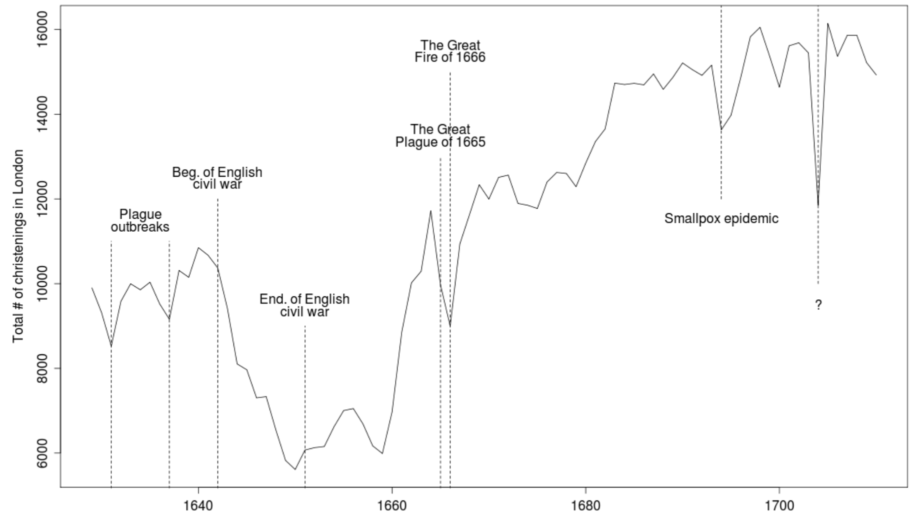
Lecture 3 - More Conditional Probability

Statistics 102

Colin Rundel

January 22, 2013

Lab 1 - Extra Credit



Basic Probability Review

$$0 \le P(E) \le 1$$

$$P(E^c) = 1 - P(E)$$

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

$$P(E \cap F) = P(E|F) \times P(F) = P(F|E) \times P(E)$$

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

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

Disjoint and Independent

$$\text{Independent} \not\Rightarrow \text{Disjoint}$$
$$\text{Disjoint} \not\Rightarrow \text{Independent}^\dagger$$

Your intuition can easily be wrong on this type of problem, always check using the definitions:

Independent:

$$P(E \cap F) = P(E) \times P(F) \text{ or } P(E|F) = P(E)$$

Disjoint:

$$P(E \cap F) = 0$$

Other Important Terms

Joint Distribution - $P(E \cap F)$

Marginal Distribution - $P(E)$

Conditional Distribution - $P(E|F)$

Sampling

Imagine an urn filled with white and black marbles
...or a deck of cards
...or a bingo cage
...or a hat full of raffle tickets

Two common options + one extra for completeness:

- Sampling without replacement
- Sampling with replacement
- Pólya's urn

What is the probability of being dealt two aces?

What if you replace the first card and reshuffle before showing the second?

What is the probability of being dealt a royal flush in poker?

And now a brief magic trick ...

If you have ever shuffled a deck of cards you have done something no one else has ever done before or will ever do again ...

Birthday Problem

Clicker question

Ignoring leap years, and assuming birthdays are equally likely to be any day of the year, what is the probability that there is at least one shared birthday among the students in this class?

As of last night there are 53 people enrolled in this course,

$$P(\text{at least one match}) = 1 - P(\text{no match}) =$$

Let A_i be the event that student i does not match any of the preceding students then

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)\cdots P(A_n|A_1, \ldots, A_{n-1})$$

Birthday Problem, cont.

