}(xx)$

The subscripted random variable is usually omitted, so you have to remember that PP is a different function for each random variable that you're working with

Probability distributions

Assuming that a random variable exhibits a fixed, characteristic probability distribution, e.g.

$$\Omega = \{aa, bb, cc\}$$

$$PP(XX = xx) = \begin{cases} 1\text{�}_3, ttoo\ xx = \{aa\}; \\ 1\text{�}_3, ttoo\ xx = \{bb\}; \\ 1\text{�}_3, ttoo\ xx = \{cc\}; \end{cases}$$

allows us justify our intuition about events from last week:

$$AA = \{aa\}$$

$$AA^{CC} = \{\ bb, cc\}$$

$$PP(AA) = \frac{|AA|}{|\Omega|}$$

$$PP(AA^{CC}) = \frac{|AA^{CC}|}{|\Omega|}$$

$$PP(AA) + PP(AA^{CC}) = \frac{|AA|}{|\Omega|} + \frac{|AA^{CC}|}{|\Omega|} = \frac{|AA| + |AA^{CC}|}{|\Omega|} = \frac{|\Omega|}{|\Omega|} = 1$$

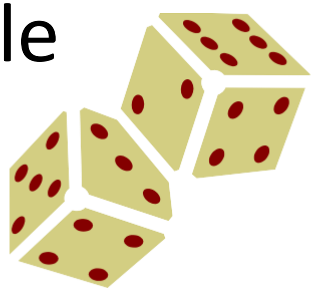
Probability distributions

- A random variable's probability distribution encapsulates both:
  - a characteristic type of "spread" or "shape" (distribution)
    - uniform
    - normal
    - etc.
  - the scaling and normalization factors that map between probabilities [0.0, 1.0] and the range of measurement values

This is why the capital letter subscript is (supposed to be) used: $PP_{XX}(XX = xx)$

Uniform distribution

- Dividing the probability mass evenly between the values of a discrete random variable creates a uniform distribution



$$aa = 1, bb = 6$$

the mean μ is the average value

$$\mu\mu = \frac{aa + bb}{2} = 3.5$$

Non-uniform distribution

(the, cat, in, the, hat)

$$XX = \{ \ tt\boldsymbol{t}bb \ vvooppr r \ \boldsymbol{vvt}ttcct \ tttt \ ttbbbbbbccttbbrr \ \}$$

$$PP_{XX}\,(XX = tt\boldsymbol{t}bb\,) = 0.4$$

$$PP_{XX}(XX = ccaatt) = 0.2$$

$$PP_{XX}(XX = ttnn) = 0.2$$

$$PP_{XX}\,(XX = \boldsymbol{t}aatt\,) = 0.2$$

$$YY = \{ \ tt\boldsymbol{t}bb \ nnbbnnbbbbpp \ oooo \ ttttnnbbtt \ \mathrm{X{=}the} \ ttnn \ 3 \ ttppttaabbtt, \ vvtttt\boldsymbol{t} \ ppbbppbbaaccbbnnbbnntt \ \}$$

$$PP_{YY}(YY = 0) = .6 \times .6 \times .6 = .216$$

$$PP_{YY}(YY = 3) = .4 \times .4 \times .4 = .064$$

$$PP_{YY}(YY = 1) = .4 \times .6 \times .6 \times \binom{3}{1} = .432$$

$$PP_{YY}(YY = 2) = .4 \times .4 \times .6 \times \binom{3}{1} = .288$$

$$PP_{YY}(YY \geq 2) = .352$$

Finite state machines

or, finite state automata

- Deterministic
- Non-deterministic

{ set of states, transitions, start state, input alphabet, final states }

- Finite state transducers
- Acceptor

Deterministic FSM

$qq \in SS$	States
$\delta\delta : SS \times \Sigma \to SS$	Transitions
$SS_0 \in SS$	Start state
$xx \in \Sigma$	Input alphabet
$FF \in SS$	Final states (or ∅)

Each state/input pair has no more than one transition

Non-deterministic FSM

$qq \in SS$	States
PP_{SS}	Transition probabilities
$\delta\delta : SS \times \Sigma \times PP_{SS} \rightarrow SS$	Transitions
$SS_0 \in SS$	Start state
$xx \in \Sigma$	Input alphabet
$FF \in SS$	Final states (or Ø)

