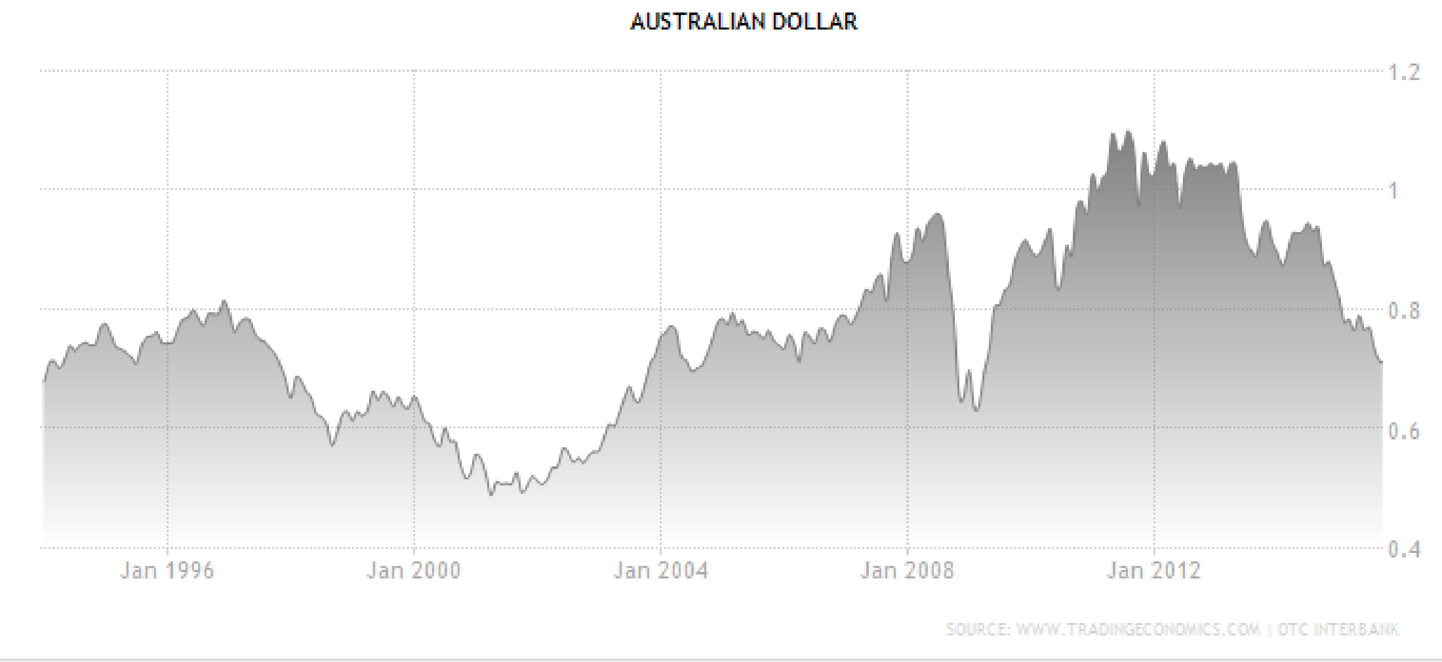
37161 Probability and Random Variables

Lecture 9

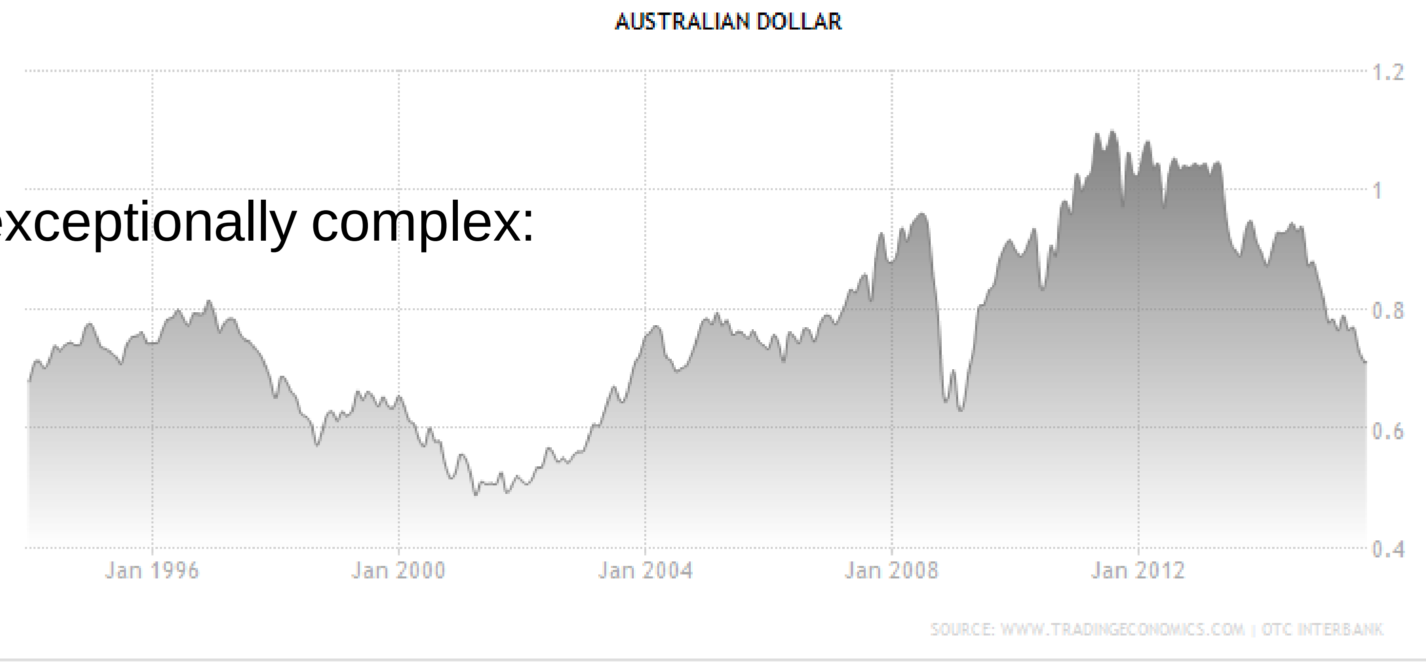
Time Series

- A **time series** is a collection of datapoints recording observations made at successive intervals of some time period.