Calculation:

$$P(A_1) = 365/365$$

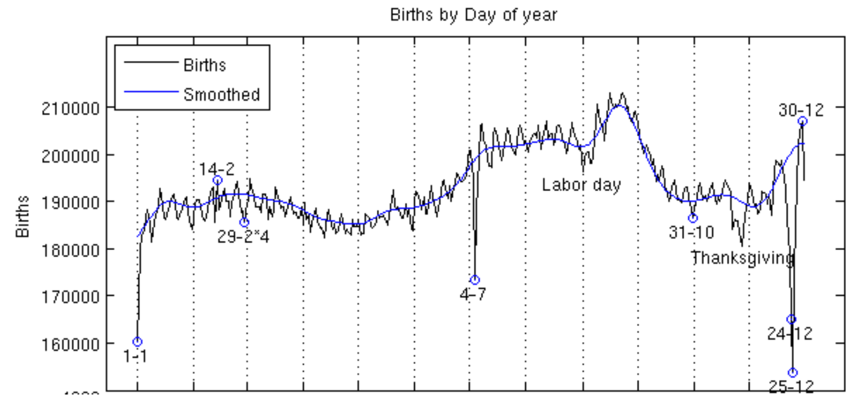
$$\begin{aligned} P(A_1, A_2) &= P(A_1)P(A_2|A_1) \\ &= (365/365) \times (364/365) \end{aligned}$$

$$\begin{aligned} P(A_1, A_2, A_3) &= P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \\ &= (365/365) \times (364/365) \times (363/365) \end{aligned}$$

$$\vdots$$

$$\begin{aligned} P(A_1, \ldots, A_n) &= \frac{365}{365}\frac{364}{365} \cdots \frac{365-(n-1)}{365} \\ &= \frac{365!}{(365-n)!\ 365^n} \end{aligned}$$

Birthday Problem, cont.



The two armed bandit ...

Imagine you walk into a casino and there are two slot machines, the floor manager tells you that one of the machines pays out 50% of the time while the other only pays out 20% of the time. Wanting to make as much money as possible the manager refuses to tell you which machine is which.

What would a good strategy be in this circumstance?

Slots - *http://en.wikipedia.org/wiki/File:Las_Vegas_slot_machines.jpg*

Setup

In order to analyze this problem it helps to define some terms:

- We will label the machines - G for good and B for bad
- We will define W and L as the events of winning and losing respectively
- What we know a priori:
  - $P(W|M = G) = 0.5$
  - $P(L|M = G) = 0.5$
  - $P(W|M = B) = 0.2$
  - $P(L|M = B) = 0.8$

Bayesian vs. Frequentist

In our setup M is what we call a parameter (we'll see more of these in the next couple of lectures).

Initially we know nothing about M but we can collect data (playing the slots) and use that data to make inferences about M (which machine is the good machine).

Frequentist paradigm - long run frequency definition:

- Either $M = G$ or $M = B$ - parameters are fixed but unknown
- Uses likelihood functions - $\mathcal{L}(M|\text{data}) = P(\text{data}|M)$

Bayesian paradigm - degree of belief

- Parameters can have a distribution
- Uses posterior and prior distribution - $P(M|\text{data})$, $\pi(M) = P(M)$

Priors and Posteriors

The Bayesian paradigm is closer to our intuitive approach to solving this problem - I start with a certain set of beliefs about which machine is the good one (e.g. 50-50) and then use data to update those beliefs.

Prior distribution - reflects our initial beliefs:

$$\pi(M = m) = \begin{cases} 0.5 & \text{if } m = G \\ 0.5 & \text{if } m = B \end{cases}$$

Posterior distribution - reflects our beliefs after taking into account data:

$$P(G|\text{data}) = \frac{P(\text{data}|G)}{P(data)}\,\pi(G)$$

Example - Winning

Assume that you pick a machine, you play and you win! What does this tell us about the machine you picked??

Example - Winning and Losing

After winning on on your first play you decide to play the same machine again, and you lose. Now what do we know about the machine you played?

Why do we care?

The two-armed (multi-armed) bandit is a very useful model when it comes to clinical trials.

We are trying one or more treatments against a control and we want to know the efficacy of those treatments. This is much more complex in practice because not only do we not know which is better ($P(G)$ in our slot example) we also don't know how much better they are (also need to estimate $P(W|G)$).