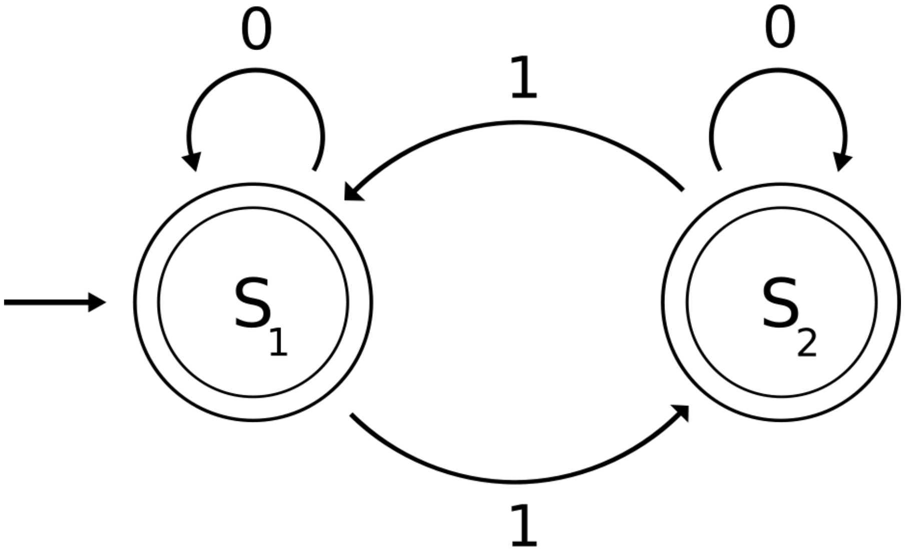
For a given state/input, there may be more than one possible transition

At runtime

- This is sufficient description of the machine. At runtime, an input stream composed of symbols from alphabet Σ is provided
- If $\delta\delta(qq,xx)$ is incomplete, the FSM is said to reject the input

Example

parity: the number of bits in a binary value that are 'set' to one (1)