- For example, the graph here show the conversion of the $US to $AUS over the past 20 years.

Time Series



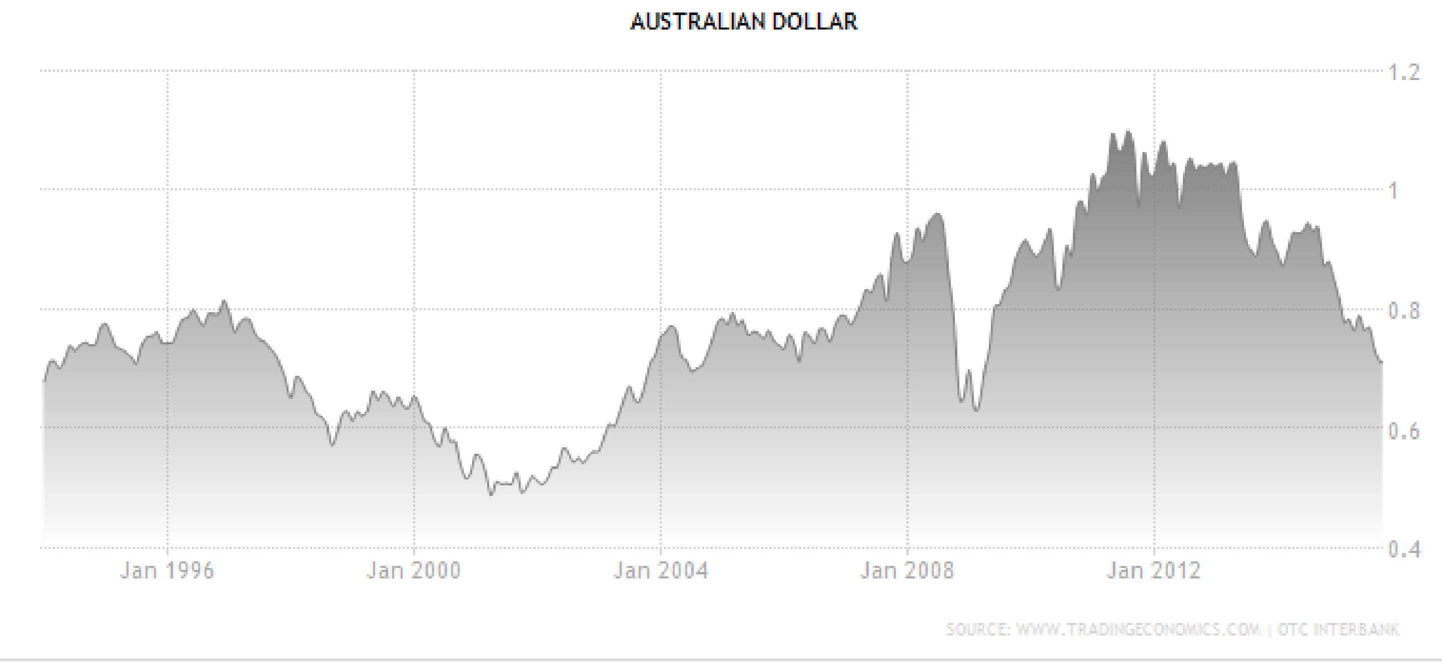
- Analysing such time series can be exceptionally complex:

