Lecture 6: Coincidences, near misses and one-in-a-million chances.

David Aldous

September 12, 2017

Almost every textbook and popular science account of probability discusses the *birthday problem*, and the conclusion

> *with 23 people in a room, there is roughly a 50% chance that some two will have the same birthday.*

This is widely used to illustrate the principle

coincidences are more common than you (intuitively) think.

It's easy to check this "birthday" prediction with real data, for instance from MLB active rosters, which conveniently have 25 players and their birth dates.

[show]

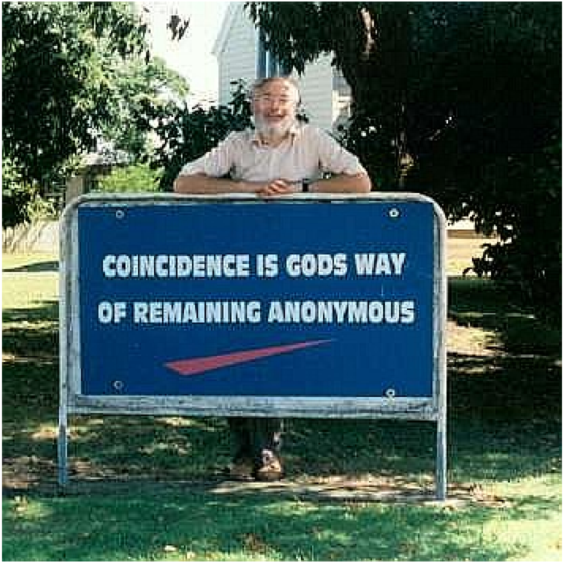
The predicted chance of a birthday coincidence is about 57%. With 30 MLB teams one expects around 17 teams to have the coincidence – can check in freshman seminar course.

The theme of this lecture, and
[Projects;] Can we apply the same sort of mathematical modeling to other real-life perceptions of coincidences?

[show UU-coincidences]

One could focus on some very specific type of coincidence. A calculation later seeks to estimate the probability of meeting someone you know in an unexpected venue. But there is a huge variety of different things we perceive as coincidences.

A long and continuing tradition outside mainstream science assigns spiritual or paranormal significance to coincidences, by relating stories and implicitly or explicitly asserting that the observed coincidences are immensely too unlikely to be explicable as "just chance".



What does math say?

The birthday problem analysis is an instance of what I'll call a *small universe* model, consisting of an explicit probability model, and in which we prespecify what will be counted as a coincidence. Certainly mathematical probabilists can invent and analyze more elaborate small universe models, but these miss what I regard as three essential features of real-life coincidences:
(i) coincidences are judged subjectively – different people will make different judgements;
(ii) if there really are gazillions of possible coincidences, then we're not going to be able to specify them all in advance – we just recognize them as they happen;
(iii) what constitutes a coincidence between two events depends very much on the concrete nature of the events.

I will show a little of the math of "small universe" models and then turn to more interesting real-world settings.

Some math calculations in “small universe” models of coincidences.

Mathematicians have put great ingenuity into finding exact formulas, but it’s simpler and more broadly useful to use approximate ones, based on the informal Poisson approximation. If events $A_1, A_2, \ldots$ are roughly independent, and each has small probability, then the random number that occur has mean (exactly) $\mu = \sum_i \mathbb{P}(A_i)$ and distribution (approximately) Poisson(μ), so

$$\mathbb{P}(\text{none of the events occur}) \approx \exp\left(-\sum_i \mathbb{P}(A_i)\right). \qquad (1)$$

So if we list all possible coincidences in a “small universe” model as $A_1, A_2, \ldots$ then

$$\mathbb{P}(\text{at least one coincidence occurs}) \approx 1 - \exp\left(-\sum_i \mathbb{P}(A_i)\right).$$