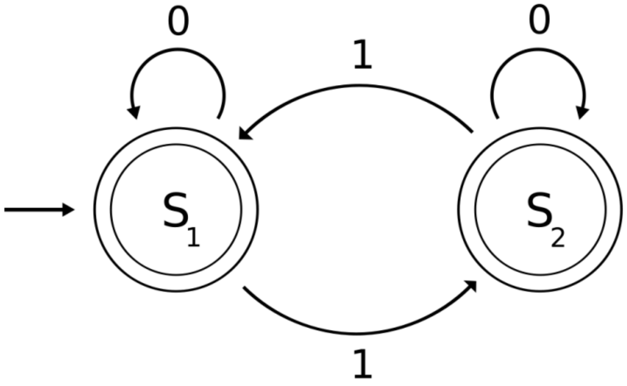


Double-circles are used to indicate accepting states

parity of the binary input:
S1 : even
S2 : odd

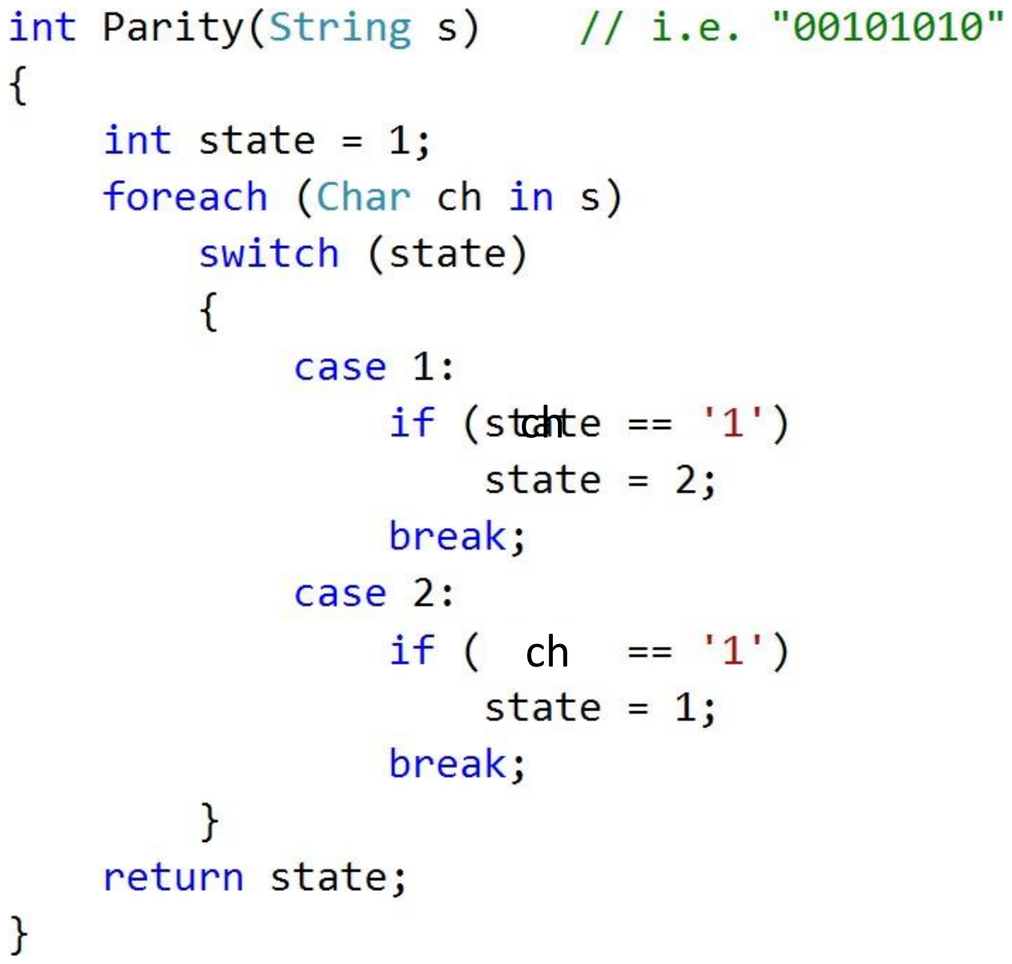
1 0 1 1 0 0 1 → S1
0 0 0 1 0 0 0 → S2

Example

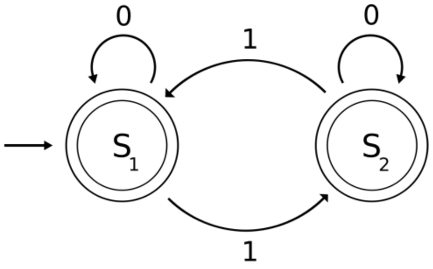


State	Transition
S1	0 → S1, 1 → S2
S2	0 → S2, 1 → S1

Programming FSTs



```
int Parity(String s)    // i.e. "00101010"
{
    int state = 1;
    foreach (Char ch in s)
        switch (state)
        {
            case 1:
                if (ch == '1')
                    state = 2;
                break;
            case 2:
                if ( ch  == '1')
                    state = 1;
                break;
        }
    return state;
}
```



Example

- Write an FSA for the RegEx:

  a[ab]*b[cd]

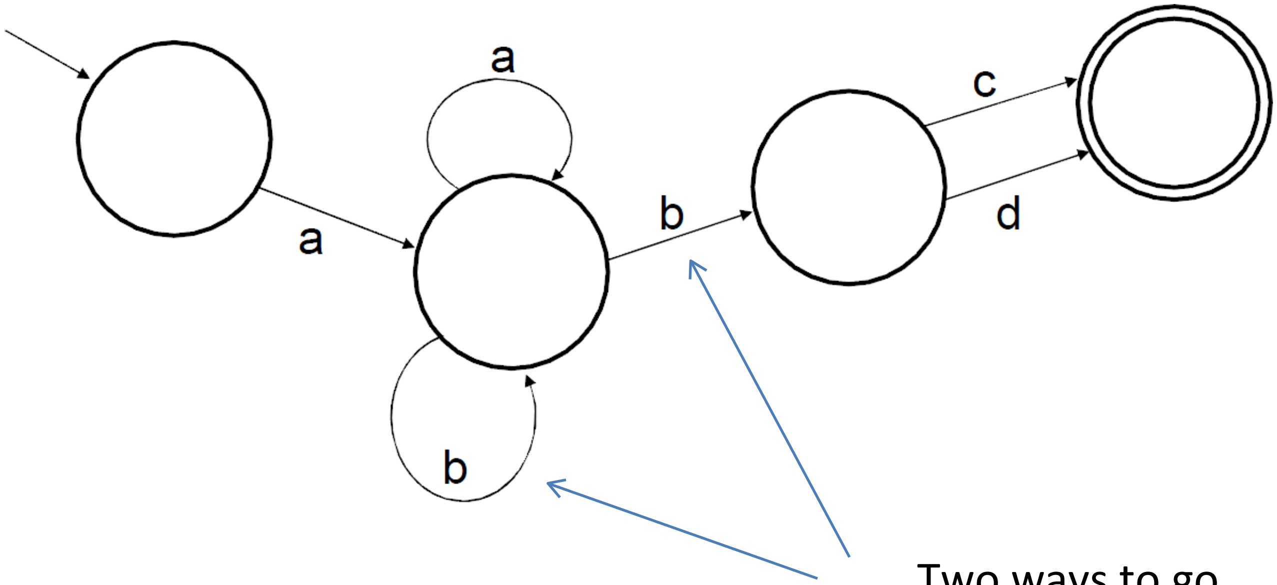
FSM example

Is your FSM deterministic or non-deterministic?

