Lecture 2: Conditional Probability

Sta 111

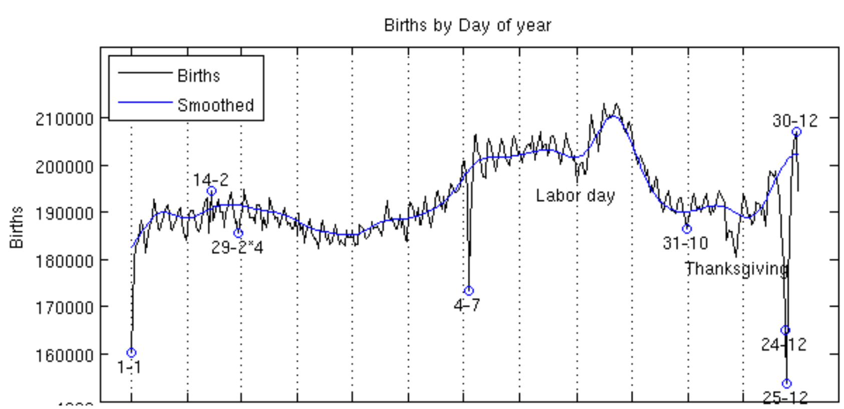
Colin Rundel

May 15, 2014

Birthday Problem

Ignoring leap years, and assuming birthdays are equally likely to be any day of the year, what is the chance of a tie in birthdays among the students in this class?

Birthday Problem, cont.



Approximating the Birthday Problem

We've already seen the birthday problem where D_n is the event that there are no matches with n people

$$P(D_n) = \frac{365}{365}\frac{364}{365}\cdots\frac{365-(n-1)}{365} = \quad \text{for } n \le 365$$

we can rewrite this as

$$\log P(D_n) = \log 1 + \log\left(1 - \frac{1}{365}\right) + \log\left(1 - \frac{2}{365}\right) + \cdots + \log\left(1 - \frac{(n-1)}{365}\right)$$

A common and useful approximation is $\log(1+x) \approx x$ for small values of x

$$\begin{aligned}
\log P(D_n) &\approx -\frac{1}{365} - \frac{2}{365} - \cdots - \frac{n-1}{365} \\
&\approx -\frac{1}{365}(1 + 2 + \cdots + (n-1)) \\
&\approx -\frac{1}{365} \times \frac{1}{2} n * (n-1) \\
P(D_n) &\approx e^{\frac{-n(n-1)}{2\times 365}}
\end{aligned}$$

Another Approximation to the Birthday Problem

Another common and useful approximation is known as Stirling's approximation

$$k! \approx \sqrt{2\pi k}\left(\frac{k}{e}\right)^k$$

we can then rewrite the probability as

$$\begin{aligned} P(D_n) &= \frac{365!}{(365-n)!\ 365^n} \\ &\approx \frac{\sqrt{2\pi}}{\sqrt{2\pi}} \frac{\sqrt{365}}{\sqrt{(365-n)}} \frac{365^{365}}{(365-n)^{365-n}} \frac{e^{-365}}{e^{-(365-n)}} \frac{1}{365^n} \\ &\approx \left(\frac{365}{365-n}\right)^{(365.5-n)} e^{-n} \end{aligned}$$

Approximation Results

n	log approx	Sterling approx	exact
10	0.88401	0.88306	0.88305
20	0.59419	0.58857	0.58856
30	0.30368	0.29369	0.29368
40	0.11801	0.10877	0.10877
50	0.03487	0.02963	0.02963
60	0.00783	0.00588	0.00588
70	0.00133	0.00084	0.00084

Why is Sterling's approximation better?

Conditional Probability

In a setting where we have equally likely outcomes we define conditional probability as:

For a finite set Ω of equally likely outcomes with events A and B then the *conditional probability* of A *given* B is

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

or the proportion of outcomes in B that are also in A.

Conditional Probability, cont.

We can rewrite the counting definition of conditional probability as

$$\begin{aligned} P(A|B) &= \frac{\#(A \cap B)}{\#(B)} \\ &= \frac{\#(A \cap B)/\#(\Omega)}{\#(B)/\#(\Omega)} \\ &= \frac{P(A \cap B)}{P(B)} \end{aligned}$$

which is the general definition of conditional probability.

Note that $P(A|B)$ is undefined if $P(B) = 0$.

