Mathematical Analysis of Subjectively Defined Coincidences; a case study using Wikipedia

David J. Aldous[*]
Department of Statistics
367 Evans Hall # 3860
U.C. Berkeley CA 94720
aldous@stat.berkeley.edu
www.stat.berkeley.edu/users/aldous

Fayd Shelley.

August 14, 2006

Abstract

Rationalists assert that real-life coincidences occur no more frequently than is predictable by chance, but (outside stylized settings such as birthdays) empirical evidence is scant. We describe a study, with a few real-life features, of coincidences noticed in reading random articles in Wikipedia. Part of a rationalist program (that one can use specific observed coincidences to infer general types of unobserved coincidence and estimate probabilities of coincidences therein) can be examined in this context, and fits our data well enough. Though this conclusion may be unremarkable, the study may provide guidance for the design of more "real-life" studies of coincidences.

`xxx in progress - on Fayd's desk!`

[*] Research supported by N.S.F Grant DMS-0203062

1 Introduction

A long and continuing tradition outside mainstream science [1, 3, 5] 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 well known *birthday paradox* [7]) 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 paradox 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;
(iii) what constitutes a coincidence between two events depends very much on the concrete nature of the events.
In this paper we seek to take 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 Diaconis-Mosteller [2]. Our "gazillions" explanation, which they call the *law of truly large numbers* and which is also called *Littlewood's law* [9], 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, mention the extensive list of coincidences recorded by Kammerer [4], 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.

The aspect we study is *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 [8] 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 in this paper, described in the next section with some details postponed to section 3, 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. How this particular study relates to the general considerations above will be discussed in xxx.

2 The study

About Wikipedia Wikipedia is an online encyclopedia in which anyone may edit existing articles or create a new article. Readers unfamiliar with it should simply experiment for a few minutes. Briefly, the kind of article topics are
(a) traditional print encyclopedia topics (every academic discipline; biographies; general reference material)
(b) popular culture, e.g. movies, TV shows, actors; musicians and groups; professional sports players; video games
(c) stereotypical nerd topics, e.g. obselete hardware and software; U.K. railway stations.

	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. As described in section 3.1, certain types of articles were excluded.

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.

See Table 1 for the results. “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.

Analysis The first step in our analysis is to assess the probability of each specific coincidence. In some examples this is easy by using lists (see section 3.1 for remarks regarding Lists and Categories) within Wikipedia. In trial 7 Loppington and Beckjay are both villages in Shropshire (U.K.). Wikipedia has a Category: Villages in Shropshire which lists 193 articles including these two, The effective number of Wikipedia articles for our purposes (see section 3.1) is 0.94 million. So we estimate the chance of this specific coincidence (for two random articles) as $(193/940000)^2 = 4.2 \times 10^{-8}$. Note the $\times 10^{-8}$ scaling in Table 1. More commonly the two articles are in related lists; for instance (trial 1) Kannappa is in Category: Hindu religious figures and Vasishtha is in Subcategory: Hindu sages. The majority of examples in Table 1 can be done using a few lists, though some require rougher estimation. For instance (trial 20) Wikipedia shows about 2,000 articles linking to the Julius Caesar article, but most are too tangential; we estimated that about 100 people articles and 100 event articles are sufficiently close that a typical pair would be noticed as a “linked to Julius Caesar” coincidence. In two trials we couldn’t do an estimate because our original description of the specific coincidence was too fuzzy, and we stick to the protocol of not using hindsight to revise the description.

The probabilities in Table 1 illustrate the range of probabilities for the specific coincidences observed. These chances are *not* used in our main statistical analysis, though do imply (section 3.2) that in a further 28 trials we would expect about 26 further distinct specific coincidences and only 2 repeats of the observed specific coincidences. It is conceptually important that we never use Wikipedia lists to *define* coincidences, merely as a counting aid in the estimation of probabilities.

The empirical coincidence rate From the study data (last column of Table 1) on number of articles read until a coincidence is noticed, it is straightforward to derive an estimate of the underlying coincidence rate

$$\lambda = \text{chance of a coincidence betwen two random articles} \qquad (1)$$

and we find $\lambda = 7.2 \times 10^{-4}$. See section 3.2.

Coincidence types We now arrive at the main issue: is it really possible to go from observed "specific coincidences" to identify "coincidence types" in such a way that we can roughly estimate the chance of a coincidence within the type, and of course where different types don't overlap much. Table 2 illustrates what we did with our 28 specific coincidences, generalizing them to derive xxx coincidence types. Here we used a strict protocol. Author Aldous derived Table 1 using his subjective judgements, and passed it to author Shelley, who used introspection and common sense to write down a detailed description of each coincidence type. After that, Shelley used lists within Wikipedia to estimate the chance of some coincidence within each type. As above, the point of the protocol is to mimic real life, where we cannot use lists to define coincidences.

xxx insert Table 2

As an example, the detailed description of type xxx was (xxx insert) and the Wikipedia lists used were (xxx insert). Corresponding information for the other types is recorded online at xxx.