- Does the distribution of possible future datapoints depend only on the current level?

- Is there "momentum" in the system i.e. should past performance be used as an predictor as well?

- Can other factors outside the series of observations also be useful predictors?

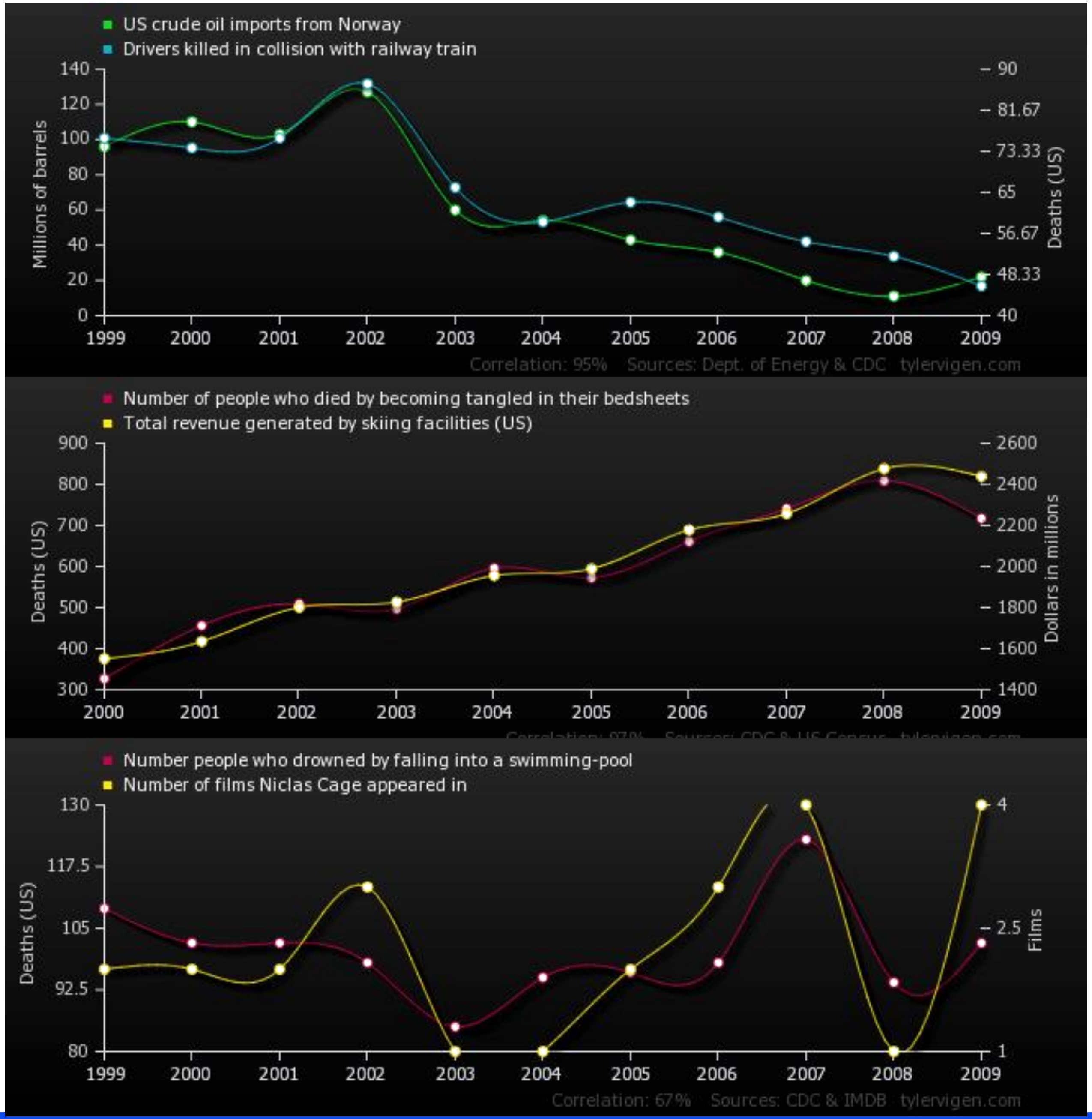
Time Series



- Most time series analyses are concerned with making predictions about possible future observations, based on both the past trends and any other relevant factors.
- Typical analyses include weather forecasting, economic/financial modelling, climate modelling etc.
- Understanding what is and is not relevant from observed datasets in predicting future trends is often highly debated.

Time Series

