Start Recording

- Today:
  - Critical Paper Review
  - Assignment 1 solutions
  - Regex Review
  - Cluster Computing
  - Probability

Resources on course website

- Charles Grinstead and J. Laurie Snell *Introduction to Probability* - online on course website
- Unicode, UTF-8, BOM reference
- Some other probability references

Project 1

- Due at 11:45 p.m. next Thursday
- Please see the updated project1.pdf
- Condor or Patas issues?
- Questions?

Assignment 2

- Due August 7th
- Probability (to be covered soon)

Writing assignment

- Due September 4^{th} , 2016
- Short Critical review of a paper from the computational linguistics literature
- Formatted according to ACL guidelines
- Any published journal or peer-reviewed paper on a comp. ling. topic is acceptable
- Send me the paper you plan to review once you have selected it for "approval"

Assignment 1

1. Thank you for your essays. Full credit

Assignment 1

2. I saw that gas can explode.

- I realized that gas (in general) is able to explode.

(ROOT (S (NP (PRP I)) (VP (VBD saw) (SBAR (IN that) (S (NN gas) (VP (MD can) (VP (VB explode)))))))(. .)))

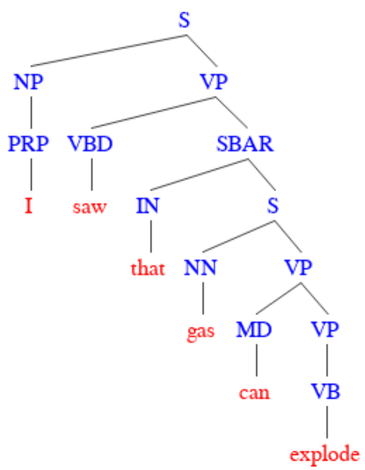
- I (literally) saw that (particular) gas container explode.

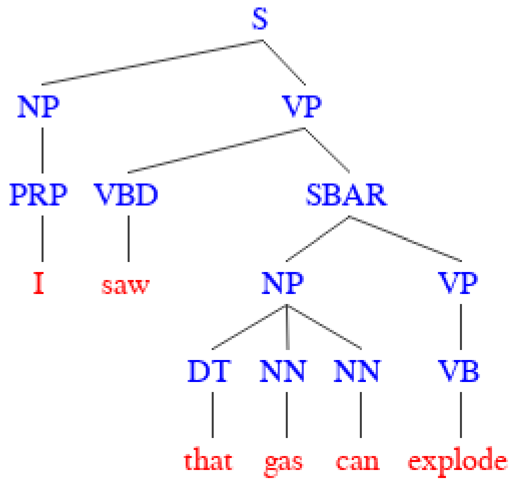
(ROOT (S (NP (PRP I)) (VP (VBD saw) (SBAR (NP (DT that) (NN gas)) (VP (MD can) (VP (VB explode))))))(. .)))

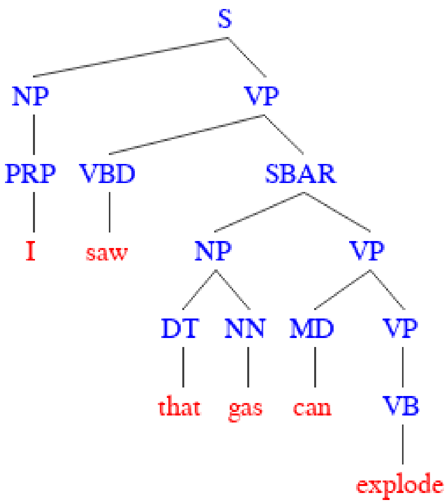
- I realized (that) that (particular) gas is able to explode.

(ROOT (S (NP (PRP I)) (VP (VBD saw) (SBAR (NP (DT that) (NN gas) (NN can)) (VP (VB explode)))))(. .)))

