Independence

Reading

- Malik Magdon-Ismail. Discrete Mathematics and Computing.
  - Chapter 17

Overview

- Independence is an assumption
  - Fermi method
  - Multiway independence
- Coincidence and the birthday paradox
  - Application to hashing
- Random walk and gambler's ruin

Independence is a Simplifying Assumption

Rensselaer

- Tosses of different coins have nothing to do with each other
  - independent
- What about two siblings' eye color?
  - (Depends on genes of parent)
  - not independent
- Making lecture slides mistakes
  - independent (assuming I am equally tired every time I work on slides)
- Cloudy and rainy days.
  - When it rains, there must be clouds.
  - not independent

Independence is a Simplifying Assumption, cont'd

- Toss two coins:

$$\mathbb{P}[Coin\ 1 = H] = \frac{1}{2}, \mathbb{P}[Coin\ 2 = H] = \frac{1}{2}, \mathbb{P}[Coin\ 1 = H\ AND\ Coin\ 2 = H] = \frac{1}{4}$$

- Toss both coins 100 times:
  - Coin 1 $\approx$ 50H
    - (of these) Coin 2 $\approx$ 25H
    - since they are independent tosses
- Independence allows us to conclude the following:

$$\mathbb{P}[Coin\ 1 = H\ AND\ Coin\ 2 = H] = \frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = \mathbb{P}[Coin\ 1 = H] \times \mathbb{P}[Coin\ 2 = H]$$

- If the variables are not independent, we can't split the probability, e.g.,

$$\mathbb{P}[rain\ AND\ clouds] = \mathbb{P}[rain] = \frac{1}{7} \gg \frac{1}{35} = \mathbb{P}[rain] \times \mathbb{P}[clouds]$$

Definition of Independence

- Events A and B are independent if "They have nothing to do with each other."
- Knowing the outcome is in B does not change the probability that the outcome is in A
- Formally, events A and B are *independent* if
$$\mathbb{P}[A \; AND \; B] = \mathbb{P}[A \cap B] = \mathbb{P}[A] \times \mathbb{P}[B]$$
- Keep in mind that in general (regardless of independence):
$$\mathbb{P}[A \cap B] = \mathbb{P}[A|B] \times \mathbb{P}[B]$$
- So independence means that
$$\mathbb{P}[A|B] = \mathbb{P}[A]$$
- Independence is a non-trivial assumption, and you can't always assume it.
- When you can assume independence

PROBABILITIES MULTIPLY

Fermi-Method: Chance of Reaching Troy from Albany

- In order to reach Troy from Albany, one needs to overcome many potential obstacles:

$$A_1 = no\ flat\ tires$$
$$A_2 = avoid\ protruding\ manholes$$
$$A_3 = avoid\ accidentally\ going\ the\ wrong\ way$$
$$A_4 = avoid\ running\ out\ of\ gas$$
$$A_5 = avoid\ getting\ stuck\ in\ traffic\ due\ to\ road\ work$$

- So $A = reach\ Troy = A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5$
  - All criteria must be met
- These are more or less independent events, so:

$$\mathbb{P}[A] = \mathbb{P}[A_1] \times \mathbb{P}[A_2] \times \mathbb{P}[A_3] \times \mathbb{P}[A_4] \times \mathbb{P}[A_5]$$

  - All events magically happen with the same probability of $\frac{364}{365}$
- So finally,

$$\mathbb{P}[reach\ Troy] = \left(\frac{364}{365}\right)^5 = 0.986$$

  - i.e., you should expect to not make it to Troy on 5 days/year

Multiway Independence

- When you have multiple events, independence can be tricky
- Suppose I have 3 fair coins. Here are all the outcomes:

Ω	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
$\mathbb{P}[\omega]$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

- Suppose we have the following events:

$$A_1 = \{coins\ 1\ and\ 2\ match\}$$
$$A_2 = \{coins\ 2\ and\ 3\ match\}$$
$$A_3 = \{coins\ 1\ and\ 3\ match\}$$

- What are their probabilities:

$$\mathbb{P}[A_1] = \mathbb{P}[A_2] = \mathbb{P}[A_3] = \frac{1}{2}$$

- How about the pairwise conjunctions ("AND"s)?

$$\mathbb{P}[A_1 \cap A_2] = \mathbb{P}[A_2 \cap A_3] = \mathbb{P}[A_1 \cap A_3] = \frac{1}{4}$$

