

On Chance and Unpredictability: lectures on the
links between mathematical probability and the real
world.

David Aldous

April 15, 2015

Contents

Preface	vii
1 Everyday perception of chance	1
1.1 References to chance in blogs	1
1.2 Queries to the search engine Bing	4
1.3 Comparison with imaginary data	6
1.4 Mathematical probability textbook examples	7
1.5 Wrap-up: which side of the Looking-Glass is the fantasy? . .	11
2 The Kelly criterion for favorable games: stock market investing for individuals	13
2.1 Some semi-historical data	13
2.2 Some conceptual background	14
2.3 Probability and financial investment	17
2.4 A simulation demo	19
2.5 The IID model and the Kelly criterion	20
2.6 Mathematics of the Kelly criterion: one risky and one safe asset	21
2.7 What about the not-so-long term?	25
2.8 Can one execute this theory?	26
2.9 Geometric Brownian motion and the Black-Scholes formula .	28
2.10 Mean-variance analysis	29
2.11 On investment advice	31
2.12 Wrap-up	32
2.13 Further reading	34
3 Coincidences, near misses and one-in-a-million chances	35
3.1 The birthday problem and its relatives	35
3.2 Using the Poisson approximation in simple models	37
3.3 Coincidences in everyday life	38

3.4	Coincidences in the news	41
3.5	Coincidences in Wikipedia	43
3.6	Classifying coincidences in everyday life	45
3.7	Near misses	46
3.8	What really has a 1 in a million chance?	48
3.9	How not to explain coincidences	50
3.10	Hot hands	51
4	Prediction markets, fair games and martingales	53
4.1	What is a prediction market?	53
4.2	Some data: baseball and elections	54
4.3	Fair games and martingales	55
4.4	Prediction markets and martingales: conceptual issues.	57
4.5	Theoretical predictions for the behavior of prediction markets	59
4.6	Were there improbably many candidates for the 2012 Repub- lican nomination whose fortunes rose and fell?	61
4.7	The halftime price principle.	64
4.8	Other martingale calculations	66
4.9	Stock markets and prediction markets	68
4.10	Wrap-up	68
4.11	Further reading	69
5	Game Theory: Introducing Nash equilibria via some actual data	71
5.1	The specific game: Dice City Roller	72
5.2	Analysis of the simplest case	73
5.3	General numbers of players and items	77
5.4	Comparing data from the DCR game with NE theory	78
5.5	Dice City Roller	81
5.6	The Least Unique Positive Integer game	83
5.7	Further reading	84
6	Short and medium term predictions and risks in global pol- itics and economics.	87
6.1	Some conceptual issues.	88
6.2	The annual WEF Global Risks assessment	89
6.3	The Good Judgment Project	95
6.4	More about the GJP	98
6.5	The cost of errors in assessing probabilities	100

7	Coding and entropy	103
7.1	Introduction	104
7.2	Entropy as a measure of unpredictability	104
7.3	Coding, compression and encryption	106
7.4	The asymptotic equipartition property	108
7.5	Entropy rate and minimum code length	110
7.6	Morse code and ASCII	111
7.7	Lempel-Ziv algorithms	112
7.8	Checking for yourself	113
7.9	...but English text is not random	114
7.10	Wrap-up and further reading	115
8	From physical randomness to the local uniformity principle	117
8.1	A glance at physical randomness	117
8.2	Dart throws as a simple example of physical randomness	119
8.3	County fair	121
8.4	The physics of coin-tossing and the fine-grain principle	123
8.5	The smooth density idealization for data and Benford's law	126
8.6	Which of 4 foundational principles do you believe?	133
8.7	The local uniformity principle and asteroid near-misses	136
8.8	Wrap-up and further reading	137
9	A glimpse at research: spatial networks over random points	139
9.1	Regular networks	139
9.2	Networks on random points	141
9.3	An optimality criterion for road networks	144
9.4	Scale-invariant random networks	146
10	The other lectures	151
10.1	Psychology of Probability: Predictable Irrationality	151
10.2	Science fiction meets science	152
10.3	Ranking and rating	154
10.4	Risk to Individuals: Perception and Reality	155
10.5	Tipping points and phase transitions	156
10.6	Luck	156
10.7	Toy models in Population Genetics: some mathematical aspects of evolution	158

Preface

This book is an extended write-up of some lectures in a course I have given in Berkeley every few years since 2002, most recently in Fall 2014. It is part of a broad project described on my web site [Probability in the Real World](#). In particular the site contains much more discussion of the goals outlined below, of the three lists mentioned below, and of the type of “Perception” material in Chapter 1.

My overall goal, compressed to 11 words, is

to articulate critically what mathematical probability says about
the real world.

By *critically* I mean with reference to actual evidence, not just repeating mathematical or rhetorical arguments – what would stand up under cross-examination in the witness box? Hidden behind *what* is an emphasis on the extraordinary *breadth* of contexts where we perceive chance or uncertainty, while seeking to understand in which of these contexts mathematics is useful. And by *real world* I mean “not mathematics or philosophy or fantasy examples”, though I do include issues of perception.

The book is addressed to two opposite audiences. First, the readership of serious popular science books. As a hobby I have read and [reviewed every \(around 100\) non-technical book on probability](#) I’ve found, so I can claim some familiarity with the genre. An author, bearing in mind the dictum *every equation will cut sales in half*, faces an immediate decision of how much math to include. Books without any math strike me as too fuzzy, tending to flit to a new topic every other page without leaving any bottom line conclusion. At the other extreme are what I call “textbook lite” works, which seek to teach a little math while recounting the more interesting parts of undergraduate courses plus popular topics such as the Monty Hall problem. These strike me as far too narrow. I adopt an intermediate policy of quoting some mathematics – model descriptions and formulas – without trying to teach readers how to do the math for themselves.

The second audience, of course, is students and their instructors. The Berkeley lectures are to undergraduates majoring in Statistics or some similar subject. These students already know the basic mathematics of probability, and a handful of standard textbook applications (games of chance, opinion polls, etc). But the difficulty in teaching any sort of applied math is that in practice the math takes over. Look at the actual content of any book with a title like *Applied Probability*, and you will see it is overwhelmingly devoted to developing the mathematics, with very scant discussion of the significance (let alone the realism) of the mathematical conclusions. Advanced college courses lead to more specialized and mathematically oriented topics, rather than to *breadth*. I doubt that any student anywhere else has been exposed in college courses to more than half of the topics in this book.

My Berkeley course consists of 20 lectures, on topics chosen to be very diverse. Here are my desiderata for an ideal topic.

- It is appropriate for the target audience: those interested in the relation between mathematics and the real world, rather than those interested in the mathematics itself¹.
- There is some interesting data-set that one can show.
- There is some concrete bottom line conclusion, which can be said in words ...
- ...but where some mathematics has been used to derive conclusions ...
- and where the mathematics leads to some theoretical quantitative prediction that my students can test by gathering fresh data.
- There is available “further reading”, both non-technical and technical, that I can recommend to students.

Very few topics permit all this, so many of the actual lectures fail to attain the ideal. This book contains extended versions of 10 lectures, on the topics for which I believe there are no comparable accounts elsewhere. The contents of the other lectures are briefly described in chapter 10 and cross-referenced in format “Lecture 10.1”: they tend to consist of material available elsewhere, and to be less mathematical. The lectures are fairly independent of each other, so the order is fairly arbitrary.

¹For reasons peculiar to Berkeley, few undergraduate majors in (theoretical) mathematics take any course in probability or statistics.

I have used the word *breadth* many times, and this relates to another part of the project, an attempt to categorize all the **contexts where we perceive chance in the real world** into around 100 categories. I don't know any comparable attempt, and it provides a testbed for evaluating asserted generalizations, mathematical or philosophical or whatever, about the nature and scope of probability.

A final Chapter ?? (which I do not lecture on) is called *Some Conceptual Issues in Probability*. I am critical of typical discussions of the Philosophy of Probability (in the popular sense, not narrow technical academic philosophy). because of its reliance on unrealistic simplistic or fantasy examples, its recycling of a small set of classical or fashionable issues and a general lack of appreciation for the *breadth* of the subject. So Chapter ?? discusses a less standard set of issues arising from thinking about substantive real-world questions. On this theme, the practical distinction between frequentist and Bayesian views is much less than many popular accounts suggest, and I use one or other without comment. The rare contexts where there *is* a difference are interesting for that reason, so I do comment there.