2.1 The main result

xxx bottom line is the sum of probs associated with all these coincidence types

3 Details and analysis

3.1 Details of the study

1. By *Wikipedia* we mean the English language version of Wikipedia. At the time (August 2006) the study was conducted, it had 1.3 million articles. About 12% of articles we found were lists or other "non-content" pages, which we discarded; another 16% referred to topics in contemporary pop culture (TV shows, music, video games) which Aldous felt unqualified to judge coincidences in, and were discarded. So the effective size of Wikipedia for the purposes of our study is around 72% × 1.3 million = 0.94 million articles:

$$N_{pop} = 0.94 \text{ million}.$$

2. It is important that we did not pre-calibrate our subjective judgement of coincidences to specific statistical knowledge. For instance, we regarded as coincidences (trial 2) football clubs in England and (trial 19) cities and townships in Minnesota without realizing there were over $1,000$ of the former and $2,500$ of the latter. The aim was to mimic the real-life situation where one lacks statistical knowledge. We terminated after 28 trials for psychological reasons, suspecting that we started to unconsciously overlook "boring" types of coincidence that would have been noted earlier.

3. Confusingly for novices, Wikipedia has distinct concepts of *Category* and *List* which on a given topic often overlap without coinciding. We used both in estimating probabilities.

4. A readily analyzable "small universe" setting (e.g. math papers in the A.M.S. *Mathematics Subject Classification*; library books in the Dewey Decimal Classification) is

> randomly sample items which have a preexisting hierarchical classification; declare a coincidence if two items are in the same bottom-level class.

Our study is different, partly because Wikipedia doesn't have this kind of fixed structure (it's more like *miscellanized piles*, in a phrase of David Weinberger) and partly because we insist on subjective judgements of coincidence..

3.2 Some statistical analysis

5. The natural approximation for the distribution of

$$T = \text{ number of articles read until a coincidence is noticed}$$

in terms of λ at (1) is

$$P(T > n) \approx \exp\left(-\lambda\binom{n}{2}\right). \tag{2}$$

Indeed, the abstract mathematical structure is

$$T = \min\{m : (\xi_i, \xi_m) \in A \text{ for some } 1 \leq i < m\}$$

$$(\xi_i) \text{ i.i.d. } S\text{-valued}, \quad A \subset S \times S, \quad \lambda = P((\xi_1, \xi_2) \in A)$$

and we recognize (2) as a consequence of the Poisson limit theorem for U-statistics [6]. Explicit bounds for the approximation (2) could be derived from explicit bounds in that limit theorem (xxx refs).

We estimated λ from the observed median value (44.5) of T in Table 1, giving

$$\lambda = 7.2 \times 10^{-4}.$$

xxx graph

6. In the entire trial (1, 413 articles) we would expect about $\frac{1}{2} \times 1,413^2/N_{pop} \approx$ 1 articles to appear twice. We didn't try to record this information, but we did notice one article (The Tornados) appearing twice in a single trial (trial 15); such an event (naively, a "one in a million" event) happening during the study had prior probability about 4%.

Table 1 suggests that the particular "NFL players" case will arise as the specific coincidence about $\frac{3000 \times 10^{-8}}{\lambda} = \frac{1}{24}$ of the time. So if we continued the study for another 28 trials then we would expect this specific coincidence to recur about $\frac{1}{24} \times 28 \approx 1$ time. The sum of probabilities of other specific coincidences in Table 1 is about 2500×10^{-8} so similarly we would expect about 1 other specific coincidence to recur; in other words, we might get about 26 new specific coincidences.

3.3 Discussion and Conclusion

xxx invite reader to repeat xxx state predictions if you

xxx maybe obvious real world has many more coincidence types; send us a list of 200!

xxx multiple affinities hard to study because of dependence, cf. forensic DNA testing

xxx in obvious lists make for boring instances

xxx power laws etc

xxx number of unseen species

References

[1] Phil Cousineau. *Soul Moments: Marvelous Stories of Synchronicity-Meaningful Coincidences from a Seemingly Random World.* Conari Press, 1997.

[2] Persi Diaconis and Frederick Mosteller. Methods for studying coincidences. *J. Amer. Statist. Assoc.*, 84(408):853–861, 1989.

[3] Carl G. Jung. *Synchronicity: An Acausal Connecting Principle.* Princeton Univ. Press, 1973.

[4] P. Kammerer. *Das Gesetz der Serie: eine Lehre von den Wiederholungen im Lebens-und im Weltgeschehen.* Deutsche Verlags-Anstalt, 1919.

[5] Arthur Koestler. *The Roots of Coincidence.* Random House, 1972.

[6] Bernard Silverman and Tim Brown. Short distances, flat triangles and Poisson limits. *J. Appl. Probab.*, 15(4):815–825, 1978.

[7] Wikipedia. Birthday paradox — wikipedia, the free encyclopedia, 2006. [Online; accessed 1-August-2006].

[8] Wikipedia. Lincoln Kennedy coincidences (urban legend) — wikipedia, the free encyclopedia, 2006. [Online; accessed 2-August-2006].

[9] Wikipedia. Littlewood's law — wikipedia, the free encyclopedia, 2006. [Online; accessed 2-August-2006].