Chapter 18

Coincidences, near misses and one-in-a-million chances

18.1 The birthday problem and its relatives

The *birthday problem* ($\mathcal{W}$) – often called the *birthday paradox* – is described in almost every textbook and popular science account of probability. My students know the conclusion

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

Rather than repeat the usual "exact" calculation I will show how to do some back-of-an-envelope calculations, in section 18.2 below. Starting from this result there are many directions we could go, so let me point out five of these.

It really is a good example of a quantitative prediction that one could bet money on. In class, and in a popular talk, I show the active roster of a baseball team[1] which conveniently has 25 players and their birth dates. The predicted chance of a birthday coincidence is about 57% With 30 MLB teams one expects around 17 teams to have the coincidence; and one can readily check this prediction in class in a minute or so (print out the 30 pages and distribute among students).

[1] e.g. atlanta.braves.mlb.com/team/roster_active.jsp?c_id=atl; each MLB team has a page in the same format

It's fun to ask students to suggest circumstances where the prediction might not be accurate. This is, if you actually see a group of strangers in a room and know roughly why they are there – people rarely go into rooms "at random" – what make you unsure of the validity of the standard calculation? Two common suggestions are
(i) if you see identical twins
(ii) that the calculation in general may be inaccurate because of non-uniformity of population birth dates over the year.
Point (i) is clear and point (ii) is discussed in the next section (plausible levels of non-uniformity turn out to have negligible effect). Other circumstances involve very creative imagination or arcane knowledge (a party of Canadian professional ice hockey players[2]). As mentioned above, it is a rare example of a mathematically simple yet reliable model!

It illustrates the theme "coincidences are more likely than you think". This is an important theme as regards people's intuitive perception of chance. But the birthday problem and other "small universe" settings, where one can specify in advance all the possible coincidences and their probabilities, are very remote from our notion of weird coincidences in everyday life. A typical blurb for popular science books is "...explains how coincidences are not surprising" while the author merely does the birthday problem. This is surely not convincing to non-mathematicians. I will drepeat this critique more forcefully in section 18.6. My own (unsuccessful) attempt to do better is recounted in section 18.3.

One can invent and solve a huge number of analogous math probability problems and I show a glimpse of such problems in section 18.2. These can be engaging as recreational math and for illustrating mathematical techniques – but I find it almost impossible to produce novel interesting data to complement such theory.

There is an opposite problem with sports data on "hot hands" for individual players, or winning/losing streaks for teams. Here there is plenty of data, but coming up with an accurate chance model is difficult; saying that we see streaks longer than predicted in an oversimplified chance model is not telling us anything concrete about the world of sports.

[2] who have substantial non-uniformity of birthdays. A 1985 paper *Birthdate and success in minor hockey* by Roger Barnsley and A. H. Thompson and subsequent work, popularized in Gladwell's *Outliers*, attributes this to the annual age cutoff for starting minor hockey.

18.2 Using the Poisson approximation in simple models

In this section I want to make the point

> mathematicians know how to do calculations in "small universe" settings, where one can specify in advance all the possible coincidences and their probabilities.

In fact while mathematicians have put great ingenuity into finding exact formulas, it is simpler and more informative 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) $\mu = \sum_i P(A_i)$ and distribution (approximately) Poisson(μ), so

$$P(\text{none of the events occur}) \approx \exp\left(-\sum_i P(A_i)\right). \tag{18.1}$$

Consider the birthday problem with k people and non-uniform distribution

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

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 (18.1)

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

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

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

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.

To illustrate robustness to non-uniformity, imagine hypothetically that half the categories were twice as likely as 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 might be considered another "paradox", and is in fact atypical of combinatorial problems in general. In the coupon collector's problem, for instance, the change would be much more noticable.

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

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

and then in the uniform case,

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

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

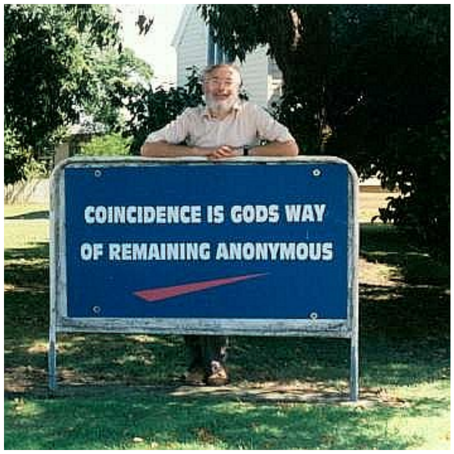
If instead of calendar days we have k events at independent uniform times during a year, and regard a coincidence as seeing two of these events within 24 hours (not necessarily the same calendar day), then the chance that a particular two events are within 24 hours is $2/N$ for $N = 365$, and we can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

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

Finding real-world instances where such theoretical predictions are applicable seems quite hard, in that the first instances one might think of – major fires in a big city, say – have noticeably non-uniform distribution.