Most of these lectures are intended to be uncontroversial accounts of a few aspects of some established academic topic, colored by my own taste in what is interesting or significant. Note they are not intended as a descriptive summary of the main points of the topic – that's what Wikipedia is for. For some topics, what I say in class is similar to material already written by someone else. It is pointless to rewrite such material in one's own words. So I don't – I refer the reader to the existing material (a policy which would have spared many forests if systematically adopted by all authors). Consequently, what appears in this write-up is intended to have little overlap with any existing writings at the same length and technical level.

Let me reiterate my goal in more rhetorical fashion. We live in several overlapping worlds. There's the natural world that exists independently of humans; the human social world; the world of human artifacts; the world of ideas and perceptions and motivations. Outside the classroom, we know that chance enters all these worlds in many ways. From the way we first meet a future spouse to the spatial fluctuations in mass density of the early universe that led to galaxy formation; from the chance of Scotland voting for independence in September 2014 to the event that Kokura was covered by clouds on August 9, 1945. Inside the classroom we forget all this, and revert to the mathematical tradition that implicitly views probability as about things that are similar to dice. So what is the connection?

Some details. In writing up lectures² I have stayed close to what I want to say in an actual lecture. One issue is that what is acceptable as a brief spoken explanation may look embarrassingly vague when written. So sometimes I have added a more careful explanation in the text, sometimes in the Notes at the end, and sometimes I’ve just left it vague.

Students in the course do a “course project”, preferably involving real data. I give many suggestions for projects in lectures and post others online. I have included some in this write up to give their flavor. “Small project” means less substantial, and “research project” means more substantial, than a course project. I tell students about many resources – books, papers, web sites – that may be helpful; again, some are included here. The order of lectures is different in the course, because I want the topics most amenable to course projects to be done early.

I am a fan of Wikipedia as a basic factual reference, outside of sophisticated mathematics, and many of the in-text links are to a Wikipedia page which may be helpful or interesting. Occasionally I use the format “Wikipedia **Spatial network** not helpful”, which is an implicit suggestion to edit that entry. References to books are typically given in-text and to papers³ as footnotes, but note these references are to “further reading for the target audience”, not academic citations to the originators of ideas.

Students in my lecture course have taken a course in mathematical probability, so I do not explain the basic mathematics from first principles but instead give “reminders” of theory. Students are inclined to view mathematics as symbolic manipulation of the $(x-y)(x+y) = x^2 - y^2$ kind; throughout I remind them that our mathematics is intended to refer to something in the real world. So I emphasize conceptual aspects – what do the assumptions mean, what do the conclusions mean – rather than the internal mathematical arguments leading from assumptions to conclusions.

Acknowledgements. Responding to criticism that in writing *I, Claudius* he had merely run together Tacitus and Suetonius’s works and added his own “vigorous fancy”, Robert Graves asserted that in fact he also borrowed from the writings of Cassius Dio, Pliny, Varro, Valerius Maximus, Orosius, Frontinus, . . . [26 in total]. This book of mine has some analogous cut-and-paste material and also draws, consciously or unconsciously, on interactions with many colleagues over many years, in particular Persi Diaconis.

²By L^AT_EX default *Lectures* are set as *Chapters*, though I refer to them in the text as lectures.

³In general I only cite papers that are readily available online.

Chapter 1

Everyday perception of chance

In what contexts do you think of elephants? I suspect you can't answer very confidently, partly because now I have put the idea of elephants into your mind with this question, it's hard to remember the previous times. Our topic today is

In what everyday contexts do “ordinary people” perceive events in terms of chance?

Just ask them! isn't a helpful way to try to answer – we humans do forget.

There is substantial academic research relating to perception of probabilities, which will be the topic of our Lecture 10.1 on “psychology” later. But this typically studies responses when subjects are *prompted* to think about chance by being asked some specific question which plainly involves chance. Our topic today is: when do you think about chance, *unprompted*?

Of course it's easy to imagine many contexts, but a main theme of this course is to look at actual data, not imaginary data:

Don't Make Stuff Up!

I will talk about two sources of actual data (and then, as evil fun, compare with some imaginary data). The first is a data collection exercise done by undergraduates that you could easily repeat yourself.

1.1 References to chance in blogs

In 2009 we did an online search through personal blogs – people writing about their everyday life and thoughts – searching for the phrase *one in*

a million chance. Below are the first 22 instances found (literal text italicized, without correcting spelling or grammar). I have divided them into 4 categories.

1. Past events that happened to writer. (6)

Finding a romantic partner. *There was this weird connection that I felt when I first met him . . . Seeing how its like a one in a million chance to find that one person you connect with.*

(similar quote omitted).

Major life events. *I have . . . syndrome. The fact that I ever became a mother was a “one in a million chance”.*

Unusual dramatic events. *. . . and they [adults] all start talking about how im too young to be going out by myself . . . But it’s not like im going to listen to them, what happened [witnessing a mall shooting] was a once in a million chance.*

Unusual minor events.

(vacation went unexpectedly well: quote omitted)

(throwing chips in drunken party: quote omitted)

2. Possible future events that might affect writer. (8)

Minor pleasant possibilities. *I’m somewhat hoping to meet friends there . . . it’s a one in a million chance.*

i’m waiting for the day they [upcoming movie/TV filming locations] say my city which is one in a million chance

. . . got this contest. It’s a one in a million chance to get some people . . . to tell me what they think of my work.

my greatest ambition is to see [a supernova] one day, though there’s probably a one in a million chance that i will. smaller than that.

I’d only be satisfied with one particular scenario and there’s maybe a 1 in a million chance of that happening . . . no, less. I would get struck by lightning before that happened, twice

Worries. *Of course if I don’t go [to the doctor about certain symptoms], there’s that one in a million chance that I’ll be sorry I didn’t.*

(similar quote omitted)

There are things that we were never told that really end up happening to most women [during pregnancy]. Instead we were told the things that we had a one in a million chance of experiencing.

3. Events affecting specific other people. (2)

On the one in a million chance that Christine actually gets hired to do costumes for ...

There is a one in a million chance that [a particular NHL player] gets picked up on waivers

4. Impersonal speculation. (6)

On the other hand, if you chase after it [a volleyball spike by opponents], who knows? It might just be one-in-a-million chance that you'll get it, but isn't that a chance worth taking?

When you're looking for the one in a million chance of getting a Beethoven you could be overlooking an Einstein.

Becoming a successful actor, singer, or dancer is a one in a million shot in the dark during a snowstorm.

(similar quote omitted)

... reflect on how [Valentine's day] has brainwashed a whole lot of people into believing that love could actually happen on that day, which is a one in a million chance by the way (which they would argue is worth the risk anyway, which is also bullshit, by the way).

First [one particular sperm] had to survive and beat out millions of other sperm ... that's like winning the lotta right there ... only one in a million, and from that point, you got to survive ...

Here are the results of [searches on other phrases](#).

Once one sees this kind of data, it may seem obvious that “this is the sort of data we expected to see”. But actually predicting such things is hard. I challenge readers to stop after the first paragraph of the next section, and try to predict what the data will show! The four particular categories above were suggested by this particular data-set, and are not really useful ways to categorize “contexts where we perceive chance”.

I suggest as a course project that students gather more data using other search terms and other regions of the online world. For this I insist on a *repeatable experiment*. You must show (a random sample of) results of a specified search, not human-selected ones. The internet is so big that one could invent examples and then search for similar ones, so *selecting* examples is little different from *inventing* them. Specifying a search protocol that gives the kind of “contexts where we perceive chance” examples is harder than it sounds.

It is important to note that I are not claiming that what we find is a statistically accurate sample of “contexts where we perceive chance”, either in

general or specifically within the blogosphere. Our purpose, more modestly, is to illustrate actual usage, as opposed to made-up examples.