- etc…







3. All possible 6-letter words (26^6)

subtract words with all consonants (21^6)

subtract words with all vowels (5^6)

$$26^6 - 21^6 - 5^6 = 223{,}134{,}030$$

4. Assuming that we consider identical characters to be indistinguishable in the output:

(萄 萄 萄 萄 橙 橙 苹 梨 蕉)

repeated groups: 萄:4, 橙:2

$$\frac{9!}{4! \times 2!} = 7{,}560$$

slide adapted from Glenn Slaydon

5. How many pairwise comparisons are possible between documents on the same topic?

$$\binom{7}{2} + \binom{9}{2} + \binom{3}{2} = 60$$

How many pairwise comparisons are possible between documents on different topics?

$$(7 \times 9) + (7 \times 3) + (9 \times 3) = 111$$

6. Extra Credit

Write an expression that gives the number of unordered sets of k items that can be formed from a set of n distinct items while allowing repetition in the output set.

example: { a, b, c, d } choose 3 (unordered), but allowing repetition in the output:

$n = 4$, $k = 3$

total sets = 20

{ a a a }, { a a b }, { a a c }, { a a d },
{ a b b }, { a b c }, { a b d }, { a c c },
{ a c d }, { a d d }, { b b b }, { b b c },
{ b b d }, { b c c }, { b c d }, { b d d },
{ c c c }, { c c d }, { c d d }, { d d d }

Divide into groups and count them (remember: $n = 4, k = 3$):

a-group

{ a a a } { a a b } { a a c } { a a d } { a b b }

{ a b c } { a b d } { a c c } { a c d } { a d d }

$$\binom{4}{2} + 4 = 10$$

b-group

{ b b b } { b b c }{ b b d } { b c c } { b c d } { b d d }

$$\binom{3}{2} + 3 = 10$$

c-group

{ c c c } { c c d }{ c d d }

$$\binom{2}{2} + 2 = 10$$

d-group

{ d d d }

$$\binom{1}{2} + 1 = 10$$

$$\sum_{i}^{n} \left(\binom{i}{k-1} + i \right) = \boxed{\binom{n+k-1}{k}} = \left(\!\binom{n}{k}\!\right)$$

This is called the multiset coefficient

Assignment 1

{a, b, c, d} multichoose 3 (remember: $n = 4, k = 3$):

Every time we choose, we should add back into the set a copy of whatever we just chose. E.g., if we choose a, we just add another a. If we choose b, we just add another b. We will do this *k - 1* times.

So we really have a set that looks like this:

{a, b, c, d, X, X}

Where each X is going to have a value equivalent to whatever the last chosen item was. So now we have a regular choose function. Our choose is:

$$\binom{n + \underbrace{k - 1}_{\text{The number of Xs}}}{k}$$

Combinatorics Summary

{ a b c }

- Permutation: how many different orderings?

  (a b c) (a c b) (b a c) (b c a) (c a b) (c b a) $n!$

- Combination: how many different subsets (i.e. of 2)?

  { a b } { a c } { b c } $\binom{n}{k}$

  allowing repetition in the output

  { a a } { a b } { a c } { b b } { b c } { c c } $\binom{n+k-1}{k}$

- Variations: how many different ordered subsets (i.e. of 2)?

  (a b) (a c) (b a) (b c) (c a) (c b) $\frac{n!}{(n-k)!}$

  allowing repetition in the output

  (a a) (a b) (a c) (b a) (b b) (b c) (c a) (c b) (c c) n^k

RegEx Review

^	matches the start of a line
$	matches the end of a line
.	matches any one character (except newline)
[*xyz*]	matches any one character from the set
[^*pdq*]	matches any one character not in the set
|	accepts either its left or its right side
\	escape to specify special characters

anything else: must match exactly

RegEx Review

*	accepts zero or more of the preceding element
	this is the canonical ‘greedy’ operator
?	accepts zero or one of the preceding element(s)
+	accepts one or more of the preceding element(s)
{*n*}	accepts *n* of the preceding element(s)
{*n*,}	accepts *n* or more of the preceding element(s)
{*n*,*m*}	accepts *n* to *m* of the preceding element(s)

(*pattern*) defines a capture group which can be referred to later via \1

RegEX Practice

- Find hyphenated words

  grep ‘[a-z]\-[a-z]’
- Find English words with a consonant doubled by the English spelling rule (eg. hitting from hit)

  grep ‘[aeiouy]([bcdfghjklmnpqrstvwxz])\1[aeiouy]’
- Find consonant clusters of two or more

  grep ‘[bcdfghjklmnpqrstvwxz]{2,}’
- Check for words preceded by the wrong form of a/an