18.3 Coincidences in Wikipedia

The material here was written as an introduction to a project we were unable to complete. Nevertheless I like this prose introduction and the data in the table, so am taking the opportunity to report it here. It will be pruned in ta later version.



As the photo[3] suggests, a long and continuing tradition outside mainstream science[4] 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". Self-described rationalists dispute this, firstly by pointing out that (as illustrated by the *birthday problem*) untrained intuition about probabilities of coincidences is unreliable, and secondly by asserting that (in everyday language) observing events with *a priori* chances of one in a gazillion is not surprising because there are a gazillion possible other such events which might have occurred. While the authors (and most readers, we imagine) take the rationalist view, it must be admitted that we know of no particularly convincing studies giving *evidence* that interesting real-life coincidences occur no more frequently than is predictable by chance. The birthday problem analysis is an instance of what we'll call a *small universe* model, consisting of an explicit probability model expressible in abstract terms (i.e. the fact that the 365 categories are concretely "days of the year" is not used) 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 we 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;

[3]The gentleman is not me.
[4]e.g. Arthur Koesler *The Roots of Coincidence*, 1972.

(iii) what constitutes a coincidence between two events depends very much on the concrete nature of the events.
Can onetake one tiny step away from small universe models by studying a setting with these three features?

Almost the only serious discussion of the big picture of coincidences from a statistical viewpoint is a 1989 paper by Persi Diaconis and Fred Mosteller[5]. Our "gazillions" explanation[6], which they call the *law of truly large numbers* and which is also called *Littlewood's law* ($\mathcal{W}$) , is one of four principles they invoke to explain coincidences (the others being hidden cause; memory, perception or other psychological effects; and counting close events as if they were identical). They summarize earlier data in several contexts such as ESP and psychology experiments, show a few "small universe" calculations, and end with the conclusion

> In brief, we argue (perhaps along with Jung) that coincidences occur in the mind of observers. To some extent we are handicapped by lack of empirical work. We do not have a notion of how many coincidences occur per unit of time or how this rate might change with training or heightened awareness. . . . Although Jung and we are heavily invested in coincidences as a subjective matter, we can imagine some objective definitions of coincidences and the possibility of empirical research to find out how frequently they occur. Such information might help us.

Let's take a paragraph to speculate what a mathematical theory of real-life coincidences might look like, by analogy with familiar random walk/Brownian motion models of the stock market. Daily fluctuations of the S&P500 index have a s.d. (standard deviation) of a little less than 1%. Nobody has an explanation, in terms of more fundamental quantities, of why this s.d. is 1% instead of 3% or 0.3% (unlike *physical* Brownian motion, where diffusivity rate of a macroscopic particle can be predicted from physical laws and the other parameters of the system). But taking daily s.d. as an empirically-observed parameter, the random walk model makes testable predictions of other aspects of the market (fluctuations over different time scales; option prices). By analogy, the observed rate of subjectively-judged coincidences in some aspect of real life may not be practically predictable in terms of more fundamental quantities, but one could still hope to develop a self-consistent theory which gives testable predictions of varying aspects of coincidences.

[5] *Methods for studying coincidences*, available by online search.
[6] Further comments will be given in section 18.5.

The simpest aspect to study is surely *single-affinity* coincidences, exemplified in real life by stories such as

> In talking with a stranger on a plane trip, you discover you both attended the same elementary school, which is in a city not on that plane route.

Call this ("same elementary school") a *specific coincidence*; one might plausibly estimate, within a factor of 2 or so, the *a priori* probability of such a specific coincidence. Now a specific coincidence like this suggests a *coincidence type*, in this case "having an affinity (both members of some relatively small set of people) with the stranger", where the number of possible affinities (attended first ever Star Trek convention; grow orchids; mothers named Chloe) is clearly very large and subjective. Nevertheless one could try to estimate (within a factor of 10, say) the chance of some coincidence within this coincidence type. Next one can think of many different specific single-affinity coincidences (finding a dollar bill in the street, twice in one day; seeing on television someone you know personally) which should be assigned to different types, and it is hard to imagine being able to write down a comprehensive list of coincidence types, even within the very restricted domain we're calling "single affinity". Finally, real life offers many different domains of coincidence, in particular *multiple affinity* coincidences (exemplified by the well known list[7] of asserted similarities between the assassinations of Presidents Lincoln and Kennedy); these are the mainstay of anecdotes but are harder to contemplate mathematically.