1.2 Queries to the search engine Bing

In 2010 I obtained, from the Bing team, a file of all (around 100,000) queries made to Bing containing the strings “chance of” or “probability of”. After excluding those which were not actually looking for the chance of something (e.g. were seeking the movie *Cloudy with a Chance of Meatballs*) I had enough patience to examine 675, sorting them into 66 groups of about 10 similar queries. Picking one from each group gives this [sample of 66 “representative” queries](#). Here I prune down further to a representative 30.

Before turning the page, I challenge readers to predict what the data will show!

- Query: what's the chance of getting pregnant after tubal litigation?
- Query: chance of pregnancy after intercourse
- Query: how to improve chance of getting pregnant
- Query: percent chance of getting pregnant with clomid
- Query: chance of getting pregnant while breastfeeding
- Query: if twins run in my family whats my chance of having them?
- Query: chance of having multiples using fertility
- Query: chance of siblings both having autism
- Query: chance of miscarage after 8 weeks
- Query: chance of bleeding with placenta previa
- Query: any chance of vaginal delivery if first ceaserian
- Query: probability of having an adverse reaction to amoxicillin
- Query: can aispirin reduce chance of a stroke
- Query: does progesterone increas chance of breast cancer
- Query: which treatment has the least chance of prostate cancer recurring?
- Query: chance of getting a brain tumor
- Query: do chargers have a chance of making the playoffs
- Query: probability of flopping a set with pocket pair
- Query: does a ring of wealth affect the chance of the dragon pickaxe drop in runescape?
- Query: percent chance of getting shot if you run from an attacker
- Query: chance of surviving severe head injury
- Query: chance of having white christmas ontario
- Query: chance of rain in september dallas texas
- Query: what are the chance of becoming a golf pro
- Query: chance of closing airports in mex because of swine flu
- Query: chance of getting a short sale
- Query: probability of winning a traffic ticket court case
- Query: chance of food spoiling if left out over night
- Query: probability of life and evolution
- Query: wich technology has the least probability of a collision

Discussion. The best descriptive phrase I can devise for these examples of queries is that they are *personal and concrete*: they typically concern near-future events with substantial significance to the person involved. The examples from blogs are loosely similar, though (as “obvious” from the nature of blogs) people also write about past personal events and general thoughts.

It is striking that around half the queries concern medical matters, and more than half of those concern pregnancy and birth control. Of course this data derived from search queries is several steps removed from the conceptual question “how do people think about chance in everyday life”. For instance people may often be interested in the chance of rain tomorrow, but few of them will type into a search box “what is the chance of rain”. So we certainly do not suggest a quantitative correspondence like “half of our everyday perception of chance involves medical matters”. Indeed a further data-set of [examples from Twitter](#) reveal a substantially different range of topics of interest, with (for instance) only 3% concerning medical matters. In retrospect such differences are to be expected, in that Twitter supports casual comments on momentary concerns.

I used the three sources of “real data” above in compiling the “everyday life” section (around contexts 10 - 20) of the [list of contexts where we perceive chance](#). Of course it would be good to find many more sets of data with different origins; so please send me suggestions or do the data collection yourself!

1.3 Comparison with imaginary data

I am confident it is easy to distinguish between real and imaginary examples. Here is a small project. Find two friends who have not seen this material. Ask one to imagine and write down ten instances of how one might use the phrase “one in a million chance” in a blog, or ten instances of “chance of” queries one might type into a search engine. Give this list, and a sample of 10 examples from our earlier lists, to the other friend. I bet the other friend will unhesitatingly identify the real list..

An intriguing source of imaginary examples is the book *Luck: The Brilliant Randomness Of Everyday Life* by Nicholas Rescher, a former President of the American Philosophical Association. This short book could be viewed as an unusually erudite blog, or as an unusually reader-friendly monograph. And the content of his musings about Luck is perfectly reasonable. But what interests me here is the examples he cites. Here is my list of [all the examples from the parts \(Introduction; Chapters 1 and 3\) closest to our “everyday life” theme](#). I categorize them as

- Specific historical events (11)
- Iconic headlines (4)
- Conventional examples of luck (20)
- Notes for a historical novel? (23)

All except the first category consist of invented examples. Here are the last 12 examples from the fourth category. Italics are exact quotes, others are paraphrases.

potential victim saved because a would-be assassin missed the bus being wounded by an assassin who mistakes one for someone else
 injured as bystander in political demonstration
you were inadvertently delayed and just missed crossing on the Hindenberg
hit by falling icicle
 fighter pilot hits ejector button instead of defroster
burglar who breaks into a house just before its owner returns well-armed from a bear hunt
the painter who produces a [long-sought] effect . . . by throwing his brush at the picture in a fit of rage . . .
coming down with a disease for which a cure has just been discovered
author whose biography of a celebrity hits the bookshops just as its protagonist is enmeshed in a highly publicized scandal . . .
 scam victim accidentally profiting
the winner of a lottery who decides to build a dream cottage on Krakatoa

In class, I cannot resist saying “Wow, Everyday Life in a Philosophy department sure seems more exciting than in a Statistics department” and then comparing to the real “everyday life” revealed by the type of data in the previous sections.

Now in one sense I am merely being humorous. You and I both know the author did not intend his examples to be literally “everyday life”; instead, he was interested in the abstract ideas surrounding *luck* and just made up illustrative hypothetical examples as he wrote.

But in another sense I am perfectly serious. The author is adopting a style of intellectual enquiry where he starts with abstract ideas and then invents hypothetical examples to justify the ideas. To what extent is this a useful style of intellectual enquiry?

1.4 Mathematical probability textbook examples

Note: *This section is here for the “popular science” reader, since it’s hardly news to my students.*

What is the picture of chance – that is, of the contexts where chance arises – that one obtains from the examples and exercises in an undergrad-

uate mathematical probability textbook? I would put them into 4 style categories, below. Illustrative exercises are taken from [Grinstead-Snell *Introduction to Probability*](#), which I regard as one of the best textbooks.

1. Purely mathematical.

Let X_1, X_2, \dots, X_n be n mutually independent random variables, each of which is uniformly distributed on the integers from 1 to k . Let Y denote the minimum of the X_i 's. Find the distribution of Y .

2. An (at least somewhat) interesting real-world question and an (at least somewhat) realistic model.

A large number, N , of people are subjected to a blood test. This can be administered in two ways: (1) Each person can be tested separately, in this case N test are required, (2) the blood samples of k persons can be pooled and analyzed together. If this test is *negative*, this one test suffices for the k people. If the test is *positive*, each of the k persons must be tested separately, and in all, $k + 1$ tests are required for the k people. Assume that the probability p that a test is positive is the same for all people and that these events are independent.

For small p , show that the value of k which will minimize the expected number of tests under the second plan is approximately $1/\sqrt{p}$.

3. Actions one could do, but with no evident purpose.

A die is rolled 30 times. What is the probability that a 6 turns up exactly 5 times?

4. A real-world story with invented data and/or a very unrealistic model. This is my main concern, so let me give several examples.

a. A student must choose exactly two out of three electives: art, French, and mathematics. He chooses art with probability $5/8$, French with probability $5/8$, and art and French together with probability $1/4$. What is the probability that he chooses mathematics? What is the probability that he chooses either art or French?

b. A restaurant offers apple and blueberry pies and stocks an equal number of each kind of pie. Each day ten customers request pie. They

choose, with equal probabilities, one of the two kinds of pie. How many pieces of each kind of pie should the owner provide so that the probability is about .95 that each customer gets the pie of his or her own choice?

c. Take a stick of unit length and break it into two pieces, choosing the break point at random. Now break the longer of the two pieces at a random point. What is the probability that the three pieces can be used to form a triangle?

d. Suppose you toss a dart at a circular target of radius 10 inches. Given that the dart lands in the upper half of the target, find the probability that

1. it lands in the right half of the target.
2. its distance from the center is less than 5 inches.
3. its distance from the center is greater than 5 inches.
4. it lands within 5 inches of the point $(0, 5)$.

e. You are in a casino and confronted by two slot machines. Each machine pays off either 1 dollar or nothing. The probability that the first machine pays off a dollar is x and that the second machine pays off a dollar is y . We assume that x and y are random numbers chosen independently from the interval $[0, 1]$ and unknown to you. You are permitted to make a series of ten plays, each time choosing one machine or the other. How should you choose to maximize the number of times that you win?

f. A small boy is lost coming down Mount Washington. The leader of the search team estimates that there is a probability p that he came down on the east side and a probability $1 - p$ that he came down on the west side. He has n people in his search team who will search independently and, if the boy is on the side being searched, each member will find the boy with probability u . Determine how he should divide the n people into two groups to search the two sides of the mountain so that he will have the highest probability of finding the boy. How does this depend on u ?