- The website Spurious Correlations (http://www.tylervigen.com/spurious-correlations) lists a few humourous examples of cases when, despite strong correlation, a past trend might not be such a good predictor of future results…



Markov Chains

- Consider a sequence of random variables $\{X_{0,} X_{1,} X_{2,} \ldots\}$ each of which takes one of a finite or countable number of values.
- We call the set of possible values for each variable the states of the system, S
- For many problems of interest, we have the property that

$$P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, \ldots X_0 = i_0) = P(X_{n+1} = j \mid X_n = i)$$

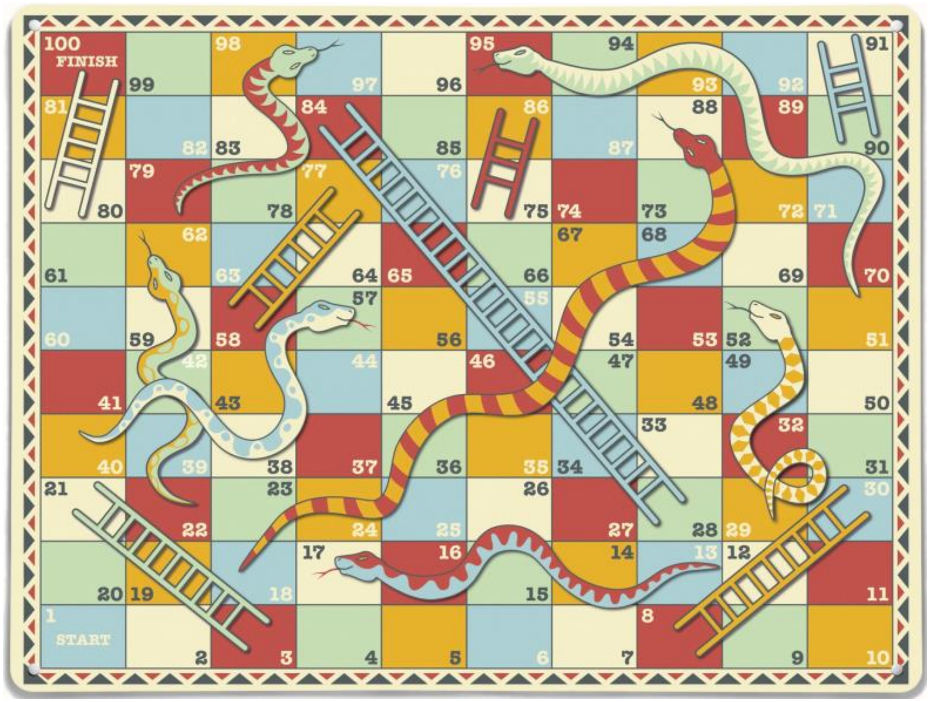
for all possible n and all possible $j, i, i_{n-1}, i_{n-2}, \ldots, i_0 \in S$

- In other words, the only necessary predictor of future states is the current one.
- Such a system is called a **Markov Chain**.