To summarize: the usual rationalist analysis of coincidences starts out by observing that estimating the *a priori* chance of some observed specific coincidence isn't the real issue; one has to think about the sum of chances of all possible coincidences. But rationalists seem to have despaired of actually doing this, and merely assert that in the end one would find that coincidences occur no more frequently than "just chance" predicts. We think this is too pessimistic an attitude; though one may not be able to prespecify all possible coincidences, surely one can learn something from observed instances?

The study we initiated consisted of noting coincidences amongst articles in Wikipedia obtained using the "random article" option. This is less "real-life" than one would like, but has the advantages of possessing the essential features (i-iii) above, while also allowing data to be gathered quickly and allowing independent replication by other people.

[7] *Lincoln-Kennedy coincidences urban legend* ($\mathcal{W}$) .

Design of study We did 28 separate trials of the procedure:

> read random articles online until noticing a first coincidence with some earlier article; record the names of the two coinciding articles and the number of articles read, and write down a phrase describing the specific coincidence observed.

"Coincidence" means some subjectively noticable close similarity in article subject or content; of course your subjective judgements might be different from mine. In principle the statistically efficient design would be to print out (say) 500 articles and carefully search them for *all* coincidences, but we are seeking to mimic real life where we notice coincidences without searching for them. We explicitly did not backtrack to re-read marterial, except to find the name of the earlier coincident article.

	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 chance that two random articles from Wikipedia would fit the specific coincidence named. The left column is trial number and the right column shows number of articles included in that trial. The total number of articles read was $1,413$. The median number of articles per trial was 44.5.

Why didn't this project work out? If one repeated the procedure, the next 28 "specific coincidences" would be almost all different from those in the table. Our goal was to formulate and list higher-level "coincidence types" so that most specific coincidences would fall into some "type"; then by counting pages in Wikipedia (using its own *lists* and *categories*) we could give a theoretical prediction of the rate of seeing coincident pages, to compare with experimental data.

We were unable to finish, partly because of the "long tail" of both types and specific coincidences within types, and partly because what a human sees as a coincidence is broader than what is picked up via such lists.

18.4 Near misses

Closely related to coincidences are a range of events that one might view as *near-misses*. That phrase originated in the setting a physicaly aiming at a target (I'll call that the *geometric* setting) but is also used in other settings I will call *combinatorial* – see examples below. The message of this section will be

> In combinatorial (rather than geometric) settings, near-misses may be much more likely than exact hits, and this phenomenon is exploited by designers of Lotto-like games.

Here is our exemplar, which will be familiar to players of Scrabble-like word games. If we pick 5 letters of the alphabet, 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. The numerical chances depend on how exactly you pick the random letters and how large your vocabulary or dictionary is, but in our small experiment chance (a) was about 18% and chance (b) was about 94%.

Near misses in geometric settings. Before trying to explain what "combinatorial settings" means, it may help (and is easier) to illustrate the opposite notion of "geometric setting". On a dartboard there is a small "bulls eye" (scoring 50 points in the traditional British game) surrounded by a ring (scoring 25 points) of twice the radius. If you have some small probability p of hitting the 50, then you will have probability about $3p$ of hitting the 25, because the area is three times larger. Similarly in the asteroid example from section 8.6, the chance an asteroid comes within 4,000 miles (the Earth's radius) of the Earth's surface will be about three times the chance of actually hitting the Earth. This is just the local uniformity principle from Lecture 8, the point being that the ratio "3" of probabilities depends only on the fact that we're dealing with a problem in two dimensions. In contrast, if we view 10 out of 10 Heads in coin tossing as a "coincidence" and 9 out of 10 as a near miss, then the ratio of probabilities is 10. But here, "10" isn't a magic number associated with coin-tossing; if we had chosen a different, rarer coincidence we would get a larger ratio.

Near-misses in Lotto picks. Instead of Scrabble or coin tossing, a more common occurrence of "combinatorial" near-misses is in Lotto-type games. If you pick 6 numbers out of 51, then when the lottery picks 6 numbers, the chance you get 5 out of 6, relative to 6 out of 6, is now $6 \times 45 = 270$ to 1. This is dramatically different from the ratio "3" we saw in geometric examples. And indeed, 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[8] 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.