Example

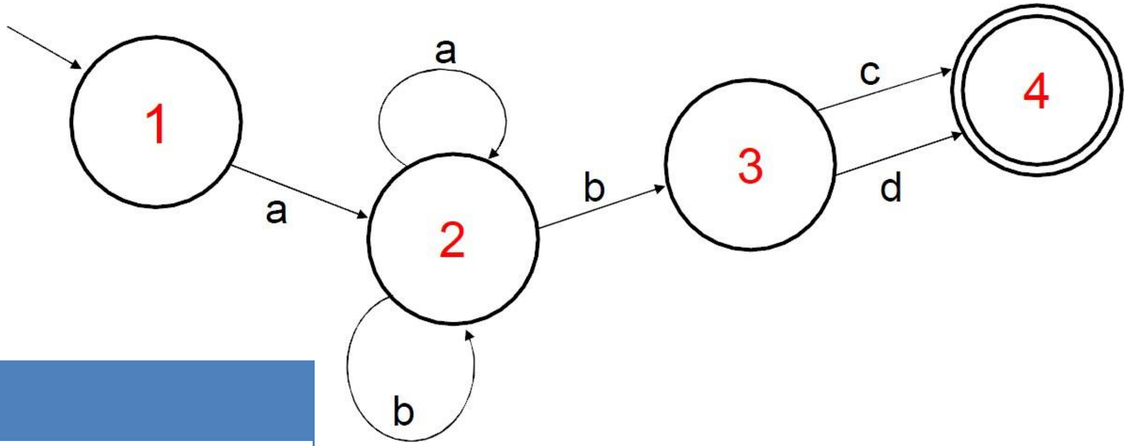
- Non-deterministic

  a[ab]*b[cd]



Two ways to go

a[ab]*b[cd]



State	Transition
1	a → 2
2	a → 2, b → 2, b → 3
3	c → 4, d → 4

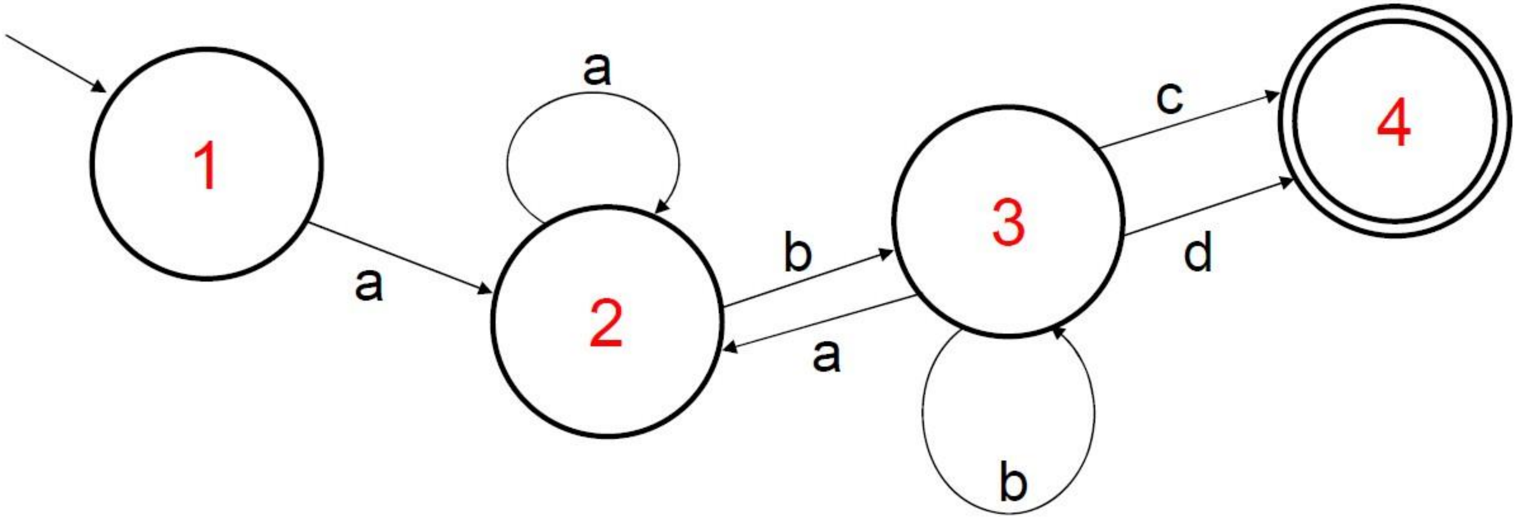
How would we implement this state machine? We would need more information on how to proceed out of state 2 when the input is ‘b’

Example: abbcd
if we choose state 3 here, we will fail to accept this pattern when we should have

Example

- deterministic

  a[ab]*b[cd]