Terminology

Conditional Probability:

$$P(A|B)$$

Joint Probability:

$$P(A \cap B)$$

Marginal Probability:

$$P(A)$$

Useful Rules

Very often we may know the probability of events and their conditional probabilities but not probabilities of the events together, in which case we can use

Multiplication rule:

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

Other cases where we do not have the probability of one of the events, we can use

Law of total probably:

For a partition $B_1, \ldots, B_n$ of Ω,

$$P(A) = P(A|B_1)P(B_1) + \ldots + P(A|B_n)P(B_n)$$

Independence

We defined events A and B to be independent when

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

which also implies that

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

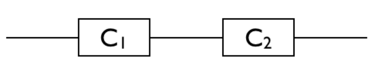
Not to be confused with mutually exclusive events where

$$P(A \cap B) = 0$$

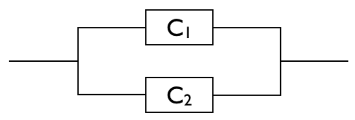
Example, Reliability

If we assume that the probability that a component C_i does not fail in the next week is p_i. What is the probability that the system continues to work for the next week if the failure of C_1 and C_2 are independent?

Series:



Parallel:



Bayes' Rule

Expands on the definition of conditional probability to give a relationship between $P(B|A)$ and $P(A|B)$

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

In the case where $P(A)$ is not known we can extend this using the law of total probability

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

Example - *House*

If you've ever watched the TV show *House* on Fox, you know that Dr. House regularly states, "It's never lupus."

Lupus is a medical phenomenon where antibodies that are supposed to attack foreign cells to prevent infections instead see plasma proteins as foreign bodies, leading to a high risk of blood clotting. It is believed that 2% of the population suffer from this disease.

The test for lupus is very accurate if the person actually has lupus, however is very inaccurate if the person does not. More specifically, the test is 98% accurate if a person actually has the disease. The test is 74% accurate if a person does not have the disease.

Is Dr. House correct even if someone tests positive for Lupus?

Generalizing Conditional Probability

These rules can be naturally extended to more than one event as follows.

For three events:

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

For n events:

$$P(\cap A_i) = P(A_1)P(A_2|A_1)P(A_3|A_1,A_2)\cdots P(A_n|A_1,\ldots,A_{n-1})$$

Return to the Birthday Problem

What we saw previously saw for the birthday problem,

$$\begin{aligned} P(\text{no match in } n) &= \left(\frac{365}{365}\right)\left(\frac{364}{365}\right)\cdots\left(\frac{365-(n-1)}{365}\right) \\ &= \frac{365!}{(365-n)!\,365^n} \end{aligned}$$

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

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

Example - Competing Slots

Imagine two slot machines that have a probability of winning of 50% and 66.6% respectively, but you do not know which machine is which. If we play one machine twice what is the probability that we are playing the good machine given the possible outcomes?

Let W_n be the event of winning on the nth pull and G be the event we played the good machine and B the bad then:

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.

Which option should you choose?

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.

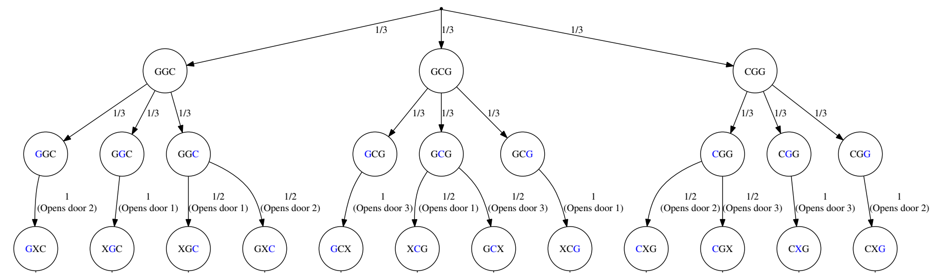
- vos Savant's solution claimed that the contestant should always switch
- About 10,000 (1,000 from Ph.D.s) letters contesting the solution
- vos Savant was right, easy to show with simulation

Moral of the story: trust the math not your intuition

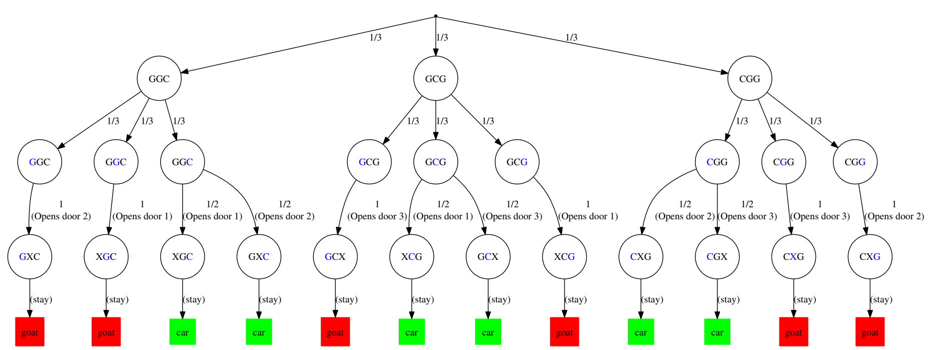
A slightly more entertaining variant of Monty Hall ...

https://www.youtube.com/watch?v=tvODuUMLLgM&t=3m25s

Monty Hall - The hard way



Monty Hall - The hard way - Stay



Monty Hall - The hard way - Switch