  - So they are independent (e.g., $\mathbb{P}[A_1 \cap A_2] = \mathbb{P}[A_1] \times \mathbb{P}[A_2]$)

Multiway Independence, cont'd

Rensselaer

- When you have multiple events, independence can be tricky
- Suppose I have 3 fair coins. Here are all the outcomes:

Ω	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
$\mathbb{P}[\omega]$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

- Suppose have the following events:

$$A_1 = \{coins\ 1\ and\ 2\ match\}$$
$$A_2 = \{coins\ 2\ and\ 3\ match\}$$
$$A_3 = \{coins\ 1\ and\ 3\ match\}$$

- What are their probabilities: $\mathbb{P}[A_1] = \mathbb{P}[A_2] = \mathbb{P}[A_3] = \frac{1}{2}$
- How about the AND of all 3 events:

$$\mathbb{P}[A_1 \cap A_2 \cap A_3] = \frac{1}{4}$$

  – not independent (why?)
  – (1,2) match and (2,3) match → (1,3) match.

Multiway Independence, cont'd

- Mutual independence for more than 2 events is stronger than just 2-way independence between all events!
- Formally, events $A_1, \dots, A_n$ are **independent** if the probability of *any intersection* of distinct events is the *product* of the event-probabilities of those events,

$$\mathbb{P}[A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}] = \mathbb{P}[A_{i_1}] \times \mathbb{P}[A_{i_2}] \times \cdots \times \mathbb{P}[A_{i_k}]$$

  - Distinct events don't have any outcomes in common

Coincidence: Let's Try to Find a FOCS-Twin

Rensselaer

- Assume we have 200 students: $S = \{s_1, s_2, \dots, s_{200}\}$
- Suppose I go in order. How many students do I need to ask until I find a FOCS-twin?
  - Assume birthdays are *independent* (no twins, triplets, etc.) and all birthdays are equally likely
- How do we go about computing the probability that at least two people were born on the same day?
  - We want the probability $\mathbb{P}[s_1, \dots, s_{200}\ have\ no\ FOCS - twin]$ (why?)
    $$\mathbb{P}[at\ least\ one\ FOCS\ - twin] = 1 - \mathbb{P}[s_1, \dots, s_{200}\ have\ no\ FOCS - twin]$$
  - I can start with s_1. What is $\mathbb{P}[s_1\ has\ no\ FOCS - twin]$?
  - Suppose s_1 was born on Jan. 1 (the actual date doesn't matter).
  - What is the probability that s_2 was not born on Jan. 1?
    $$\mathbb{P}[s_2\ not\ born\ on\ Jan. 1] = \frac{365}{366}$$
  - In general, what is the probability that all students weren't born on Jan. 1?
    $$\mathbb{P}[s_2, \dots, s_{200}\ not\ born\ on\ Jan. 1] =$$
    $$= \mathbb{P}[s_2\ not\ born\ on\ Jan. 1] \times \cdots \times \mathbb{P}[s_{200}\ not\ born\ on\ Jan. 1]$$

Coincidence: Let's Try to Find a FOCS-Twin, cont'd

Rensselaer

- Assume we have 200 students: $S = \{s_1, s_2, \dots, s_{200}\}$
- Suppose I go in order. How many students do I need to ask until I find a FOCS-twin?
  - Assume birthdays are *independent* (no twins, triplets, etc.) and all birthdays are equally likely
- How do we go about computing the probability that at least two people were born on the same day?
  - Probability that student s_1 has no FOCS twin

$$\mathbb{P}[s_1\ has\ no\ FOCS-twin] = \left(\frac{B-1}{B}\right)^{N-1} = \left(\frac{365}{366}\right)^{199}$$

- Now what about the next student? What is $\mathbb{P}[s_1, s_2\ have\ no\ FOCS-twin]$?
  - Let's use the definition of conditional probability

$$\mathbb{P}[s_1, s_2\ have\ no\ FOCS-twin] =$$
$$= \mathbb{P}[s_2\ has\ no\ FOCS-twin | s_1\ has\ no\ FOCS-twin] \times$$
$$\times \mathbb{P}[s_1\ has\ no\ FOCS-twin]$$