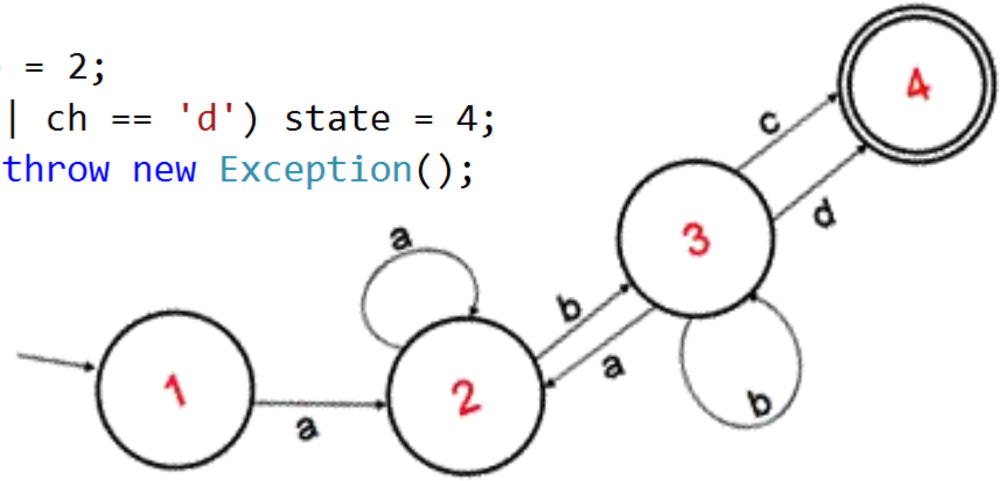


Finite state transducer (FST)

- Add an output function (per state) to an FSM
- The function fires upon arriving at a state
  - Or, when "transitioning"
- Output models:
  - Mealy model: the output depends on both the current input and the state
  - Moore model: the output depends only on the state

FST

```
IEnumerable<int> FST(String input) {
    int state = 1;
    foreach (Char ch in input) {
        switch (state) {
            case 1:
                if (ch == 'a') state = 2;
                else throw new Exception();
                break;
            case 2:
                if (ch == 'b') state = 3;
                else if (ch != 'a') throw new Exception();
                break;
            case 3:
                if (ch == 'a') state = 2;
                else if (ch == 'c' || ch == 'd') state = 4;
                else if (ch != 'b') throw new Exception();
                break;
            case 4:
                yield break;
        }

    }
}
```



Using C#

With your UW-NetID, you can download and install the full version of Microsoft Visual Studio 2013 Professional for free: http://www.dreamspark.com

To compile a C# program on patas:

```
/home2/joe-student$ gmcs project2-b.cs
```

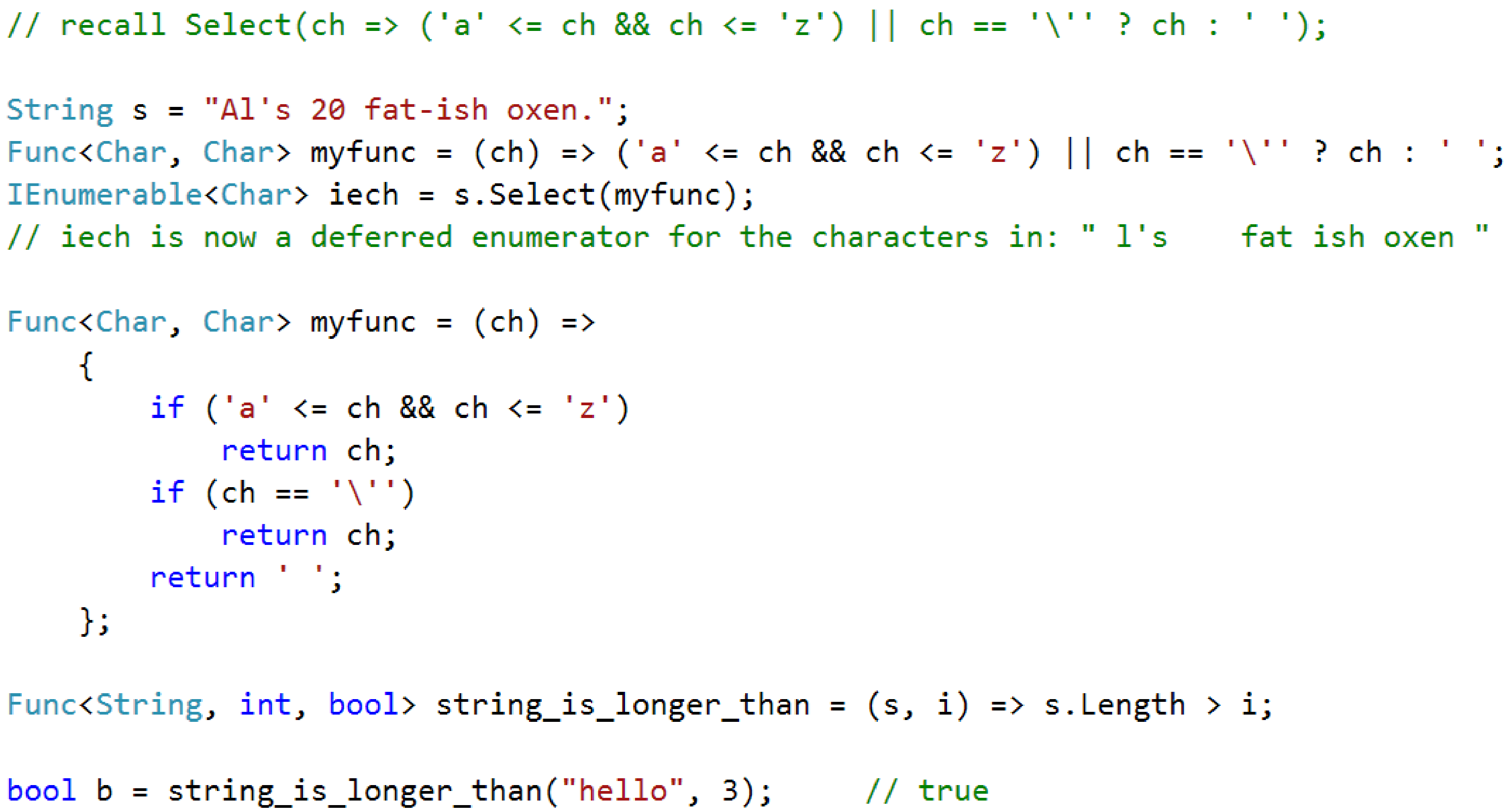
To run it:

```
/home2/joe-student$ mono project2.exe
```