As a minor course project one could study near-misses in bingo with many players – when one person wins, how many others will have lines with 4 out of 5 filled?

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

[8] See e.g. a 1986 paper by R.L. Reid *The psychology of the near miss* for discussion.

> 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.

18.5 What really has a 1 in a million chance?

This is fun to do in class. First I ask students

> If you overheard the phrase "1 in a million chance" in someone else's casual conversation, what might they be talking about?

and students typically offer both iconic examples (winning the lottery, struck by lightning) and more imaginative suggestions. Then I ask

> How could we get data on actual casual usage of the phrase "1 in a million chance"?

and neither the students nor I can think of anything much more practical than searching in blogs, some results of which were shown at the start of Lecture 1. Finally I ask for suggestions for

> events that we can convince ourselves really do have a 1 in a million chance

(up to a factor of 2, let's say). Then I go through the students' suggestions; can we quantify the chances, and (if so) are they around 1 in a million?

Here are just a few examples. The classroom 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 2007 estimate [9] puts the chance at about 1% per year, so the chance (i) is indeed around 1 in a million. Next consider

(ii) One of the next 25 babies born in the U.S. will become President.
The U.S. birth rate is currently about 4.3 million per year. If we guess a President will serve on average about 6 years, then it is reasonable to figure

[9] http://pubs.usgs.gov/of/2007/1437/

that about 1 in 6 times 4.3 million = 25 million babies will someday be President.

For many other examples one would need to rely on population percentage data. Using such data as estimates for individuals is a big topic that might be discussed in more detail in another lecture. If "you" is interpreted as "a randomly-picked 20-year-old in the U.S." then the chance

(iii) you will die (sometime) by being struck by lightning

is roughly 1 in 100,000, from population statistics. But if I point to one of my students as "you", it is not true – the chance depends so much on that individual's behavior that I cannot assess the chance, just like I can't assess the chance of you-the-reader winning the lottery sometime (I guess you buy fewer lottery tickets than the average person, but have no idea how many).

As a practical matter one can use common sense to guess how variable the chance is between individuals, and use population data when you guess it's not greatly variable (recall we are allowing a factor of two error). 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 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. 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.

18.6 How not to explain coincidences

> Being a professional mathematician, [Littlewood] . . . defined a miracle as an event that has special significance when it occurs, but occurs with a probability of one in a million. This definition agrees with our common-sense understanding of the word "miracle. Littlewood's Law of Miracles states that in the course of any normal person's life, miracles happen at a rate of roughly one per month. The proof of the law is simple. During the time that we are awake and actively engaged in living our lives, roughly for eight hours each day, we see and hear things happening at a rate of about one per second. So the total number of events that happen to us is about thirty thousand per day, or about a million per month. With few exceptions, these events are not miracles

> because they are insignificant. The chance of a miracle is about one per million events. Therefore we should expect about one miracle to happen, on the average, every month. Broch tells stories of some amazing coincidences that happened to him and his friends, all of them easily explained as consequences of Littlewood's Law.
>
> *Freeman Dyson*, in a review in the New York Review of Books.

To me, this is mind-bogglingly awful prose – an exemplar of how *not* to write for the public. That is not the usual meaning of the word miracle ("an effect or extraordinary event in the physical world that surpasses all known human or natural powers and is ascribed to a supernatural cause"), so using that word creates needless confusion. It is difficult to determine which real events have a 1 in a million chance, so invoking unspecified hypothetical events is hardly convincing. But the main point is that we are discussing a *quantitative* issue – those who assign spiritual or paranormal significance to *some* coincidences would hardly deny that "ordinary" coincidences also happen, but assert that some occur that are so very unlikely that they cannot be explained as just chance. One may believe, as part of a rationalist world-view, the assertion "amazing coincidences *might be* explicable as consequences of Littlewood's Law". But to demonstrate they are thus explicable, rather than merely assert it, would require an actual quantitative argument from real-world data.

18.7 Wrap-up

There has been considerable study of hot hands and streaks; this is the topic of Chapter 1 of Grinstead-Peterson-Snell's *Probability Tales*, which could be used as a lecture in this course. A blog by Alan Reifman[10] discusses ongoing streaks, and it's a good topic for student projects.

I am fond of the style of "back-of-an-envelope" calculations in section 18.2, and indeed my 1989 book *Probability Approximations via the Poisson Clumping Heuristic* consists of 100 examples of such calculations, within somewhat more complicated models.

[10] hehothand.blogspot.com