Complex optimization problem where we must allocation a limited number of subjects to properly balance:

- Exploration - estimate the payoff of each treatment
- Exploitation - get the best outcome for the most patients

Back to House and Lupus

Last time we worked through a problem on the probability of a patient having lupus given they test positive. We were given

- $P(L) = 0.02$
- $P(+|L) = 0.99$
- $P(-|L^c) = 0.74$

From which we calculated that

$$P(L|+) = \frac{P(L \cap +)}{P(+)} = \frac{P(+|L)P(L)}{P(+|L)P(L) + P(+|L^c)P(L^c)} = \frac{0.02 \times 0.99}{0.02 \times 0.99 + 0.98 \times 0.26} = 0.072$$

If the patient gets a second test, how should our belief in the probability of having lupus, $P(L)$, change?

Let's Make a Deal...

Monty Hall Problem

You are offered a choice of three doors, there is a car behind one of the doors and there are goats behind the other two.

Monty Hall, Let's Make a Deal's original host, lets you choose one of the three doors.

Monty then opens one of the other two doors to reveal one of the goats.

You are then allowed to stay with your original choice or switch to the other door.

Clicker Question

Which option should you choose?

(a) stay (b) switch (c) it does not matter

A Little History

First known formulation comes from a 1975 letter by Steve Selvin to the American Statistician.

Popularized in 1990 by Marilyn vos Savant in her "Ask Marilyn" column in Parade magazine.

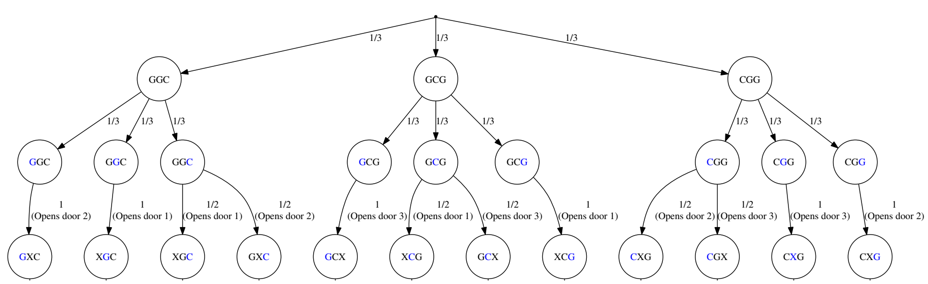
- vos Savant's solution claimed that the contestant should always switch
- About 10,000 (1,000 from Ph.D.s) letters contesting the solution
- vos Savant was right, easy to show with simulation

Moral of the story: trust the math not your intuition

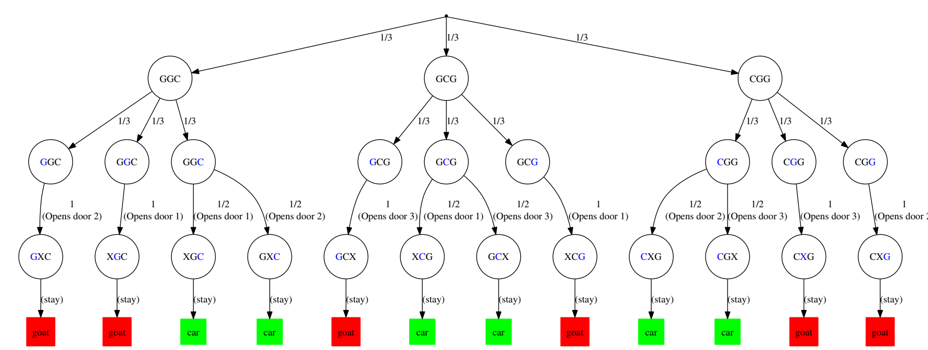
A slightly more entertaining variant of Monty Hall ...

http://www.youtube.com/watch?feature=player_detailpage&v=vDjm4VLfG_g#t=1722

Monty Hall - The hard way



Monty Hall - The hard way - Stay



Monty Hall - The hard way - Switch