Coincidence: Let's Try to Find a FOCS-Twin, cont'd

- Assume we have 200 students: $S = \{s_1, s_2, \dots, s_{200}\}$
- Suppose I go in order. How many students do I need to ask until I find a FOCS-twin?
  - Assume birthdays are *independent* (no twins, triplets, etc.) and all birthdays are equally likely
- How do we go about computing the probability that at least two people were born on the same day?
  - Probability that student s_1 has no FOCS twin

$$\mathbb{P}[s_1 \; has \; no \; FOCS-twin] = \left(\frac{B-1}{B}\right)^{N-1} = \left(\frac{365}{366}\right)^{199}$$

- Now what about the next student?

$$\mathbb{P}[s_2 \; has \; no \; FOCS-twin | s_1 \; has \; no \; FOCS-twin] =$$

$$= \left(\frac{B-2}{B-1}\right)^{N-2} = \left(\frac{364}{365}\right)^{198}$$

  - The two birthdays are independent, but the events $A_1 = \{s_1 \; has \; no \; FOCS - twin\}$ and $A_2 = \{s_2 \; has \; no \; FOCS - twin\}$ are dependent!
    - s_1 constrains the values s_2 can take (can't equal s_1)

Coincidence: Let's Try to Find a FOCS-Twin, cont'd

- How do we go about computing the probability that at least two people were born on the same day?

$$\mathbb{P}[s_1\ has\ no\ FOCS - twin] = \left(\frac{B-1}{B}\right)^{N-1} = \left(\frac{365}{366}\right)^{199}$$

$$\mathbb{P}[s_2\ has\ no\ FOCS - twin | s_1\ has\ no\ FOCS - twin] = \left(\frac{B-2}{B-1}\right)^{N-2} = \left(\frac{364}{365}\right)^{198}$$

$$\mathbb{P}[s_3\ has\ no\ FOCS - twin | s_1, s_2\ have\ no\ FOCS - twin] = \left(\frac{363}{364}\right)^{197}$$

$$\vdots$$

$$\mathbb{P}[s_k\ has\ no\ FOCS - twin | s_1, \dots, s_{k-1}\ have\ no\ FOCS - twin] = \left(\frac{366-k}{366-k+1}\right)^{200-k}$$

- Let's see what we can do with this information. We want

$$\mathbb{P}[s_1, \dots, s_k\ have\ no\ FOCS - twin]$$

- How do we proceed?

Coincidence: Let's Try to Find a FOCS-Twin, cont'd

Rensselaer

- Let's see what we can do with this information. We want

$$\mathbb{P}[s_1, \dots, s_k \text{ have no FOCS} - \text{twin}]$$

  - Use the definition of conditional probability

$$\begin{aligned}
\mathbb{P}[s_1, \dots, s_k \text{ have no FOCS} - \text{twin}] &= \mathbb{P}[s_k \text{ has no FOCS} - \text{twin} | s_1, \dots, s_{k-1} \text{ have no FOCS} - \text{twin}] \times \\
&\quad \times \mathbb{P}[s_1, \dots, s_{k-1} \text{ have no FOCS} - \text{twin}] \\
&= \mathbb{P}[s_k \text{ has no FOCS} - \text{twin} | s_1, \dots, s_{k-1} \text{ have no FOCS} - \text{twin}] \times \\
&\quad \times \mathbb{P}[s_{k-1} \text{ has no FOCS} - \text{twin} | s_1, \dots, s_{k-2} \text{ have no FOCS} - \text{twin}] \times \\
&\quad \times \mathbb{P}[s_1, \dots, s_{k-2} \text{ have no FOCS} - \text{twin}] \\
&= \mathbb{P}[s_k \text{ has no FOCS} - \text{twin} | s_1, \dots, s_{k-1} \text{ have no FOCS} - \text{twin}] \times \\
&\quad \dots \times \mathbb{P}[s_2 \text{ has no FOCS} - \text{twin} | s_1 \text{ has no FOCS} - \text{twin}] \\
&\quad \times \mathbb{P}[s_1 \text{ has no FOCS} - \text{twin}]
\end{aligned}$$

- So,

$$\mathbb{P}[s_1, \dots, s_k \text{ have no FOCS} - \text{twin}] = \left(\frac{365}{366}\right)^{199} \times \left(\frac{364}{365}\right)^{198} \times \cdots \times \left(\frac{366-k}{366-k+1}\right)^{200-k}$$

- Probability goes up very quickly

Finding a FOCS-twin by the kth student with class size 200

k	1	2	3	4	5	6	7	8	9	10	23	25
chances (%)	42.0	66.3	80.4	88.6	93.3	96.1	97.7	98.7	99.2	99.5	99.999	100