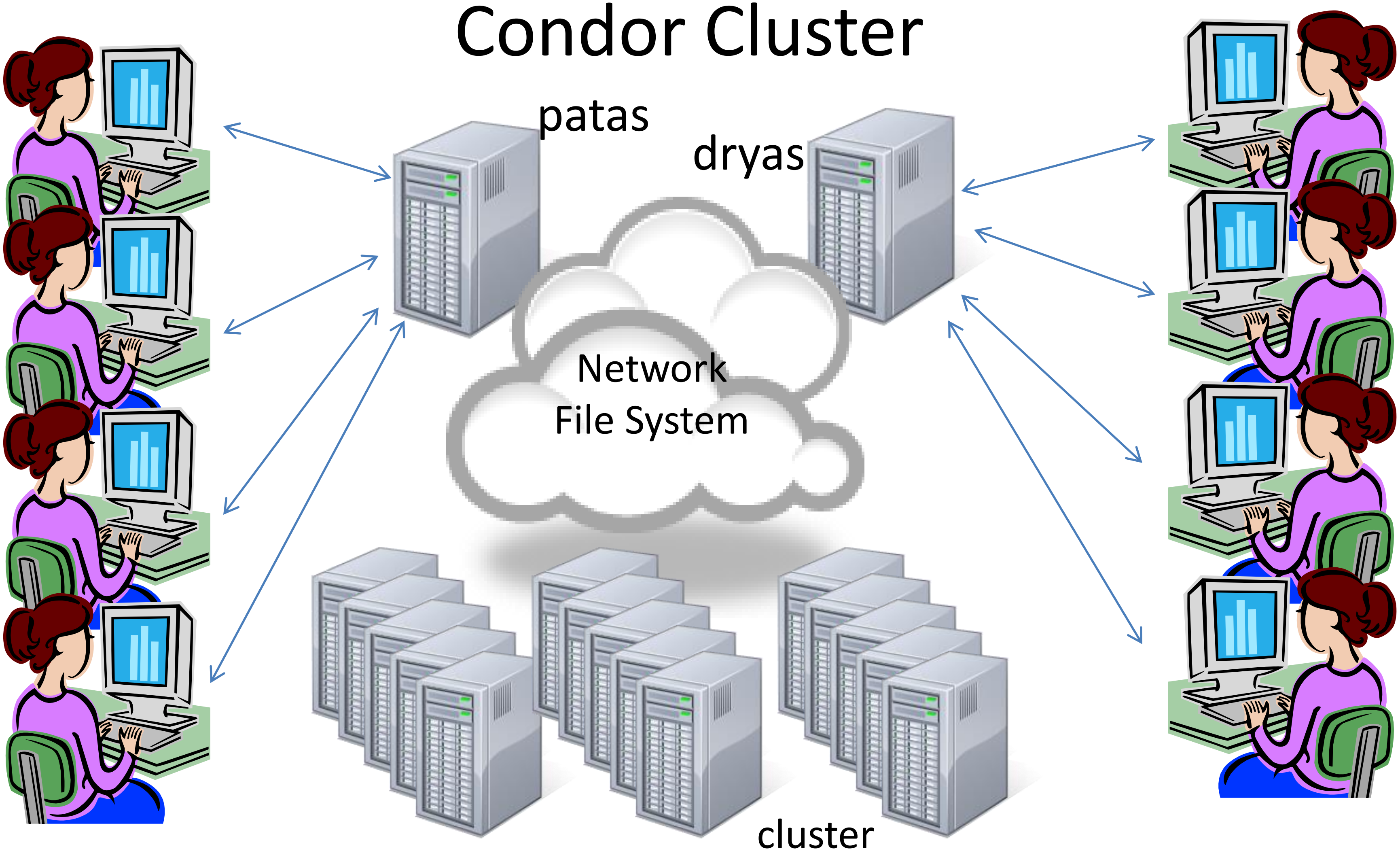
  grep ‘\s[Aa]n [bcdfghjklmnpqrstvwxz]’

  grep ‘\s[Aa] [aeiouy]’