Andrey Markov

(1856 -1922)

Markov Chains



- For example, with a standard game of snakes and ladders, with squares numbered 1-100, the square that a given player is on at the end of each turn forms a Markov Chain with states $S = \{1, 2, 3, \ldots, 100\}$

- This is because if we were told the full history of all a player's positions, for example

$$X_{11} = 41, X_{10} = 36, X_9 = 31, X_8 = 26, X_7 = 24, X_6 = 19, X_5 = 15, X_4 = 12, X_3 = 7, X_2 = 4, X_1 = 2, X_0 = 1$$

and asked to calculate the distribution of X_{12}, the only relevant information is that $X_{11} = 41$.

- The possible moves from square 41 do not depend on how the player got there.

Markov Chains

- Which of the following might reasonably be modelled with Markov Chains through time?:
- The proportion of people in a country's population who have green eyes .
  - Markov Chain. No reason to assume proportion is more likely to increase or decrease in time.
- The proportion of people in a country's population who have a harmful genetic condition.
  - Not Markov Chain. If the condition is harmful, fewer people with it may be able to reproduce.
- The volume of water in a bucket, given that 1 Litre is added per minute with none removed.
  - Markov Chain. We could know future volumes from knowing only current volume.
- The amount of money a gambler has playing a sequence of independent identical games.
  - Depends… If it's a game of pure chance, then would be a Markov Chain. If there is skill involved, then prior states would give evidence of the player's possible future performance.

Random Walks

- One of the most commonly studied types of Markov Chain is a **random walk**.

- A random walk is a sequence defined on the integers such that a probability P_{ij} is assigned to the probability of the (n+1)th state being j, given that the nth state is i.

- Consider for example the simple game of repeatedly betting $1 on the flip of a fair coin to win $2 if the result is Heads and $0 if the result is Tails.

- This defines a Markov Chain since, for example, the chance of going from $15 to $16 is completely independent of any previous flips, i.e. whether you started with $1 and kept winning until you got $15 or whether you started with $100 and kept losing until you got $15.

A Simple Random Walk

- For example, consider a game whereby a player bets $1 on the flip of a fair coin. He/she wins $2 is the result is Heads and $0 if the result is Tails.

- Staring with $10, he/she bets repeatedly until he/she either has $0 or $20.

- We can set up a **difference equation** in W_k for the probability that he eventually leaves in profit, given that he has $k \in \{0,1,2,.....20\}$ dollars at a given time.

- The probability can be broken down conditional on the next game. Given that he/she has $*k*, he/she has a 50% chance of having (*k*+1) dollars after the next game and a 50% chance of having $(*k* – 1).

A Simple Random Walk

- We know that if we have two events B_1 and B_2 which together form a partition of the sample space, i.e. $P(B_1 \cup B_2) = 1$ and $P(B_1 \cap B_2) = 0$ then, for any event A, $P(A) = P(A \cap B_1) + P(A \cap B_2)$
- Applying this same idea for each individual game, we have that the probability that a player wins overall is equal to the probability that he/she wins the next game and wins overall plus the probability that he/she loses the next game but wins overall.
- This is $W_k = P(\text{loses next game}) \times W_{k-1} + P(\text{wins next game}) \times W_{k+1}$
- For a game where the probability of winning and losing $1 are each 50%, the difference equation is $W_k = \frac{1}{2} W_{k+1} + \frac{1}{2} W_{k-1}$
- This is solved, subject to boundary conditions $W_0 = 0$ since if he ever has $0, he cannot win overall (probability 0) and $W_{20} = 1$ since if he reaches $20

Difference Equations

- A **first order difference equation** (i.e. one in which the highest and lowest numbered terms differ by exactly 1) with constant coefficients can be easily solved by inspection.

- For example, if $C_{n+1} = 2C_n$, we can easily see that $C_{n+2} = 2(C_{n+1}) = 2^2 C_n$ and $C_{n+3} = 2(C_{n+2}) = 2^3 C_n$.

- In general, $C_{n+1} = 2C_n$ is solved by $C_n = K2^n$ for some constant K.

- Similarly, for constant M, $C_{n+1} = MC_n$ is solved by $C_n = KM^n$.

Difference Equations

- How in general do we solve a **second order difference equation** (i.e. one in which the highest and lowest numbered terms differ by exactly 2)?
- To solve, for example $C_{n+2} - 7C_{n+1} + 10C_n = 0$, we look for a solution of the form $C_n = KM^n$ for constants *K* and *M*.
- If we get two different values for *M*, then the general solution is $C_n = K_1M_1{}^n + K_2M_2{}^n$.
- If we get one repeated value for *M*, then the general solution is $C_n = K_1M^n + K_2nM^n$.
- Compare this with solving second order differential equations of the form $\frac{d^2y}{dx^2} + A\frac{dy}{dx} + y = 0$ to find solutions of the form $y = K_1e^{m_1x} + K_2e^{m_2x}$ (or $y = K_1e^{mx} + K_2xe^{mx}$ if one repeated value of *m* is found.)