For the usual birthday problem, people often ask whether the fact that birthdays are not distributed exactly uniformly over the year makes any difference. So let's consider k people and non-uniform distribution

$$p_i = \mathbb{P}(\text{born of day } i \text{ of the year}).$$

For each *pair* of people, the chance they have the same birthday is $\sum_i p_i^2$, and there are $\binom{k}{2}$ pairs, so from (1)

$$\mathbb{P}(\text{no birthday coincidence}) \approx \exp\left(-\binom{k}{2} \sum_i p_i^2 \right).$$

Write *median-k* for the value of k that makes this probability close to 1/2 (and therefore makes the chance there *is* a coincidence close to 1/2). We calculate [board]

$$\text{median-}k \approx \tfrac{1}{2} + \frac{1.18}{\sqrt{\sum_i p_i^2}}.$$

For the uniform distribution over N categories this becomes

$$\text{median-}k \approx \tfrac{1}{2} + 1.18\sqrt{N}$$

which for $N = 365$ gives the familiar answer 23.

$$\text{median-}k \approx \tfrac{1}{2} + \frac{1.18}{\sqrt{\sum_i p_i^2}}.$$

To illustrate the non-uniform case, imagine hypothetically that there were twice as many births per day in one half of the year as in the other half, so $p_i = \frac{4}{3N}$ or $\frac{2}{3N}$. The approximation becomes $\frac{1}{2} + 1.12\sqrt{N}$ which for $N = 365$ becomes 22.

The smallness of the change ("robustness to non-uniformity") is in fact **not** typical of combinatorial problems in general. In the coupon collector's problem, for instance, the change would be much more noticeable.

Here are two variants. If we ask for the coincidence of *three* people having the same birthday, then we can repeat the argument above to get

$$\mathbb{P}(\text{no three-person birthday coincidence}) \approx \exp\left(-\binom{k}{3}\sum_i p_i^3\right)$$

and then in the uniform case,

$$\text{median-}k \approx 1 + 1.61 N^{2/3}$$

which for $N = 365$ gives the less familiar answer 83.

If instead of calendar days we have k events at independent uniform times during a year, and regard a coincidence as seeing two of these events within 24 hours (not necessarily the same calendar day), then the chance that a particular two events are within 24 hours is $2/N$ for $N = 365$, and we can repeat the calculation for the birthday problem to get

$$\text{median-}k \approx \tfrac{1}{2} + 1.18\sqrt{N/2} \approx 16.$$

A **project** is to look for real-world data for such simple “time” coincidences for events one might expect to happen at random times during a year.

Here are three recent examples.

[show Cancer]

In the context of “deaths linked to illnesses caused by toxic dust issuing from wreckage at Ground Zero” this coincidence is not surprising.

For the more specific context of “deaths **of firefighters** linked to **cancer** caused by toxic dust issuing from wreckage at Ground Zero” I don’t have data. Hypothetically, if rate of such deaths is 20 per year the chance of this triple coincidence is 1% per year.

But we can’t say this is “significant” because one can imagine many other “more specific coincidences” that didn’t happen.

There were 3 passenger jet crashes in 8 days in summer 2014 (Air Algerie July 24th, TransAsia July 23rd, Malaysian Airlines July 17). How unusual is this?

Data: over the last 20 years, such crashes have occurred at rate $1/40$ per day, so under the natural math model (Poisson process)

$$N = \text{number crashes in a given 8 days}$$

has approx Poisson(0.2) dist. and

$$\mathbb{P}(N = 3) \approx e^{-0.2} 0.2^3/6 \approx 1.1 \times 10^{-3}.$$

So how often should we see this "3 crashes in 8 days" event, purely by chance?

General method in my 1989 book *Probability Approximations via the Poisson Clumping Heuristic* for doing approximate calculations in such contexts. Can also do by simulation.

Conclusion. We expect to see this coincidence, purely by chance, on average once every 7 years.

The main conceptual point about coincidences.
We have a context – plane crashes – and we model an observed coincidence as an instance of some “specific coincidence type” – here “3 crashes in 8 days”. But there are many other “ specific coincidence types” that might have occurred, in the context of plane crashes. We could consider a longer window of time – a month or a year – and consider coincidences involving

- same airline
- or same region of the world
- or same airplane model.