Primitive tokenization

```
$ cat moby_dick.html |              # echo the text
tr [:upper:] [:lower:] |            # convert to lower case
tr ' ' '\n' |                       # put each word on a line
grep -v ^$ |                        # get rid of blank lines
grep -v '<' |                       # get rid of HTML tags
grep -o "[a-z']*" |                 # only want letters and '
sort |                              # sort the words
uniq |                              # find the vocabulary
wc -l                               # count them
  3956
```

Condor Cluster



Condor

```
$ condor_submit myjob.cmd
```

```
universe            = vanilla
executable          = /usr/bin/python
getenv              = true
input               = myinput.in
output              = myoutput.out
error               = myerror.err
log                 = /tmp/kphowell/mylogfile.log
arguments           = myprogram.py -x
transfer_executable = false
queue
```

The system will send you email when your job is complete.

Using variables in Condor files

`flexible.job`

```
file_ext             = $(depth)_$(gain)
universe             = vanilla
executable           = /opt/mono/bin/mono
getenv               = true
output               = acc_file.$(file_ext)
error                = q4.err
log                  = /tmp/gslayden/q4.log
arguments            = myprog.exe model_file.$(file_ext) sys_file.$(file_ext)
transfer_executable = false
queue
```

```
$ condor_submit -append "depth=20” -append "gain=4" flexible.job
```

Probability

- So far, we have considered counting and manipulating sets of items (entities, elements)
- Probability looks at the event of selecting an item from a set
- This event could also be called an observation or a trial
- We call the set the sample space and assign it a probability mass of 1.0

Sample Spaces

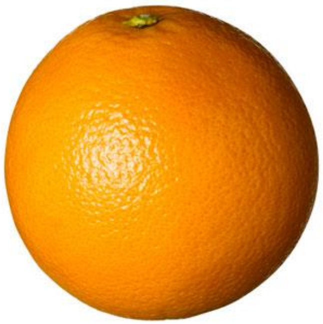
- Ω (omega) is often used to represent the sample space
- Sample spaces can be discrete or continuous

  discrete:

  Ω = (apple, banana, banana, orange)

  continuous:

  Ω = { *the mass of an orange* }



- P(Ω) = 1 (all possible events are accounted for)

most applications in computational linguistics involve discrete probabilities

Outcomes

- Often, events are often notated with a italic capital letter corresponding to a single outcome (a lower-case letter)

  Ω = (a a b c)

  A is the event of selecting ‘a’ from Ω

  B is the event of selecting ‘b’ from Ω

  C is the event of selecting ‘c’ from Ω

- A^C denotes the complement of event *A*
  - $A + A^C$ takes up the entire sample space: P(A or A^C) = 1
- An event can also be any subset of Ω
  - All individual outcomes are events, but events can also be combinations of individual outcomes

  e.g. *Q* is the event of selecting ‘b’ and ‘a’ from Ω, in that order

Rolling 2 dice

- The single occurrence of rolling a red die and a black die must have the one following outcomes (red, black)

$$\Omega = \{ (1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), \\ (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), \\ (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), \\ (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), \\ (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), \\ (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6) \}$$

- There are 6 × 6 = 36 outcomes; they are mutually exclusive and collectively exhaustive
- But there are many other events that we can talk about…

Some 2-dice Events

- A particular outcome

  A = { (3, 6) }

- Both dice are the same

  E = { (1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6) }

- The total is 5

  F = { (1, 4), (2, 3), (3, 2), (4, 1) }

- The total is prime

  G = { (1, 1), (1, 2), (1, 4), (1, 6), (2, 1), (2, 3), (2, 5), (3, 2), (3, 4), (4, 1), (4, 3), (5, 2), (5, 6), (6,1), (6, 5) }

Definition of Probability

- Let P be a function that satisfies the following:

  $P(\Omega) = 1$

  all possible outcomes are accounted for

  $\forall\, A \subseteq \Omega : 0 \leq \mathrm{P(A)} \leq 1$

  probabilities are non-negative real numbers less than 1

  $\forall\, \{ A, B \} \subseteq \Omega, A \cap B = \emptyset : P(A \cup B) = P(A) + P(B)$

  for any pair of events that are mutually exclusive, the union of their occurrence is the sum of their probabilities