Difference Equations

- To solve, for example $C_{n+2} - 7C_{n+1} + 10C_n = 0$, we look for a solution of the form $C_n = KM^n$ for constants K and M.
- If such a solution exists, $C_{n+1} = KM^{n+1} = (KM^n)M = C_n M$ and $C_{n+2} = KM^{n+2} = (KM^n)M^2 = C_n M^2$.
- Substituting these in gives $M^2 C_n - 7MC_n + 10C_n = 0$ or $C_n\left[M^2 - 7M + 10\right] = 0$.
- Apart from the trivial solution $C_n = 0$, we have that $M^2 - 7M + 10 = (M-5)(M-2) = 0$ so $M = 5$ or 2.
- The general solution is therefore $C_n = K_1 5^n + K_2 2^n$.
- Boundary conditions are needed to work out the two constants.

Difference Equations

- Returning to our initial problem of $W_k = \frac{1}{2}W_{k+1} + \frac{1}{2}W_{k-1}$, we seek solutions of the form $W_k = AM^k$.
- We find M by solving $W_k = \frac{1}{2}MW_k + \frac{1}{2}\frac{W_k}{M}$ or $M^2 - 2M + 1 = 0$.
- As this gives $(M-1)(M-1) = 0$, we solve to find that $M = 1$.
- This gives a repeated value for M, so $W_k = A_1 1^k + A_2 k 1^k$ or $W_k = A_1 + A_2 k$.
- The boundary conditions of $W_0 = 0$ and $W_{20} = 1$ give that $W_0 = A_1 = 0$ and $W_{20} = A_1 + 20A_2 = 1$.
- Together, we get that $W_k = \frac{k}{20}$ so, for example, the probability of winning \$20 rather than \$0 after starting with \$10 is, perhaps obviously 50%. This is because the starting position is equally spaced from both boundaries. .

Asymmetric Random Walks

- The cases when the probabilities of winning and losing each game are not equal is a little simpler to solve as it does not involve a repeated value of *M*.

- For example, consider a European style roulette wheel. For a \$1 bet, a gambler gets \$2 back with probability $\frac{18}{37}$ and \$0 back with probability $\frac{19}{37}$.
- If a gambler plays repeatedly until first getting to \$*D* or \$0, then the difference equation which describes his overall chance of winning, given that he has \$*k* at a given point is

$$W_k = \frac{18}{37} W_{k+1} + \frac{19}{37} W_{k-1} \text{ with boundary conditions } W_0 = 0 \text{ and } W_D = 1$$

Asymmetric Random Walks

- We find M via $18M^2 - 37M + 19 = 0$ so $(18M - 19)(M - 1) = 0$.
- $M = \dfrac{19}{18}$ or 1 hence the general solution is

$$W_k = A_1\left(\frac{19}{18}\right)^k + A_2(1)^k = A_2 + A_1\left(\frac{19}{18}\right)^k$$

- The boundary conditions give

$$W_0 = A_2 + A_1 = 0 \text{ and } W_D = A_2 + A_1\left(\frac{19}{18}\right)^D$$

Asymmetric Random Walks

- The solution of $W_k = \frac{18}{37} W_{k+1} + \frac{19}{37} W_{k-1}$ with boundary conditions $W_0 = 0$ and $W_D = 1$ is therefore $W_k = \frac{1 - \left(\frac{19}{18}\right)^k}{1 - \left(\frac{19}{18}\right)^D}$.
- So, even though the odds are very nearly fair ($\frac{18}{37} \approx 48.65\%$), if starting with \$100 and playing until reaching either \$200 or \$0, the probability of ending in profit is only

$$W_{100} = \frac{1 - \left(\frac{19}{18}\right)^{100}}{1 - \left(\frac{19}{18}\right)^{200}} \approx 0.45\%$$

- For a truly fair game, this would be 50%, of course. .

Simple Random Walks: How Many Games?