Discussion. My 4 categories are rather fuzzy – do “drawing balls from urns” problems fit category 1 or 3? So I haven’t tried to examine textbooks to find the percentages in each category. But I suspect every introductory probability textbook has an extremely low percentage in category 2. Good textbooks on *statistics* do rather better. For instance Freedman - Pisani -

Purves - Adhikari *Statistics* has about 260 examples with cited sources and another 100 for which one could easily find data of the given type¹.

What's unrealistic about the examples in category 4 is (I hope) clear to the reader. In (a) people don't choose at random, and (b) breaks the commandment that should be given in a first class on probability: thou shalt not assume different possibilities are equally likely, without some darn good reason. In (c) I suspect it's physically impossible to break at a uniform random point, and in (d) the implied uniform distribution only applies if you are *very* bad at throwing a dart. The casino in (e) would rapidly go bankrupt, and in (f) it's hard to justify independence.

And of course the authors know these exercises are unrealistic, just as the philosopher author knew his examples weren't really "everyday life".

Now I must admit that in teaching such courses, I use the same style of examples as do the textbooks. Indeed, part of the reason for teaching this completely separate course is that it's too hard to put realistic material into a conventional mathematics-focused course. But what *should* one do in a first course, ideally? Certainly one needs some purely mathematical examples to illustrate math techniques, and the justification for phrasing examples in terms of dice and urn is to provide a more concrete visualization than would a purely mathematical formulation.

On the other hand, if you look at a textbook for a course being taught on (say) history or biology, you will see page after page of declarative sentences, and – amazingly enough – there is nothing that the author has “just made up” in the style of our textbook examples.

Returning to our [list of 100 contexts](#), in a typical probability textbook you will find extensive occurrence of the context “explicit games of chance based on artifacts with physical symmetry” and briefer occurrences of “random sampling for representativeness” and a few others. This – in my opinion – indicate the huge disconnect between introductory textbooks on probability and the big picture of the role of chance.

On a positive note, the best instance I know of academics engaging broad topics of popular interest is the web site [Understanding Uncertainty](#) from Cambridge U.K. The site is centered around issues of health (recall our Bing data showed that about half the searches were related to health) but covers other popular angles (coincidences, lotteries etc) and gives commentaries on risk and statistics items in the news. The latter is the focus of another valuable site, [Chance News](#).

¹Though I am amazed how many introductory statistics texts have minimal and stereotyped real data.

1.5 Wrap-up: which side of the Looking-Glass is the fantasy?

A common view amongst academics involved with probability is that ordinary people are pretty dopey when it comes to understanding randomness. They are superstitious about luck or coincidences. Under irresponsible media influence they have a completely distorted notion of which risks in everyday life are substantial and which are negligible. They gamble when the odds are against them (horses, lotteries, casinos) but not when the odds are in their favor (increase your insurance deductible). They waste billions on useless stock market advice or managed mutual fund fees. They misconstrue positive medical diagnostic test results for rare diseases. And so on.

Now all this is true. But it's one side of a picture. What about the academic side? I should clarify that I am discussing *the impression given by textbooks and papers written by academics*, not what individual academics think. The elements of fiction we have seen in the textbook math exercises and in the examples of luck are mostly harmless in themselves, but set a tone that invites one to approach the world with a *suppose*:

suppose these events are independent

This starts down a slippery slope, liable to end with the implicit belief one can learn something about the real world primarily from setting up and studying models, rather than primarily from data and experiment. We will return to this theme several times during these lectures.

Chapter 2

The Kelly criterion for favorable games: stock market investing for individuals

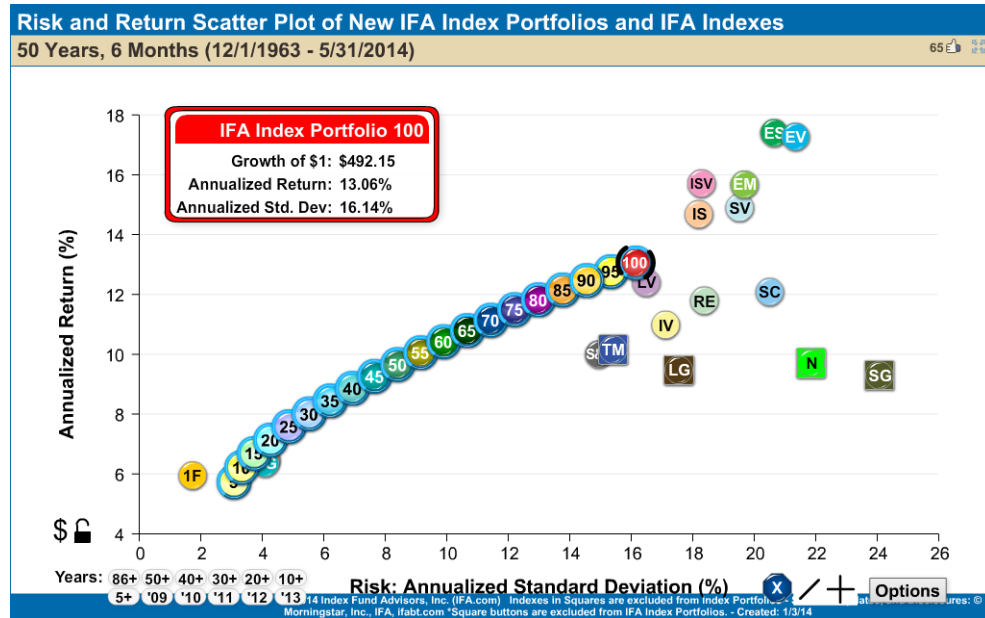
Finance provides inexhaustible data and is a popular topic for student projects. This lecture focusses on the Kelly criterion, a topic both intellectually interesting and useful to you as a real-life amateur investor, but curiously neglected in textbooks. In class I jump quickly to section 2.4; the previous material is intended for the general reader.

2.1 Some semi-historical data

There are many web sites that provide combinations of historical data, hypothetical data and (not so mathematical) theory relevant to this topic. Figure 2.1, from the website of IFA (Index Funds Advisors) at which you can find a [current version](#)¹ and much more graphical data. IFA and similar sites start out by trying to assess the individual's subjective risk tolerance using a questionnaire. The site then suggests one of a range of 21 portfolios, represented on the slightly curved line in the figure. The horizontal axis shows standard deviation of annual return, (3% to 16%), and the vertical axis shows mean annual return (6% to 13%). Of course this must be historical data, in this case over the last 50 years. Notwithstanding the standard

¹Of course I am not endorsing this outfit in particular; I just like their graphics.

Figure 2.1: Historical return vs volatility



“past performance does not guarantee future results” legal disclaimer, the intended implication is that it is reasonable to expect similar performance in future.

So how does this relate to any theory? Should you believe this implication if you invest your own money?

2.2 Some conceptual background

Games vs models. One of my 100 contexts where we perceive chance in the real world is rather fussily called “explicit games of chance based on artifacts with physical symmetry” exemplified by dice, roulette, lotteries, playing cards, etc. Textbooks for a first course in mathematical probability lean heavily on such artifacts for examples. Amongst the vast literature, both elementary and detailed game-specific, regarding probabilities in such games, let me mention only Ethier’s 2010 monograph *The Doctrine of Chances*, which gives an encyclopedic cross-section of less-elementary mathematics for games of chance. Of course the mathematical attraction of these exam-

ples is that we know underlying probabilities, and so for sufficiently simple games with these devices (craps or roulette or lotteries, but not games like poker involving complex strategy) we can calculate probabilities of winning purely mathematically without any input from empirical data. Our “list of contexts” spotlights how uncommon this is; I only know one other interesting case where such calculations are possible (see Lecture 4).