$\emptyset$ denotes the empty set, { }

Definition of Probability

- For every trial, an event either occurs, or does not occur

  $\forall A \subseteq \Omega : P(A^C) = 1 - P(A)$

- Each event $A \subseteq \Omega$ can be thought of as partitioning the probability space

Mutual Exclusivity

- It is impossible for two mutually exclusive events to co-occur on the same trial
- For the 2 dice example, each of the 36 basic outcomes are mutually exclusive with each other, and the entire set is collectively exhaustive
- Therefore, one way of defining *P* is to assume that these outcomes are all equally likely:

$$E = \{\,(1,6)\,\}$$
$$P(E) = \frac{1}{|\Omega|} = \frac{1}{36} = .0278$$

Compositional Events

- Events which are not in the set of mutually-exclusive, collectively-exhaustive events can be composed from them
- Compositional events are handy for grouping together certain types of events that we might be interested in
- If the function *P* describes a valid probability space, then the definition of well-formed *P* allows us to calculate *P* for mutually exclusive compositional events

$$\forall \{ A, B \} \subseteq \Omega, A \cap B = \emptyset : P(A \cup B) = P(A) + P(B)$$

- Every trial has an outcome, which may satisfy multiple events; this can be illustrated with Venn Diagrams

2 Dice Events

- Both dice are the same

  E = { (1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6) }

  $P(E) = .0278 \times 6 = .1667$

- The total is 5

  F = { (1, 4), (2, 3), (3, 2), (4, 1) }

  $P(F) = .0278 \times 4 = .1111$

- The total is prime

  G = { (1, 1), (1, 2), (1, 4), (1, 6), (2, 1), (2, 3), (2, 5), (3, 2), (3, 4), (4, 1), (4, 3), (5, 2), (5, 6), (6,1), (6, 5) }