Even if a coincidence within any one “specific type” is unlikely, the chance that there is a coincidence in some one of them – somewhere within the context of plane crashes – may be large.

In other words, claims that “what happened is so unlikely that it couldn’t be just chance” rely on an analysis of the specifics of what did happen which does not consider other similar coincidences that didn’t happen (I write **generic** coincidence as opposed to **specific** coincidence).

Moral: “Someone must win the lottery”.

Another email to me. U.S. District Court Judge (Washington DC) Richard Leon handled 3 cases involving the FDA and tobacco companies.

- In January 2010 he prevented the Food and Drug Administration from blocking the importation of electronic cigarettes.
- In February 2012 he blocked a move by the FDA to require tobacco companies to display graphic warning labels on cigarette packages.
- In July 2014 he ruled in favor of tobacco companies and invalidated a report prepared by an FDA advisory committee on menthol.

The question asked of me by a journalist: What are the chances that one judge would pull these major cases when cases are supposedly assigned randomly?
(Not discussing the merits of the judgments)

The implicit question: Is this just coincidence, or does it suggest maybe these cases were not assigned randomly?

It turns out there are 17.5 (explain) judges in this court, so (if random assignment) the chance all 3 cases go to the same judge is $1/17.5 \times 1/17.5 \approx 1/300$.

But there were over 10,000 cases in the period. Imagine looking at all those cases and looking to see where there is a group of 3 cases which are ”very similar” in some sense. The sense might be “same plaintiff and same issue”, as here, but one can imagine many other types of possible similarity. There is surely a huge number N of such groups-of-3, and so there must be a large number N/300 groups assigned to the same judge.

Now of course the FDA-tobacco issue is unusually interesting. A more precise analysis would to go through the 10,000+ cases and find out the number of groups-of-3 that were "very similar" in some sense *of interest to a journalist*. This is some number N, and the chance that some group "of interest to a journalist" were all assigned to the same judge (by pure chance) is N/300. Now I have no idea what N is, but

> *experience with other kinds of coincidence says that there are many more occurrences and more types of "very similar in an interesting way" than you would imagine.*

This is the central point in thinking about real-world coincidences. Here is my best attempt at a concrete illustration of this key point. We used the "random article" link on Wikipedia and looked for coincidences in the topics of the pages found.

	article	article	specific coincidence	chance $\times 10^{-8}$	
1	Kannappa	Vasishtha	Hindu religious figures	12	56
2	Harrowby United F.C.	Colney Heath F.C.	Engl. am. Football Clubs	160	120
3	Delilah	Paul of Tarsus	Biblical figures	20	30
4	USS Bluegill (SS-242)	SUBSAFE	U.S. submarine topics	6	18
5	Kindersley-Lloydminster	Cape Breton-Canso	Canadian Fed. Elec. Dist.	110	23
6	Walter de Danyelston	John de Stratford	14/15th C British bishops	1	81
7	Loppington	Beckjay	Shropshire villages	4	55
8	Delivery health	Crystal, Nevada	Prostitution	9	46
9	The Great Gildersleeve	Radio Bergeijk	Radio comedy programs	4	23
10	Al Del Greco	Wayne Millner	NFL players	3000	77
11	Tawero Point	Tolaga Bay	New Zealand coast	3	32
12	Evolutionary Linguistics	Steven Pinker	Cognitive science	???	36
13	Brazilian battleship Sao Paulo	Walter Spies	Ironic ship sinkings	< 1	28
14	Heap overflow	Paretologic	Computer security	???	52
15	Werner Herzog	Abe Osheroff	Documentary filmmakers	1	92
16	Langtry, Texas	Bertram, Texas	Texas towns	180	53
17	Crotalus adamanteus	Eryngium yuccifolium	Rattlesnake/antidote	< 1	80
18	French 61st Infantry Division	Gebirgsjäger	WW2 infantry	4	45
19	Mantrap Township, Minnesota	Wykoff, Minnesota	Minnesota town(ship)s	810	41
20	Lucius Marcius Philippus	Marcus Junius Brutus	Julius Caesar associate	4	91
21	Colin Hendry	David Dunn	Premier league players	150	62
22	Thomas Cronin	Jehuda Reinharz	U.S. College presidents	32	44
23	Gösta Knuttson	Hugh Lofting	Authors of children's lit.	32	31
24	Sergei Nemchinov	Steve Maltais	NHL players	900	16
25	Cao Rui	Hua Tuo	Three Kingdoms people	37	18
26	Barcelona May Days	Ion Moţa	Spanish Civil War	5	116
27	GM 4L30-E transmission	Transaxle	Auto transmissions	3	37
28	Tex Ritter	Reba McEntire	Country music singers	8	24