As much as possible I avoid repeating such standard “games of chance” material in these lectures. My point is that outside such settings, **probability calculations depend on models**. As Taleb wrote in [The Black Swan](#):

the sterilized randomness of games does not resemble randomness in real life;

another viewpoint is that

setting up a model presupposes you know the rules, but Life (unlike games) does not come with a rule book.

Professional statisticians like to quote [Box’s dictum](#)

all models are wrong, but some are useful.

Let us recall here, as the most familiar general-purpose model we see in introductory mathematical statistics, the *IID model*², in which observed past data, or unobserved future variables, are modeled as independent realizations from the same chance process, analogous to throws of a die. In certain textbook contexts (random sampling, randomized controlled trials) this is true by fiat: the statistician supplied the randomness. But a generation or two of 20th century statistics textbooks have left the impression (without explicitly saying so) that any set of data can be modeled as IID. So many people, faced with quantitative data, proceed to doing procedures (tests of significance, confidence intervals etc) which only make sense under the IID model (or some more complex explicit model) without ever asking whether the IID assumption is conceptually reasonable or supported by the data. In these lectures I try to be more careful.

Expectation and gambling. Recalling some basic mathematical setup, write $\mathbb{P}(\cdot)$ for probability and $\mathbb{E}[\cdot]$ for expectation. Regarding gambling, any bet has (to the gambler) some random profit X (a loss being a negative profit), and we say that an available bet is (to the gambler)

²Independent and identically distributed.

favorable if $\mathbb{E}[X] > 0$
unfavorable if $\mathbb{E}[X] < 0$
 and *fair* if $\mathbb{E}[X] = 0$.

Note the word *fair* here has a specific meaning. In everyday language, the rules of team sports are *fair* in the sense of being the same for both teams, so the better team is more likely to win. For 1 unit bet on team B, that is a bet where you gain some amount b units if B wins but lose the 1 unit if B loses,

$$\mathbb{E}[\text{profit}] = bp - (1 - p); \quad p = \mathbb{P}(\text{B wins})$$

and so to make the bet is fair we must have $b = (1 - p)/p$. (Confusingly, mathematicians sometimes say “fair game” to mean each player has chance $1/2$ to win, but this is sloppy language).

Several issues hidden beneath this terminology should be noted. Outside of games we usually don’t know probabilities, so we may not know whether a bet is favorable, aside from the common sense **Sky Masterson principle** that most bets offered to us will be unfavorable to us. The terminology comes from the law of large numbers fact that if one could repeat the same bet with the same stake independently, then in the long run one would make money on a favorable bet but lose money on an unfavorable bet. Such “long run” arguments ignore the issues of (rational or irrational) risk aversion and utility theory, a topic in Lecture 10.1 recalled briefly later in this lecture, section 2.8. In essence, we are imagining settings where your possible gains or losses are small, in your own perception.

Unfavorable bets. Roughly speaking, there are two contexts in which we often encounter unfavorable bets. One concerns most activities we call *gambling*, e.g. at a casino, and the other concerns insurance. Regarding the former, mathematicians often say ridiculous things such as

Gambling against the house at a casino is foolish, because the odds are against you and in the long run you will lose money.

What’s wrong is the *because*. Saying

Spending a day at Disneyland is foolish, because you will leave with less money than you started with

is ridiculous, because people go to Disneyland for entertainment, and know they have to pay for entertainment. And the first quote is equally ridiculous. Casino gamblers may have irrational ideas about chance and luck, but in the U.S. they typically regard it as entertainment with a chance of winning, not as a plan to make money. So it’s worth being more careful and saying

Gambling against the house at a casino *and expecting to make money* is foolish, because the odds are against you and in the long run you will lose money.

The second context is that buying insurance is mathematically similar to placing an unfavorable bet – your expected gain is negative, because the insurance company wants to cover its costs and make a profit. But the whole point of buying insurance is risk aversion, so this needs to be treated in the setting of utility theory and psychology of probability (Lecture 10.1).

So where can I find a favorable bet? The wisecrack answer “start your own casino or insurance company” is not so practical, but a variant of the latter is. For those who can, following the advice

increase your insurance deductibles to the maximum you can comfortably afford to lose

is a favorable bet, likely to save you money over a lifetime. In this lecture we consider investing in the stock market as mathematically similar to making a sequence of favorable bets (and letting your winnings ride). Exactly *why* one could consider this a favorable bet could be debated endlessly – standard economic theory asserts that investors need to be rewarded for taking risk rather than using alternative risk-free investments, while empiricists observe that, in countries without anti-capitalist revolutions, the historical performance of stock markets actually has been better than those alternatives.

2.3 Probability and financial investment

As the reader surely knows, the huge growth of financial markets since the 1970s has been driven in part by a huge increase in the use of sophisticated quantitative strategies for trading. There has been a parallel growth of the associated academic field of **mathematical finance**, much of which involves probability models. The most famous result in this field is the **Black-Scholes formula** for option pricing, which we mention later (section 2.9). But this result, and almost all of textbook mathematical finance, is simply irrelevant to the typical individual investor, more specifically to my students who are not considering a career in finance.

So what should such students know about financial investment? There is a simple first answer: they should read Malkiel’s classic book *A Random*

Walk Down Wall Street. But rather than paraphrase all the useful advice therein, this lecture will focus (after some more preliminaries) on a topic that I find both mathematically interesting and actually useful to know, the *Kelly criterion* for betting on a favorable game, as applied to financial investment. Between short-term speculation (at the modern extreme, **high-frequency trading**) and long-term investment lies a spectrum of intermediate activities with no clear dividing line, but the ends of the spectrum are very different. Novice investors are told to view the stock market as a place for *long-term* investment. This excellent advice is unfortunately rather neglected in most mathematically oriented discussions, but a great virtue of the Kelly approach is that it both emphasizes the long term while saying something explicit about the short term (section 2.7).

What is the long term? Because notions of *long term* versus *short term* play an important role in investment, let's start with a brief discussion. In everyday language, a job which will only last six months is a short term job; someone who has worked for a company for fifteen years is a long term employee. Joining a softball team for a summer is a short term commitment; raising children is a long term commitment. We judge these matters relative to human lifetime; *long term* means some noticeable fraction of a lifetime.

Table 2.1: Effect of 7% interest, compounded annually.

year	0	4	8	12	16	20
simple interest	1000	1,280	1,560	1,840	2,120	2,400
compound interest	1000	1,311	1,718	2,252	2,952	3,870

Turning to money matters, consider the difference between simple interest and compound interest. Table 2.1 compares the value, after increasing numbers of years, of an initial \$1,000 earning 7% interest. One of several possible notions of *long term* in financial matters is “the time span over which compounding has a noticeable effect”. Rather arbitrarily interpreting “noticeable effect” as “10% more” and taking the 7% interest rate, this suggests taking 8 years as the cut-off for *long term*. Being about 10% of a human lifetime, this fortuitously matches reasonably well the “noticeable fraction of a lifetime” criterion above. And indeed in matters pertaining to individuals, financial or otherwise, most writers use a cut-off between 5 and 10 years for “long term”.

Aside: the one fact from freshman calculus of most substantial relevance to

your personal life is the inequality

$$1 + \rho(e^{rt} - 1) > e^{\rho rt}.$$

This shows the value of unit investment, with interest rate r and tax rate $1 - \rho$, is greater when tax is deferred until the sale time t than if tax is paid as the interest is earned.

2.4 A simulation demo

The stock market is really “a market of stocks”, but for most of this lecture I use a conventional shorthand of representing the U.S. market by the S&P500 index (essentially an actual investment possibility, via a low-expense index fund).

In class I ask students

Suppose you invest \$1,000 today in the stock market, more precisely in an S&P500 index fund. What do you guess the investment will be worth in 10 years?

Converting their answers to annual percentage growth, the spread of their guesses is indicated in Table 2.2.

Table 2.2: Student guesses annual S&P500 growth next 10 years

date	25th percentile	median	75th percentile
9/17/2008	-0.5%	5.4%	8.4%
9/6/2011	-3.5%	4.1%	6.1%
9/3/2014	4.2%	7.8%	11.2%

This is consistent with the often-remarked observation that amateur investors tend to base their expectations on the previous several years performance.