  $P(G) = .0278 \times 15 = .4167$

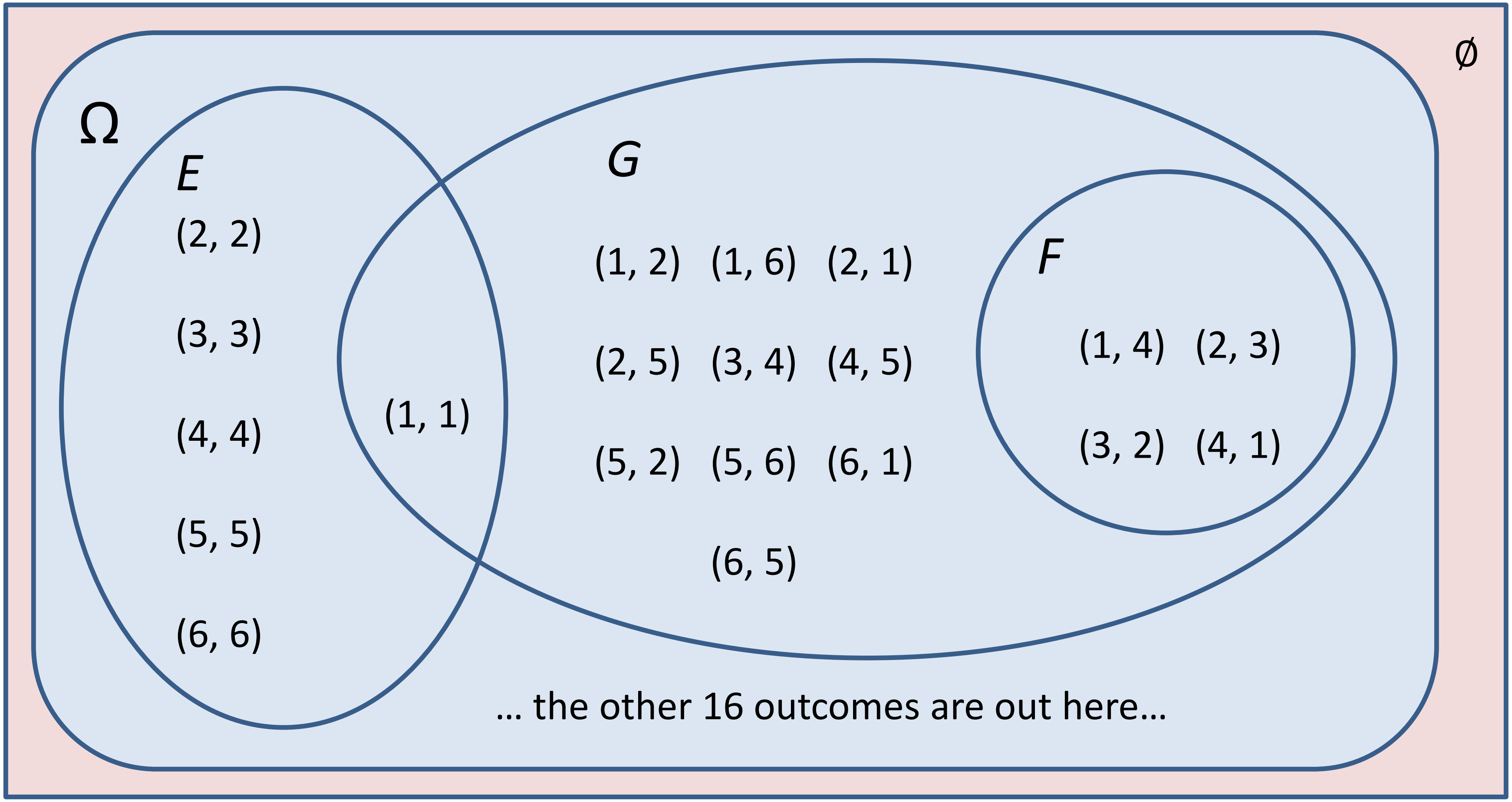
Outcomes in Probability Space

Ω

∅

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Event Composition



Intersecting Events

- The previous slide shows that compositional events can be mutually exclusive

  E and F are mutually exclusive

  $E \cap F = \emptyset$

  E and G are not mutually exclusive

  $E \cap G = \{ (1, 1) \}$

  F and G are not mutually exclusive

  $F \cap G = \{ (1, 4), (2, 3), (3, 2), (4, 1) \} = F$

More on Adding Probabilities

- We have seen how to calculate probability of P(*A* or *B*) when *A* and *B* are mutually exclusive

$$P(A \cup B) = P(A) + P(B),\ A \cap B = \emptyset$$

- If they are not, we can subtract the probability of the intersecting area

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Probability of both dice being the same, *or* their total being prime in a single trial

$$P(E \cup G) = P(E) + P(G) - P(E \cap G)$$
$$= .1667 + .4167 - .0278$$
$$= .5556$$

Joint Probability

- On the previous slide we knew that $P(E \cap G) = .0278$ by noting that only 1 of the 36 mutually exclusive, collectively exhaustive outcomes is in the set intersection $E \cap G$
- More generally though, how can we compute $P(E \cap G)$ from $P(E)$ and $P(G)$?
- $P(E \cap G)$, or $P(E \text{ and } G)$, or $P(EG)$ is the probability that two events both occur in the same trial
- This is called the joint probability
- For mutually exclusive events, the joint probability is obviously zero:

$$\forall \{ A, B \} \subseteq \Omega, A \cap B = \emptyset : P(A \cap B) = 0$$

Joint Probability

Recall our example

$$E = \{ (1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6) \}$$

"both dice are the same"

$$F = \{ (1, 4), (2, 3), (3, 2), (4, 1) \}$$

"the total is 5"

$$\forall \{ A, B \} \subseteq \Omega, A \cap B = \emptyset : P(A \cap B) = 0$$

$$E \cap F = \emptyset, \;\; \therefore P(E \cap F) = 0$$

The probability is zero, meaning it is not possible for both dice to be the same and for the total to be 5 on the same trial.

Joint Probability

- Perhaps it is the case that

  $$P(E \cap G) = P(E)\, P(G)$$

  Let's try it

  $$.0278 \stackrel{?}{=} .1667 \times .4167$$

  $$.0278 \stackrel{?}{=} .0694$$

  No. This means that events *E* and *G* are not independent

Independent Events

- 2 events are mutually exclusive if they cannot both occur as the outcome of a single trial
- 2 events are independent if the occurrence of one does not affect the probability of the other occurring in the trial
- Does event A provide any information that would bias the outcome of event B?
  - If so, A and B are *not* independent events; they are dependent
- Events E, F and G in the 2-dice example are *not* independent of each other ($\{E, F\}$, $\{F, G\}$ and $\{E, G\}$)
  - Even though E and F are mutually exclusive

Adding an independent event to our example

Let’s start with event *F* and try to think of an event that would be independent of *F*

$F = \{ (1, 4), (2, 3), (3, 2), (4, 1) \}$

“the total is 5”

$P(F) = .1111$

It’s not so easy to come up with an event that, in the same trial, will give us no information about *F*. Such an event must meet the following criteria:

- Since F does not partition Ω equally, a event that is independent of F must partition Ω equally, so as not to bias for or against F.