Closures

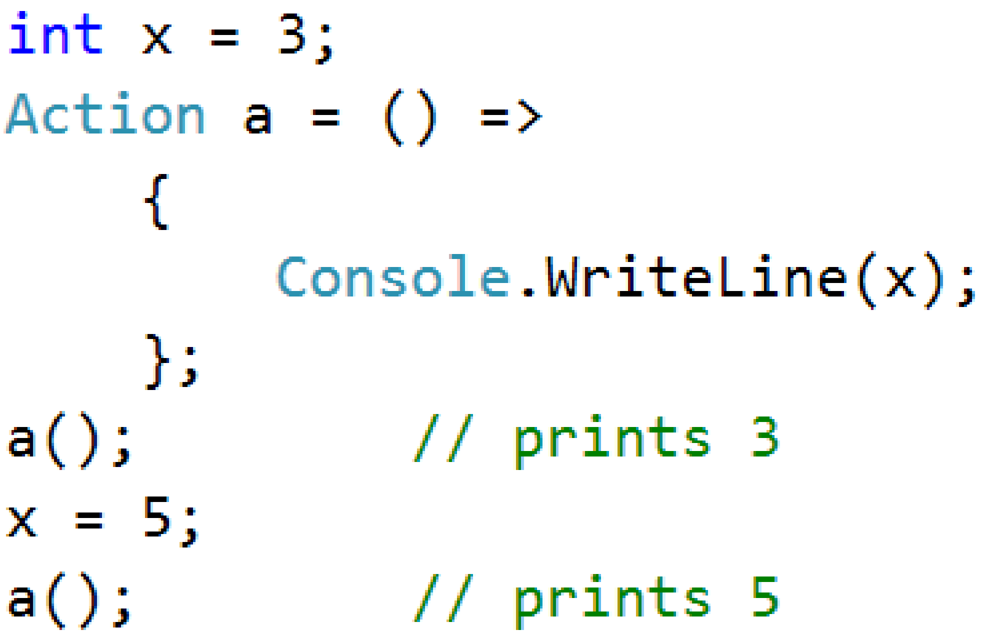
- Lambda expressions automatically capture local variables that they reference, which are then passed around as part of the lambda variable
  - Caution: languages do this differently with respect to reference (the lambda expression will modify the original value) versus value (the lambda expression has a snapshot of the value)
- This can lead to interesting scoping issues