- We have calculated the probabilities of either quitting in profit or bankrupt, but how many games should the player expect to have to play before concluding his bets, assuming a fair game?

- Clearly, if he starts at \$10 and leaves at \$0 or \$20, he must play a minimum of 10 games (either 10 losses or 10 wins) but could keep playing a lot longer than that.

- Let C_k be a "counter" of how many games he expects to play until he finishes, given that he has \$$k$.

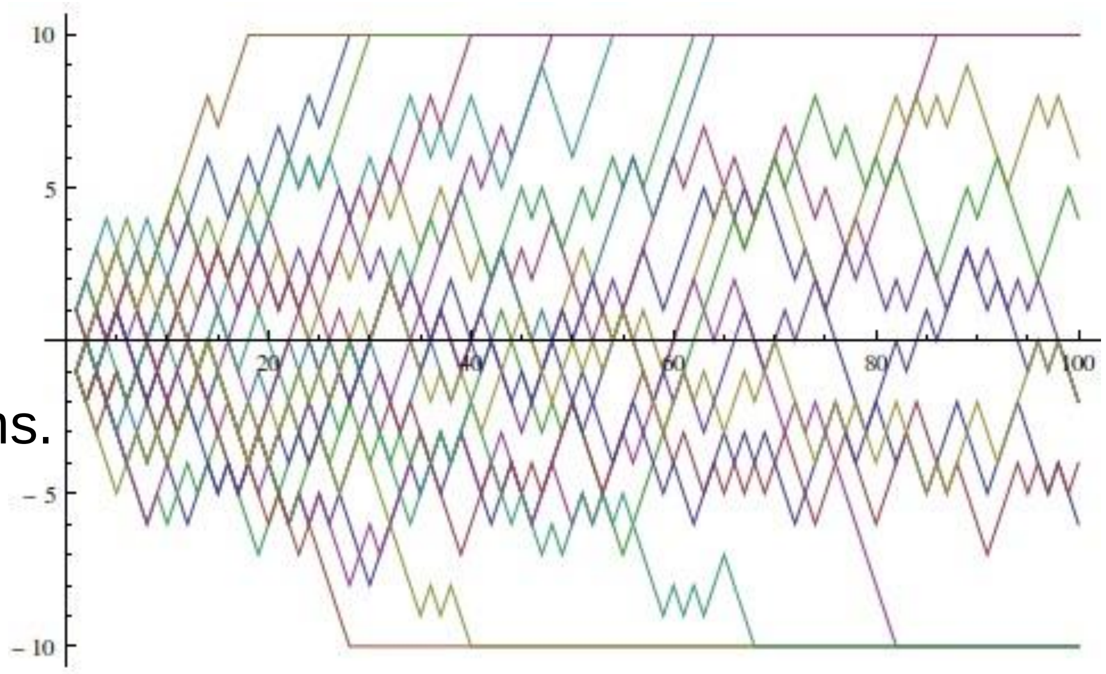
- Again, conditioning on the next move, he either wins \$1 (with probability 0.5) or loses \$1 (with probability 0.5). Either way, the "counter" moves on by one.

- We therefore get $C_k = \frac{1}{2}[C_{k+1} + 1] + \frac{1}{2}[C_{k-1} + 1]$.

Simple Random Walks: How Many Games?

- $C_k = \frac{1}{2}[C_{k+1} + 1] + \frac{1}{2}[C_{k-1} + 1]$.
- The boundary conditions are $C_0 = 0, C_{20} = 0$ since he requires no games to quit if he starts with $0 or $20.
- The solution of this is found via a nonhomogeneous difference equation with repeated roots – beyond the scope of this subject. (For completeness, the solution is $C_k = 20k - k^2$)
- Again, perhaps unsurprisingly, this confirms that the expected number of games is maximised when $k = 10$ i.e. when the starting position is furthest from either boundary.
- What do we do in such cases when we know the mechanisms of a model, but are unable to calculate an analytical solution?

Simulation



- Computer packages can be used to generate large numbers of realisations of random systems.
- Expectations of random variables can then be estimated from averaging the outcomes of these simulations.
- For example, the following is the output of 20 realisations of a random walk beginning with $10 and playing until first reaching $20 ($10 profit) or $0 (shown here as $ –10 profit.)
- As well as being able to estimate probabilities of winning etc. expected numbers of games or other more obscure properties can be estimated.