I remind them that, on being asked what the stock market will do, **J. P. Morgan famously said**

It will fluctuate.

I emphasize this point in class as follows. I have a deck of cards on which are pasted the annual total returns of the S&P500 index over each of the 52 years 1956 through 2007. I say “let’s suppose the annual returns over

the next ten years are statistically like random years from the past; we can track our hypothetical investment value over the next ten years by shuffling and dealing ten cards”. Doing this once in the 2008 class, the hypothetical investment grew from \$1,000 to \$1,839, while fluctuating noticeably from year to year. I then say “guess what will happen if I repeat this simulation” and most guesses are within 20% of the first outcome. In fact different realizations vary much more widely than students guess, so (at some risk to the credibility of theory!) I do repeat, hoping that in fact the outcome is indeed substantially different, as it likely will be.

We now start the mathematics, but let me summarize with some bullet points.

- The theme of this lecture is the nature of compounding when gains and losses are unpredictable.
- The relevant arithmetic is **multiplication** not addition: a 20% gain followed by a 20% loss combine to a 4% loss, because $1.2 \times 0.8 = 0.96$.
- Models assume the future will be statistically similar to the past.

2.5 The IID model and the Kelly criterion

Turning to mathematics, let us make explicit the type of model used implicitly above. A “return” $x = 0.2$ or $x = -0.2$ in a year means a 20% gain or a 20% loss.

The IID model. Write X_i for the return in year i . Suppose the (X_i) are IID random variables. Then the value Y_n of your investment at the end of year n is

$$Y_n = Y_0 \prod_{i=1}^n (1 + X_i) \quad (2.1)$$

where Y_0 is your initial investment.

To analyze this model we take logs and divide by n :

$$n^{-1} \log Y_n = n^{-1} \log Y_0 + n^{-1} \sum_{i=1}^n \log(1 + X_i)$$

and the law of large numbers says that as $n \rightarrow \infty$ the right side converges to $\mathbb{E}[\log(1+X)]$. We want to compare this to an investment with a non-random return of r . For such an investment (interest rate r , compounded annually)

we would have $Y_n = Y_0(1+r)^n$ and therefore $n^{-1} \log Y_n \rightarrow \log(1+r)$. Matching the two cases gives the conclusion

In the IID model, the long term growth rate is

$$\exp(\mathbb{E}[\log(1+X)]) - 1. \quad (2.2)$$

The formula looks strange, because to compare with the IID annual model we are working with the equivalent “compounded annually” interest rate. It is mathematically nicer to use instead the “compounded instantaneously” interest rate, which becomes just $\mathbb{E}[\log(1+X)]$.

The main impact of this result is that what matters about the random return X is *not* precisely the mean $\mathbb{E}[X]$, but rather its “multiplicative” analog $\mathbb{E}[\log(1+X)]$. Let us note, but set aside for a while, the points

- Is the model realistic for stock market investing?
- The phrase *long term* here refers to the applicability of the law of averages as an approximation to finite time behavior – this is a third meaning of the phrase, logically quite distinct from the two previous meanings in section 2.3.

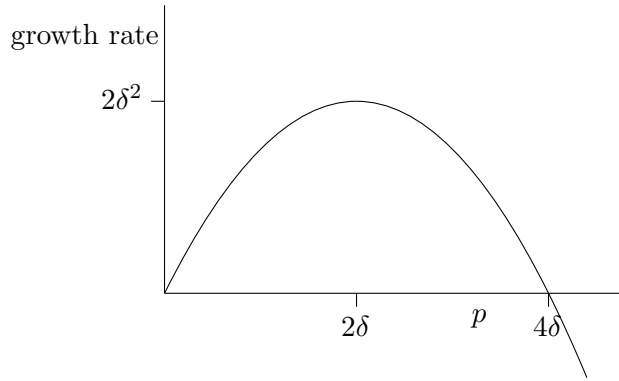
We can jump to the main mathematical point

The Kelly Criterion. *Given a range of possible investment portfolios, that is a range of ways to allocate money to different risky or safe assets, where way α will produce return X_α , choose the way α that maximizes the long term growth rate (2.2), that is the way that maximizes $\mathbb{E}[\log(1+X_\alpha)]$.*

Let us show some of the mathematics that can be done using the criterion.

2.6 Mathematics of the Kelly criterion: one risky and one safe asset

Formula (2.2) applies if we invest all our money in “a stock” (meaning an index fund, and assuming the multiplicative model for stock market returns). But suppose there’s a risk-free alternative investment (a “bond”, in the usual terminology) that pays a fixed interest rate r . Suppose we choose some number $0 \leq p \leq 1$ and at the start of each year we invest a proportion p

Figure 2.2: Schematic for the growth rate $G(p)$ at (2.4).

of our total “investment portfolio” in the stock market, and the remaining proportion $1 - p$ in the bond. In this case formula (2.1) becomes

$$Y_n = Y_0 \prod_{i=1}^n (1 + X_i^*)$$

where $X_i^* = pX_i + (1 - p)r$. The long term growth rate is now a function of p :

$$\text{growth}(p) = \exp(\mathbb{E}[\log(1 + pX + (1 - p)r)]) - 1. \quad (2.3)$$

The Kelly criterion says: choose p to maximize $\text{growth}(p)$. Let’s see two examples. In the first X is large, and we end up with p small; in the second X will be small, and we end up with large p . In these two examples we take the time unit to be 1 day instead of 1 year (which doesn’t affect math formulas).

Example: pure gambling. Imagine a hypothetical bet which is slightly favorable. Suppose each day we can place a bet of any size s ; we will either gain s (with probability $0.5 + \delta$) or lose s (with probability $0.5 - \delta$), independently for different days (here δ is assumed small). Take $r = 0$ for the moment. What proportion p of our portfolio do we want to bet each day?

Here, for small δ ,

$$\mathbb{E}[\log(1 + pX)] = \left(\frac{1}{2} + \delta\right) \log(1 + p) + \left(\frac{1}{2} - \delta\right) \log(1 - p)$$

$$\begin{aligned} &\approx \left(\frac{1}{2} + \delta\right)(p - p^2/2) + \left(\frac{1}{2} - \delta\right)(-p - p^2/2) \\ &= 2\delta p - p^2/2. \end{aligned}$$

Thus the asymptotic growth rate is approximately the quadratic function of p

$$G(p) = 2\delta p - p^2/2 \tag{2.4}$$

shown in Figure 2.2. The Kelly criterion says to choose $p \approx 2\delta$ and then your long term growth rate will be $\approx 2\delta^2$.

Now recall that we simplified by taking $r = 0$; when $r > 0$, the fact that a proportion $1 - p \approx 1$ of the portfolio not put at risk each day can earn interest, brings up the optimal growth rate to $r + 2\delta^2$; the quantity $2\delta^2$ represents the *extra* growth one can get by exploiting the favorable gambling opportunity.

To give a more concrete mental picture, suppose $\delta = 1\%$. The model matches either of the two following hypothetical scenarios.

(a) To attract customers, a casino offers (once a day) an opportunity to make a roulette-type bet with a 51% chance of winning.

(b) You have done a statistical analysis of day-to-day correlations in some corner of the stock market and have convinced yourself that a certain strategy (buying a portfolio at the start of a day, and selling it at the end) replicates the kind of favorable bet in (a).

In either scenario, the quantity $2\delta^2 = 2/10,000$ is the “extra” long term growth rate available by taking advantage of the risky opportunity. Note this growth rate is much smaller than 2% “expected gain” on one bet. On the other hand we are working “per day”, and in the stock market case there are about 250 days in a year, so the growth rate becomes about 5% per year; recalling this is “5% above the risk-free interest rate”, it seems a rewarding outcome. But if δ were instead 0.5% then the extra growth rate becomes $1\frac{1}{4}\%$, and (taking into account transaction costs and our work) the strategy hardly seems worth the effort.

Implicit in Figure 2.2 is a fact that at first strikes everyone as counter-intuitive. The curve goes negative when p increases above approximately 4δ . So even though it is a favorable game, if you are too greedy then you will lose in the long run!