Lambda expressions



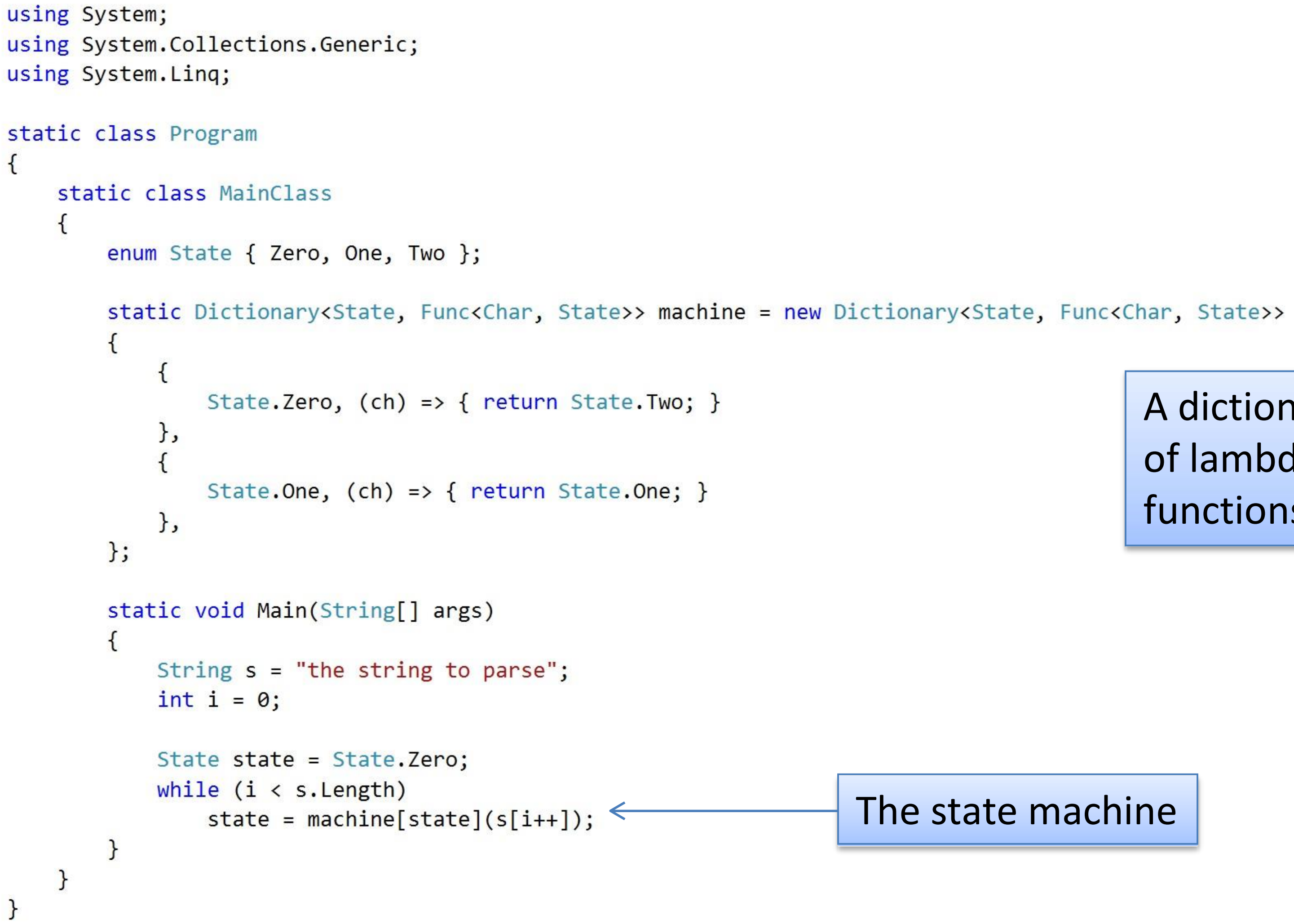
```
// recall Select(ch => ('a' <= ch && ch <= 'z') || ch == '\'' ? ch : ' ');

String s = "Al's 20 fat-ish oxen.";
Func<Char, Char> myfunc = (ch) => ('a' <= ch && ch <= 'z') || ch == '\'' ? ch : ' ';
IEnumerable<Char> iech = s.Select(myfunc);
// iech is now a deferred enumerator for the characters in: " l's    fat ish oxen "

Func<Char, Char> myfunc = (ch) =>
    {
        if ('a' <= ch && ch <= 'z')
            return ch;
        if (ch == '\'')
            return ch;
        return ' ';
    };

Func<String, int, bool> string_is_longer_than = (s, i) => s.Length > i;

bool b = string_is_longer_than("hello", 3);     // true
```

Closure example



```
int x = 3;
Action a = () =>
    {
        Console.WriteLine(x);
    };
a();        // prints 3
x = 5;
a();        // prints 5
```

State machine example



```
using System;
using System.Collections.Generic;
using System.Linq;

static class Program
{
    static class MainClass
    {
        enum State { Zero, One, Two };

        static Dictionary<State, Func<Char, State>> machine = new Dictionary<State, Func<Char, State>>
        {
            {
                State.Zero, (ch) => { return State.Two; }
            },
            {
                State.One, (ch) => { return State.One; }
            },
        };

        static void Main(String[] args)
        {
            String s = "the string to parse";
            int i = 0;

            State state = State.Zero;
            while (i < s.Length)
                state = machine[state](s[i++]);
        }
    }
}
```