Table 1. Coincidences observed in our study. "Chance" is our estimate of the

Another everyday type of coincidence is meeting someone you know in a "unexpected" place on a trip away from home district – not somewhere where either of you would usually be found. The short article by G.J. Kirby estimates the frequency this should happen "by chance" to himself as follows.

- Number of people he knows and would recognize: 212
- Total number of people encountered in a typical trip: 460
- Number of trips away from home district per year: 30
- Adult population (U.K.): 40 million

So his chance of such a "coincidence" meeting in a year

$$\approx 212 \times 460 \times 30/(40 \text{ million }) = 1/14.$$

In fact any calculation of this kind is likely to be an underestimate, because the people you know tend to be similar to you and therefore are more likely to be encountered than a random person.

Bottom line: for most of the real-world coincidences that people find intriguing, we can't model them mathematically well enough to verify the "rationalist" view that they are indeed "just chance".

Further reading.

- 2014 book *The Improbability Principle: Why Coincidences, Miracles, and Rare Events Happen Every Day* by David Hand.
- 1989 paper *Methods for Studying Coincidences* by Persi Diaconis and Frederick Mosteller.

Near misses can be regarded as a kind of coincidence. Common sense often says that near misses may be more likely than hits, but sometimes they are **much** more likely.

Example. Pick 5 letters of alphabet at random, proportional to frequency. What are the chances that
(a) The letters can be arranged to form an English word?
(b) The letters can be arranged to form an English word, if we are allowed to change one letter (our choice of letter) into any other letter we choose?

As intuition suggests, (a) is unlikely but (b) is likely. In our small experiment, the chance (a) was about 18% and the chance (b) was about 94%.

Example. Suppose a just detected asteroid is going to come close to Earth – meaning within the orbit of the Moon. What is the chance it will hit Earth?

[do on board – first ignore bad science fiction]

Example: Near-misses in Lotto picks. In a simple "6 numbers out of 51" type Lotto game, there is 1 winning combination out of $\binom{51}{6} \approx 18$ million. But the number of combinations with 5 out of 6 correct is

$$[\text{board}] \quad 6 \times 45 = 270.$$

Part of the reason for designing lotteries in this "pick k numbers out of n" format is to ensure many near-misses, on the reasonable assumption that observing near-misses will encourage gamblers to continue playing. If, instead, lottery tickets simply represented each of the 18 million possibilities as a number like 12,704,922 between 1 and 18 million, then (counting a near-miss as one digit off) there would be only around 64 near-misses.

A previous student project was **near-misses in bingo** with many players [link] – when one person wins, how many others will have lines with 4 out of 5 filled? Other possible **projects:** near misses in soccer?

Manipulation of near-misses. Exploiting mathematics to design games with many near-misses is generally considered to be within ethical boundaries (every game has rules designed to make it interesting), but other schemes have arguably crossed the boundary. The 2005 book *License to Steal* by Jeff Burbank devotes a chapter to the following story, (summary from an amazon.com review).