Example: small X I first give a nicer algebraic way of dealing with the interest rate r . Set

$$X = r + (1 + r)X^*$$

and interpret $X^* = (X - r)/(1 + r)$ as “return relative to interest rate”. Then a couple of lines of algebra let us rewrite (2.3) as

$$\text{growth}(p) = (1 + r) \exp(\mathbb{E}[\log(1 + pX^*)]) - 1 \quad (2.5)$$

and the optimization problem now doesn’t involve any r . If we imagine the stock market on a daily time-scale and suppose changes X^* are small, with mean μ and variance σ^2 , then we can use the series approximation

$$\log(1 + pX^*) \approx pX^* - \frac{1}{2}(pX^*)^2$$

to calculate

$$\mathbb{E}[\log(1 + pX^*)] \approx p\mu - \frac{1}{2}p^2(\mu^2 + \sigma^2) \approx p\mu - \frac{1}{2}p^2\sigma^2$$

(the latter because μ and σ^2 are in practice of the same order, so μ^2 is of smaller order than σ^2). So the Kelly criterion says: choose p to maximize $p\mu - \frac{1}{2}p^2\sigma^2$, that is choose

$$p = \mu/\sigma^2. \quad (2.6)$$

This is another remarkable formula, and let us discuss some of its mathematical implications.

1. The formula is (as it should be) *time-scale free*. That is, writing $\mu_{day}, \mu_{year}, \sigma_{day}^2, \sigma_{year}^2$ for the means and variances over a day and a N -day year, then (because compounding has negligible effect over a year) $\mu_{year} \approx N\mu_{day}$ and $\sigma_{year}^2 \approx N\sigma_{day}^2$, so we get the same value for μ/σ^2 whether we work in days or years.

2. Even though we introduced the setup by stating that $0 \leq p \leq 1$, the model and its analysis make sense outside that range. Economic theory and experience both say that the case $\mu < 0$ doesn’t happen (investors are risk averse and so would buy no stock; this would cause the current price of stock to drop), but if it did then the formula $p = \mu/\sigma^2 < 0$ says that not only should we invest 100% of our wealth in the bond, but also we should “sell short” (i.e. borrow) stock and invest the proceeds in the bond.

3. More interesting is the case $p > 1$. Typical values given for the S&P500 index (as noted later, stating meaningful historical values is much harder than one might think) are $\mu = 5.6\%$ and $\sigma = 20\%$, in which case the Kelly criterion says to invest a proportion $p = 140\%$ of your wealth in the stock market, i.e. to borrow money (at fixed interest rate r) and invest your own and the borrowed money in the stock market.

2.7 What about the not-so-long term?

We started with the multiplicative model, which assumes that returns in different time periods are IID. This is not too realistic, but the general idea behind the Kelly criterion works without any such assumption, as we now explain.

Going back to basics, the idea

to invest successfully in the stock market, you need to know whether the market is going to go up or go down

is just wrong. Theory says you just need to know the *probability distribution* of a future return. So suppose (a very big **SUPPOSE**, in practice!) at the beginning of each year you could correctly assess the probability distribution of the stock return over the coming year, then you can use the Kelly criterion (2.3) to make your asset allocation. The fact that the distribution, and hence your asset allocation, would be different in different years doesn't make any difference – this strategy is still optimal for long-term growth.

The numbers for growth rates that come out of the formula of course depend on the distributions of each next year's returns, but there's one aspect which is "universal". In any situation where there are sensible risky investments, following the Kelly strategy means that you accept a short-term risk which is always of the same format:

40% chance that at some time your wealth will drop to only 40% of what it you started with.

The magical feature of this formula is that the percents always match: so there is a 10% chance that at some time your wealth will drop to only 10% of what you started with.

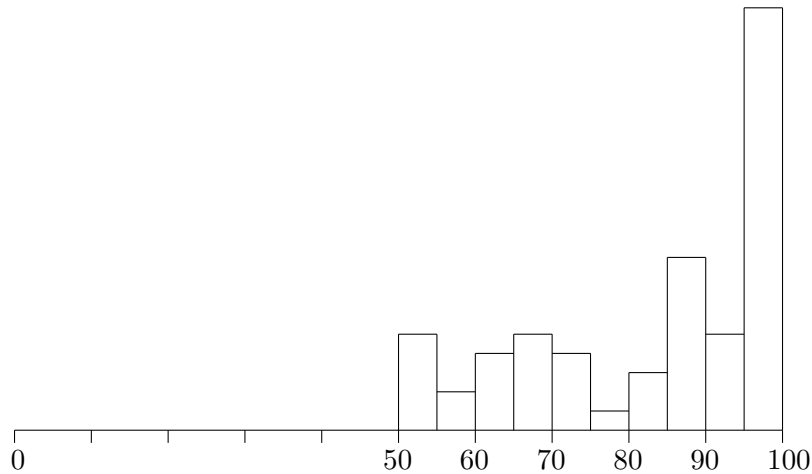
For an individual investor, it is perfectly OK to be uncomfortable with this level of medium-term risk and to be less *aggressive* (in investment jargon) by using a *partial Kelly* strategy, that is using some smaller value of

p = proportion of your assets invested in stocks

than given by the Kelly criterion. Theory predicts you will thereby get slower long-term growth but with less short-term volatility.

How one might expect this theory (based on assuming known true probability distributions for the future, and on seeking to optimize long-term growth rate) to relate to the actual stock market is not obvious, but one can certainly look at what the actual percentages have been – Figure 2.3.

Figure 2.3: Historical distribution (based on hypothetical purchase of S&P500 index on first day of each year 1950-2009 and on subsequent monthly closing data) of the minimum future value of a 100 investment.



This is obviously very different from the “Kelly” prediction of a flat histogram over $[0, 100]$. This data is not adjusted for inflation or for comparison with a risk-free investment, and such adjustments (a possible small project) would make the histogram flatter, but still not close to the Kelly prediction. We mentioned before that over the historical long term it has been more profitable to borrow to invest more than 100% of your assets in the market. Both observations reflect the fact that the stock market fluctuates *less* than would the fortune of a Kelly-optimizing speculator.

2.8 Can one execute this theory?

Relating the mathematical theory to actual stock market investing raises many issues.

Understanding the past. The first is an issue you might not have expected:

it is very hard to pin down a credible and useful number for the

historical long-term average growth rate of stock market investments.

Over the 60 years 1950-2009 the S&P500 index rose at (geometric) average rate³ 7.2%. Aside from the (rather minor) point that we are using a particular index to represent the market what could possibly be wrong with using this figure? Well,

- it ignores dividends
- it ignores expenses
- it is sensitive to choice of start and end dates; starting in 1960 would make the figure noticeably lower, whereas ending in 1999 would make it noticeably higher.
- to interpret the figure we need to compare it to some alternative investment, by convention some “risk-free” investment.
- it ignores inflation
- it ignores taxes.

So one can get very different numbers, depending on which of these factors is taken into account. For instance, taking two of these factors into account, the same site gives, for the same period 1950-2009, annual averages

$$\begin{aligned} & \text{inflation-adjusted total return} \\ & = \text{price change (7.2\%)} + \text{dividends (3.6\%)} - \text{inflation (3.8\%)} \\ & = 7.0\% . \end{aligned}$$

The past as a guide to the future. Forests have been felled for paper for philosophers to discuss the **problem of induction**. In the context of investment, my take is

As a default, assume the future will be statistically similar to the past. Not because this is true in any Platonic sense, but because anyone who says different is trying to sell you something.

As often remarked, the four most dangerous words in finance are *this time it's different*.

Asking how long of a past time period to use *for statistical purposes* as a guide for the future is a question with no right answer. Asking how far into the future one should care about does have an answer – until you're age 80 or so (that is, 60 years ahead for my students).

³Data here and below from [this site](#)

Psychology in executing Kelly. So if you do set up a historically-based model for the stock market (and at least one alternative investment) and assume the future will be statistically similar to the past, then the Kelly criterion tells you how to divide your money between these investments for maximum long term growth, assuming the model were correct. But some practical issues still remain. What is your personal trade-off between long term reward and short-term risk? How should this change with age – typically one is advised to become more risk-averse as one ages. (In principle one could introspect your utility function and then calculate your optimal trade-off, but I suspect few people have ever done so.) In a different vein, long term strategies can only work if you avoid changing your mind partway through, so how does one plan to avoid changing one’s mind later?