- For the same reason, the event must also partition F equally.

Any ideas?

An event that is independent of *F*

"the red die shows an odd number"

$H = \{ (1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6),$
$(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6),$
$(5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6) \}$

$P(H) = .5$

$F = \{ (1, 4), (2, 3), (3, 2), (4, 1) \}$

"the total is 5"

$P(F) = .1111$

$H \cap F = \{ (1, 4), (3, 2) \}$

$$\mathrm{P}(H \cap F) = .0555 \stackrel{?}{=} P(H)\, P(F) \stackrel{?}{=} .5 \times .111$$

Independent Events

When two events are independent, the probability of both occurring in the same trial is

$$P(A \cap B) = P(A)\,P(B)$$

- Actually, the reverse of this is the *definition* of independence
- This is how we can test for independence of events
  - we can compare the probability $P(A \cap B)$—obtained from counting—to the product of $P(A)$ and $P(B)$. If they are equal, the events are independent

Conditional Probability

- But what if two events are not independent? How do we compute $P(A \cap B)$ from $P(A)$ and $P(B)$?
- We must know how the events are related
- $P(A|B)$ is notation for the probability of event *A*, assuming that event *B* has co-occurred in the same trial
- This is called conditional probability
- “the probability of A, given B”
- Think of a constrained probability space which contains only those outcomes which satisfy event *B*
  - *or a ‘pre-filter’ which selects only outcomes which satisfy B*

Conditional Probability

- Because the reduced sample space is limited to events which satisfy *B*, we exclude from *A* any outcomes that do not satisfy *B*: $P(A \cap B)$
- This lets us express the conditional probability in terms of the reduced sample space

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

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

- If $P(B)$ is 0, then $P(A|B)$ is undefined

Marginal Probability

- Conditional probability introduces the idea that you might have information about *part* of a trial
- In the following equation, we are assuming that we can estimate or provide $P(B)$ for an incomplete trial

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

- P(B) here is called the marginal probability

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

joint probability = conditional probability × marginal probability

Conditional probability and independence

- Note that conditional probability degrades gracefully in the case of independent events
- Assuming *A* and *B* are independent events:

$$P(AB) = P(A)\,P(B)$$
$$P(AB) = P(A|B)\,P(B)$$
$$P(A)\,P(B) = P(A|B)\,P(B)$$
$$P(A) = P(A|B)$$

$$P(AB) = P(A)\,P(B)$$
$$P(AB) = P(B|A)\,P(A)$$
$$P(A)\,P(B) = P(B|A)\,P(A)$$
$$P(B) = P(B|A)$$

If *A* and *B* are independent, then what you may know about one doesn’t affect the probability of the other

Summary of Event Probability

- $P(A^C) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- If $P(A \cap B) = P(A)\,P(B)$, then A and B are called independent events
- Otherwise

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

- Conditional probability

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

Probability Practice

A month of the year is selected at random for an annual conference

- What is the probability that the conference will be in summer (defined as June-August)? $\frac{1}{4}$
- What is the probability that the month will start with a J? $\frac{1}{4}$
- What is the probability that it will be in summer or start with a J? $P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{3}{12} + \frac{3}{12} - \frac{2}{12} = \frac{4}{12}$
- Given that the conference will be in summer, what is the probability that it will start with a J? $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2/12}{1/4} = \frac{2}{3}$
- Are these two events independent?

$P(A \cap B) = P(A)\ P(B)$ $\frac{2}{12} \neq \frac{1}{4} * \frac{1}{4}$ no

Next Week

- Tuesday
  - Random Variables, the Chain Rule, and Probability Distributions
- Next Thursday
  - Project 1 due at 11:45 p.m.