> *. . . a slot machine manufacturer had programmed its machines to make it look as if losing spins had just missed being winners – "near misses." The owners claimed that the machine wheels would spin randomly, as they are supposed to, but that once the spin had randomly been determined to be a loser, the wheels would re-adjust to show a near miss. This made it more exciting for the player, who would play more. But the regulators thought it might compromise the appearance of randomness. They decided the near miss feature would not be allowed, but when the company appealed on the grounds that retrofitting thousands of machines would be too expensive, the [Nevada Gaming] Commission cut them some slack. They still went bankrupt.*

The final topic of this lecture is

What really has a 1 in a million chance?

I would like someone to do this, to present at Cal Day in April – could be part of a larger **course project**.

For Cal Day poster, write out 9 events, and tell people

- for 3 of these, the chance really is **about** 1 in a million
- for another 3, chance is considerably **less** than 1 in a million
- for the other 3, chance is considerably **more** than 1 in a million.

What are interesting or counter-intuitive events to choose?

Evans Hall is a few hundred yards from the faultline, so consider

(i) A major (> 6.7 magnitude) earthquake on the Hayward fault in the next 50 minutes.

A 2008 USGS estimate puts the chance at about 1% per year, so the chance (i) is indeed around 1 in a million.

(ii) One of the next 24 babies born in the U.S. will become President.

The U.S. birth rate is currently about 4.0 million per year. If we guess a President will serve on average about 6 years, then it is reasonable to figure that about 1 in 6×4.0 million $=$ 24 million babies will someday be President. So the chance (ii) is indeed around 1 in a million.

[discuss subtle point here]

What about “struck by lightning”? The number of reported U.S. deaths and injuries combined is about 330 per year, so in the sense of population statistics

(iii) being injured or killed by lightning in the next 12 months

has chance around 1 in a million. But as discussed in a previous lecture, this doesn’t make much sense for **you** – it depends on your choice of activities.

Just as I can’t say anything about the chance that **you** win the lottery, unless I know how many tickets you buy.

As a practical matter we can declare that by the phrase “1 in a million chance” we mean “. . . . . . up to a factor of 2 or 3” and we can use common sense to guess how variable the chance is between individuals, and then we can allow ourselves to use population data when we guess it’s not greatly variable. In this sense

(iv) being killed during a 150 mile auto trip in California

has a 1 in a million chance.

Finally, for a memorable instance where people underestimate a chance, I point to a *male* student in the class who is not paying attention and ask for the chance

(v) you get breast cancer sometime.

Finally, for a memorable instance where people underestimate a chance, I point to a *male* student in the class who is not paying attention and ask for the chance

(v) you get breast cancer sometime.

It's rare in men, but not so rare as they think, about 1 in 1,000 lifetime incidence [link]. It may well be greatly variable with family history, so I can't say that 1 in 1000 is the chance for "you", but it's way more than 1 in a million.

“Coincidences on everyday life” is one of our “list of 100 contexts where we perceive chance”. Is there a useful way to categorize coincidences? Here is part of one categorization, from the Cambridge Coincidences Collection page. Does this work well on their examples? Can you do better?

- **Surprising repetitions:** for instance when you’ve had not contact with someone for ages, then find two connections to them very close together in time. Or when over several years multiple members of the same family are born with the same birthday. Or even a repetition of a really rare event – like winning the lottery twice, or your life being saved twice by the same person!
- **Simultaneous events:** for example when two people phone each other at exactly the same time.
- **Parallel lives:** such as when two people in a small group find they share a birthday or an unusual name, or when two people discover their lives match each other in bizarre details.
- **Uncanny patterns:** imagine picking letters in Scrabble that spell your name.
- **Unlikely chains of events:** perhaps you lost your false teeth overboard and found them inside a fish you caught twenty years later?

Miscellaneous

In early 2016 the Powerball lottery prize reached the record level of almost 1.6 billion dollars. A single ticket had a 1 in 292 million chance of winning. There were many short news articles comparing this chance to other chances, mostly in meaningless ways [link].

A recent U.K. article [link] gives a statistical analysis of *Was 2016 especially dangerous for celebrities?* – **project** to repeat.