A dictionary of lambda functions

The state machine

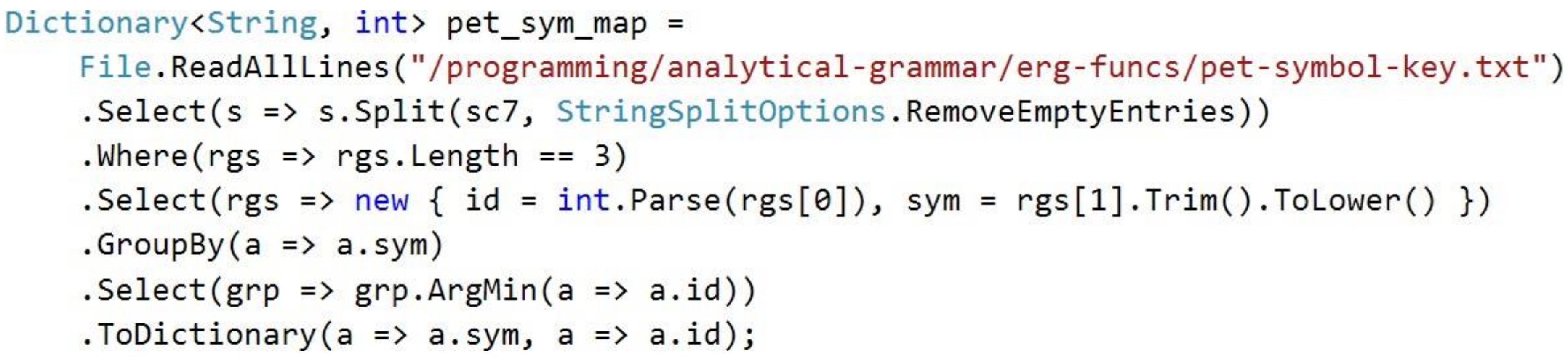
LINQ in C#

- Sequences: IEnumerable<T>
- Deferred execution
- Type inference
- Strong typing
  - despite the ‘var’ keyword
  - (C# 4.0 now has the `’dynamic’` keyword, which allows true runtime typing where desired)

LINQ operators

- Filter/Quantify:
  `Where, ElementAt, First, Last, OfType`
- Aggregate:
  `Count, Any, All, Sum, Min, Max`
- Partition/Concatenate:
  `Take, Skip, Concat`
- Project/Generate:
  `Select, SelectMany, Empty, Range, Repeat`
- Set:
  `Union, Intersect, Except, Distinct`
- Sort/Ordering:
  `OrderBy, ThenBy, OrderByDescending, Reverse`
- Convert/Render:
  `Cast, ToArray, ToList, ToDictionary`

Example

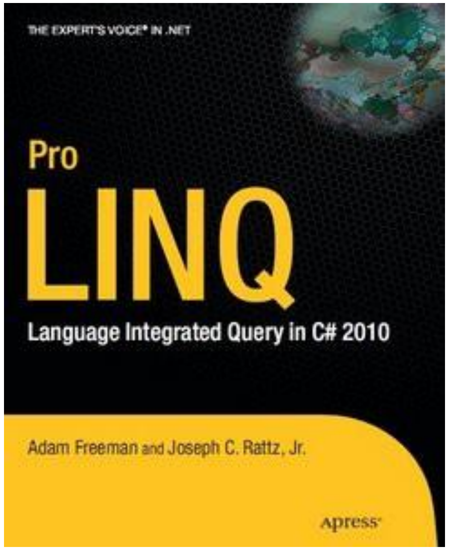


```
Dictionary<String, int> pet_sym_map =
    File.ReadAllLines("/programming/analytical-grammar/erg-funcs/pet-symbol-key.txt")
    .Select(s => s.Split(sc7, StringSplitOptions.RemoveEmptyEntries))
    .Where(rgs => rgs.Length == 3)
    .Select(rgs => new { id = int.Parse(rgs[0]), sym = rgs[1].Trim().ToLower() })
    .GroupBy(a => a.sym)
    .Select(grp => grp.ArgMin(a => a.id))
    .ToDictionary(a => a.sym, a => a.id);
```

C# LINQ (Language Integrated Query)

- Declarative operations on sequences
- Recommendation:

Joseph C. Rattz, Jr. (2007) *Pro LINQ: Language Integrated Query in C# 2008*. Apress.


THE EXPERT'S VOICE® IN .NET
Pro
LINQ
Language Integrated Query in C# 2010
Adam Freeman and Joseph C. Rattz, Jr.
Apress®