The Birthday Paradox

Rensselaer

- In a party of 50 people, what are the chances that two have the same birthday?
- Same as asking for

$$\mathbb{P}[s_1, \dots, s_{50}\ have\ no\ FOCS - twin]$$

- Answer:

$$\mathbb{P}[no\ social\ twins] = \left(\frac{365}{366}\right)^{49} \times \left(\frac{364}{365}\right)^{48} \times \cdots \times \left(\frac{315}{316}\right)^{0} \approx 0.03$$

- Chances are about 97% that two people share a birthday!
- **Moral:** when *searching* for something among many options (1225 pairs of people), *do not be surprised* when you find it
  - Why 1225 pairs?

$$\#\ \text{pairs} = \frac{50!}{2!\,48!}$$

- Also known as the infinite monkey problem:
  - Given enough time "typing", a monkey will eventually type any given text, including the complete works of Shakespeare

Search and Hashing

- Search is a fundamental part of the modern internet
- It still the main part of Google's business
- To search fast, we need hashing
  - Comparing strings is waaay too slow
- Consider the 3 pages on the right

https://page.1
It snows too much in Troy

https://page.2
It snows more in Hamilton

https://page.3
It's always sunny in Philadelphia

- Google could just build a sorted list of each word, and the pages it appears in
  - How long would search take?
    - With binary search, $O(\log n)$
  - Too slow for large n, want $O(1)$

always	{page.3}
Hamilton	{page.2}
in	{page.1, page.2,page.3}
is	{page.3}
it	{page.1, page.2,page.3}
more	{page.2}
much	{page.1}
Philadelphia	{page.3}
snows	{page.1, page.2}
too	{page.1}
Troy	{page.1}

Search and Hashing, cont'd

Rensselaer

https://page.1
It snows too much in Troy

https://page.2
It snows more in Hamilton

https://page.3
It's always sunny in Philadelphia

- Instead of sorting, hash words into a table
  - For example, take a word, raise each letter to a large prime power, e.g., 17, and take the remainder w.r.t. another large prime, e.g., 11
- E.g.,

$$HASH('Troy') = 20^{17} + 18^{17} + 15^{17} + 25^{17} \equiv 2 \ (mod\ 11)$$

- Then, when someone searches for Troy, hash and look up page number
- But

$$HASH('it') = 9^{17} + 20^{17} \equiv 2 \ (mod\ 11)$$

- This is called a collision
  - Needs to be resolved, o.w. search fails
- Good hash function maps words independently and randomly.
- No collisions → $O(1)$ search
  - (constant time, not $\log n$)

0	Hamilton → {page.2}
1	
2	it → {page.1,page.2,page.3}, Troy → {page.1}
3	too → {page.1}
4	always → {page.3}
5	
6	much → {page.1}, sunny → {page.3}
7	
8	in → {page.1, page.2, page.3}, snows → {page.1, page.2}, Philadelphia → {page.3}
9	
10	more → {page.2}

Search and Hashing, cont'd

- There are many ways to resolve collisions
  - E.g., linear search
  - Won't discuss them in detail in this class
- We assume we're given a good hashing function
  - Talk to a number theory expert about various options

Hashing and FOCS-Twins

- The two problems are surprisingly similar

  Words $w_1, w_2, \dots, w_N$ and Hashing $\leftrightarrow$ Students $s_1, s_2, \dots, s_N$ and Birthdays

  $w_1, \dots, w_N$ hashed to rows $0, 1, \dots, B-1 \leftrightarrow s_1, \dots, s_N$ born on days $0, 1, \dots, B-1$

  No collisions, or hash-twins $\leftrightarrow$ No FOCS-twins

- Example: Suppose you have $N = 10$ words $w_1, w_2, \dots, w_{10}$
  - Suppose $B = 10$ (hash table has as many rows as words)

$$\mathbb{P}[no\ collisions] = \left(\frac{9}{10}\right)^9 \times \left(\frac{8}{9}\right)^8 \times \cdots \times \left(\frac{1}{2}\right)^1 \times \left(\frac{0}{1}\right)^0 \approx 0.0004$$

  - Suppose $B = 20$ (hash table has twice as many rows as words)

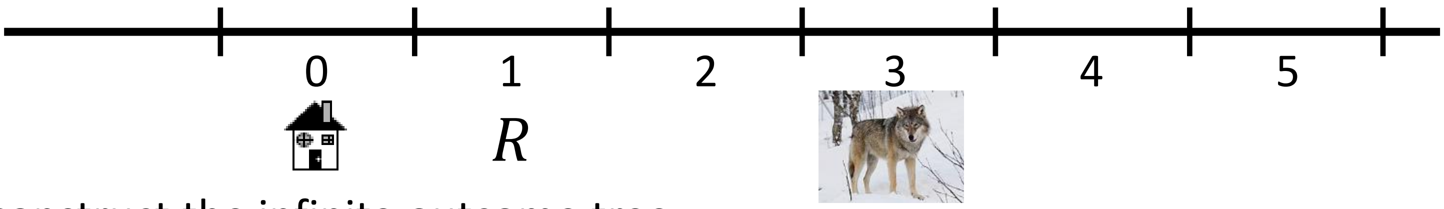
$$\mathbb{P}[no\ collisions] = \left(\frac{19}{20}\right)^9 \times \left(\frac{18}{19}\right)^8 \times \cdots \times \left(\frac{11}{12}\right)^1 \times \left(\frac{10}{11}\right)^0 \approx 0.07$$

B	10	20	30	40	50	60	70	80	90	100	500	1000
$\mathbb{P}$[no collisions]	0.0004	0.07	0.18	0.29	0.38	0.45	0.51	0.56	0.60	0.63	0.91	0.96

- B large enough $\rightarrow$ chances of no collisions are high. How large should B be?
- **Theorem.** If $B \in \omega(n^2)$, then $\mathbb{P}[no\ collisions] \rightarrow 1$

Random Walk: What Are the Chances I Return Home?

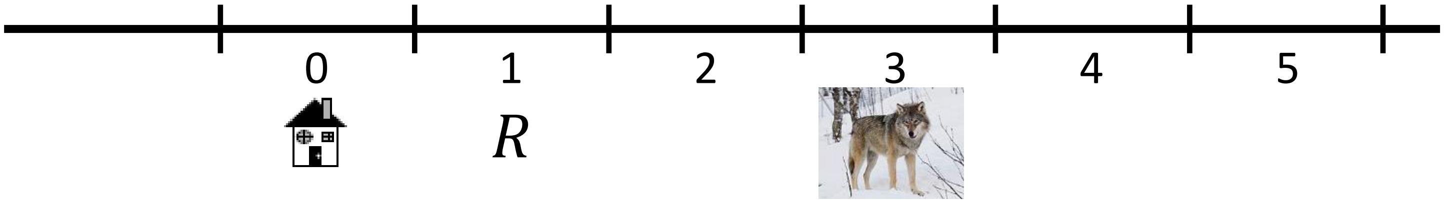
- When I was in college (at Colgate), there were many foggy days
  - On one Thanksgiving, I was the only person on campus
  - I walked outside and immediately got lost (couldn't see anything)
  - If I randomly go left or right, what are the chances I get back home?



- First, construct the infinite outcome tree
- Sequences leading to home
  - L, RLL, RLRLL, RLRLRLL, RLRLRLRLL
  - Corresponding probabilities are $\frac{1}{2}, \left(\frac{1}{2}\right)^3, \left(\frac{1}{2}\right)^5, \left(\frac{1}{2}\right)^7$
- So, the sequences look like: $\mathbb{P}\left[(RL)^{\bullet i}L\right] = \left(\frac{1}{2}\right)^{2i+1}$

$$\mathbb{P}[home] = \frac{1}{2} + \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^5 + \cdots = \frac{\frac{1}{2}}{1 - \frac{1}{4}} = \frac{2}{3}$$

Random Walk: What Are the Chances I Return Home?



- A slick approach is to use the Law of Total Probability

$$\mathbb{P}[home] = \mathbb{P}[L] \cdot \mathbb{P}[home|L] + \mathbb{P}[RR] \cdot \mathbb{P}[home|RR] + \mathbb{P}[RL] \cdot \mathbb{P}[home|RL]$$

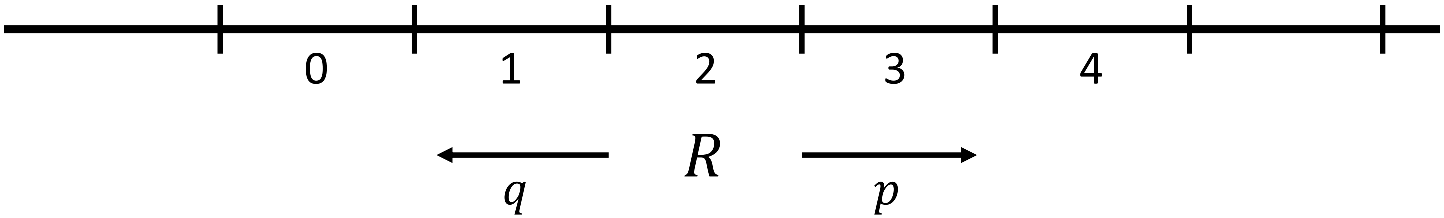
$$= \frac{1}{2} \times 1 + \frac{1}{4} \times 0 + \frac{1}{4}\mathbb{P}[home]$$

- Solving for $\mathbb{P}[home]$, we get:

$$\left(1 - \frac{1}{4}\right)\mathbb{P}[home] = \frac{1}{2}$$

$$\mathbb{P}[home] = \frac{\frac{1}{2}}{1 - \frac{1}{4}} = \frac{2}{3}$$

A Random Walk at the Casino



- Suppose you go to the casino to play roulette
  - You start with \$2 and say you will quit if you make it \$4
  - You bet \$1 on red every time
    - i.e., a random walk with left/lose probability q and right/win probability p
    - What do you think will happen?
- Let P_i be the probability to win in the game if you have \$$i$ (and TINKER!!)

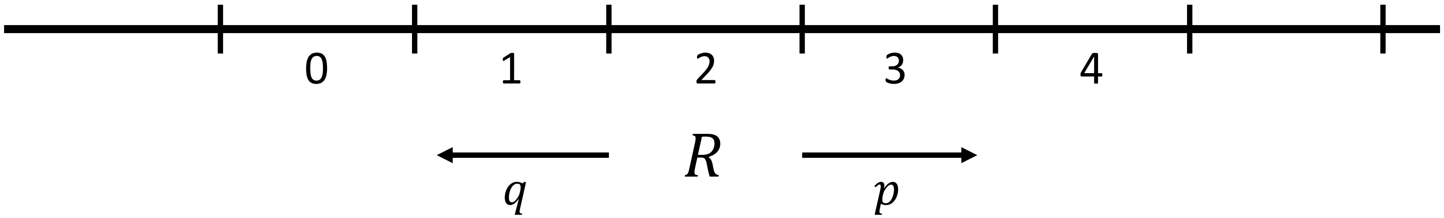
$$P_1 = qP_0 + pP_2 = pP_2 \qquad \textbf{[total probability]}$$
$$P_2 = qP_1 + pP_3 = qpP_2 + pP_3$$

i.e., $P_2 = \frac{pP_3}{1-qp}$

$$P_3 = qP_2 + pP_4 = \frac{qpP_3}{1-qp} + p \qquad [\boldsymbol{P_4 = 1} \textbf{ because we win}]$$

$$P_3 = \frac{p(1-pq)}{1-2pq}$$

A Random Walk at the Casino, cont'd



- Let P_i be the probability to win in the game if you have $\$i$ (and TINKER!!)

$$P_1 = qP_0 + pP_2 = pP_2 \quad \text{[total expectation]}$$

$$P_2 = qP_1 + pP_3 = qpP_2 + pP_3$$

i.e., $P_2 = \frac{pP_3}{1-qp}$

$$P_3 = qP_2 + pP_4 = \frac{qpP_3}{1-qpP_2} + p \quad [\boldsymbol{P_4 = 1} \textbf{ because we win}]$$

$$P_3 = \frac{p(1-pq)}{1-2pq}$$

- This means that

$$P_2 = \frac{pP_3}{1-qp} = \frac{p^2}{1-2pq}$$

  – When $q = 0.6, p = 0.4$, $P_2 \approx 0.31$ (i.e., 69% chance of losing your money!)

A Random Walk at the Casino, cont'd

- **Exercises.**
  - What if you are trying to double up from $3?
    - (Answer: 77% chance of RUIN)
  - What if you are trying to double up from $10?
    - (Answer: 98% chance of RUIN)
  - Suppose you have infinite money, and your goal is to win any positive amount or at least break even. You start by betting $2 and keep doubling your bet every time you lose (or leave if you win). What happens then?