Bottom line. Going through the procedures above by oneself will have little appeal to a typical individual, so what is the closest practical option? There are many internet sites that provide combinations of historical data, hypothetical data and (not so mathematical) theory relevant to this topic, though usually presented in the more conventional mean-variance format of section 2.10. As mentioned in the opening section. IFA and other sites start out by trying to assess the individual’s subjective risk tolerance using a questionnaire. The site then suggests one of a range of 21 portfolios, roughly 0% to 100% Kelly in 5% increments, based on historic data.

2.9 Geometric Brownian motion and the Black-Scholes formula

Most introductory accounts of mathematics related to the stock market do not focus on the Kelly criterion, but instead on the (related) topics of this and the next section. The IID model involved choosing a time unit – we chose one year – and the resulting strategy involves rebalancing your portfolio at the end of each year but not more frequently. But the choice of one year is arbitrary – why not a month or a week or a day or an hour? It is perhaps more natural to seek to model as a random process what you see in a stock chart, the fluctuation of prices with time, regarding time as a continuous variable. It turns out, as a remarkable mathematical fact, that the only possible model one can use as the continuous analog of the IID model (and have prices move continuously, without substantial instantaneous jumps) is **geometric Brownian motion**. Instead of using data from the past directly (as in my “deck of cards” class demonstration) this theoretical setup allows

and encourages you to assume that the probability distribution of a 1-year return has the **log-Normal distribution**, then to estimate the two parameters of that distribution, and then to use the associated geometric Brownian motion model for making predictions about the future.

This geometric Brownian motion model captured the attention of mathematically inclined researchers in the early 1970s as a setting where it is possible to write down formulas for aspects of the future price fluctuations (assuming the model is correct). The prototype is the following **Black-Scholes formula**. If a stock price is currently S_0 and its future prices S_t are assumed to follow geometric Brownian motion

$$dS_t = \mu S_t dt + \sigma S_t dW_t \quad (2.7)$$

then the fair price of an option to buy at time t in the future at price K is

$$S_0 \Phi(d_1) - Ke^{-rt} \Phi(d_2) \quad (2.8)$$

where r is the risk-free interest rate, $\Phi(\cdot)$ is the standard Normal distribution function,

$$d_1 = \frac{\log(S_0/K) + (r + \sigma^2/2)t}{\sigma\sqrt{t}}, \quad d_2 = \frac{\log(S_0/K) + (r - \sigma^2/2)t}{\sigma\sqrt{t}}.$$

This certainly belongs on our list of the **top ten formulas in applied probability**, but its interpretation is more subtle than the others. The fair price in question is by definition

$$e^{-rt} \mathbb{E} \max(S_t - K, 0).$$

The formula for this quantity, applied to geometric Brownian motion (2.7), depends on the drift μ , but the Black-Scholes formula (2.8) does not. The key issue is that an extra “no arbitrage” assumption, loosely analogous to the “martingale” assumption for prediction market prices in Lecture 4, is used to establish (2.8), by imagining a portfolio consisting of time-varying amounts of the stock and the option but with a total value that does not vary with the fluctuations of the stock price.

2.10 Mean-variance analysis

We earlier discussed diversification in the very simple context of one safe and one risky asset. The general case of diversifying over many assets is usually presented as *mean-variance analysis*, or (more pretentiously) as **modern portfolio theory**. This should be viewed as a medium-term theory – it

doesn't explicitly involve compounding – a viewpoint perhaps influenced by traditional economics focus on medium term issues like the business cycle.

Here's the conceptual setup. A portfolio has a *reward*, which is mean annual return, and a *risk*, which is the s.d. of annual return. Different investors have different *risk tolerances*; the goal of theory is to produce a spectrum of portfolios which provide maximum reward for each given level of risk.

The mathematics of mean-variance analysis involves the kind of matrix algebra that one learns as an undergraduate Statistics or Economics major. Rather than give the algebra I will illustrate with simple hypothetical numbers and then jump to the bottom line.

Suppose there were only two stocks, and historical data for annual returns showed

stock	mean	s.d.
A	8%	20%
B	8%	20%

One's first thought is that there's no difference between investing in A or in B. But the point is that one can invest 50% of one's portfolio in each; this preserves the mean of 8% but reduces the s.d., that is reduces the risk. If the stock price fluctuations were independent then the s.d. of the portfolio would be reduced to around 14%. In practice stock returns are typically positively correlated, so the reduction in s.d. is smaller, but still desirable.

For the next simplest hypothetical example, some the s.d.s are unequal

stock	mean	s.d.
A	8%	15% = σ_A
B	8%	20% = σ_B

Again one might first think that one should invest entirely in A. But a mixture – proportion p in A and proportion $1 - p$ in B – has s.d. σ given (in the independent case) by

$$\sigma^2 = p^2\sigma_A^2 + (1 - p)^2\sigma_B^2$$

and this is minimized by taking $p = \sigma_B^2 / (\sigma_A^2 + \sigma_B^2)$. With the numbers shown, take $p = 0.64$ to get $\sigma = 12\%$.

As a third hypothetical example consider

stock	mean	s.d.
A	8%	15% = σ_A
B	6%	20% = σ_B

Here a 50-50 mixture has $\mu = 7\%$ and $\sigma = 12.5\%$, and one might prefer that trade-off to the stock A parameters.

(In class I continue, discussing the *efficient frontier*, following Wikipedia’s [Capital asset pricing model](#)).

Analogous to the Kelly criterion one can identify one of these as the *mean-variance optimal portfolio*. By typing several stocks, e.g. `apple` `exxon` `coca-cola`, into [WolframAlpha](#) one can see this portfolio on these stocks plus S&P500, bonds and T-bills. (xxx not working 6/14 - 2/15).

Linking mean-variance analysis and the Kelly criterion Writing X as before for return in one year, and writing $\log(1 + X) \approx X - X^2/2$ so that $\mathbb{E} \log(1 + X) \approx \mu - (\mu^2 + \sigma^2)/2$, we see that the Kelly criterion corresponds (approximately) to choosing, over possible portfolio combinations, the one whose (μ, σ) value maximizes $\mu - (\mu^2 + \sigma^2)/2$. The approximation here is that we are ignoring the possibility of unusually large changes.

2.11 On investment advice

A search on “diet” in amazon.com books produces⁴ a claim of 63,932 results, though the listings actually stop at number 1201 (*The Raw Secrets: The Raw Vegan Diet in the Real World* by Patenaude). Similarly, a search on “investment” produces a claim of 78,686 results, while these listings also stop at number 1201 (evidently an Amazon policy). A naive visiting Martian might think the former reflected a vast diversity of human dietary requirements for some genetic, environmental or occupational reasons, and wonder what analogous factors might require such a wide variety of investment strategies. The reader, as a skeptical human, may suspect such books exist merely because enough people are willing to buy them. Here are two of the many strategies.

Fundamental analysis of an individual stock seeks to assess the “intrinsic value” of the business by analyzing its likely future profits, and thereby find stocks which the current market price overvalues or undervalues.

Market timing seeks to analyze aggregate economic data (business cycle, inflation, unemployment, corporate profits) as well as *sentiment* (opinions about the near future) to decide when to switch between stocks/bonds/cash or between subsectors of the markets.

Almost all investment passes through some kind of “professional advisor” (including mutual fund managers, analysts etc). So as a simple matter of arithmetic, the average return to investors must equal the average return of the markets, minus expenses and fees paid to the advisors. I shall not

⁴May 2011

go into details about the **efficient-market hypothesis** (EMH). Treating EMH as an ideology or a law of physics strikes me as silly, but simply asserting that it is at least *very difficult* to consistently beat the market – just as it is very difficult to be one of the best people in the world at anything substantial – is much more plausible. Concretely, what the EMH predicts is that if you take any well-defined strategy that has been used by a group of advisors, then over the long term, the average return to investors must again equal the average return of the markets, minus expenses and fees. Numerous academic studies of this prediction have been done, and generally confirm the prediction (reviewing such studies is a natural student project).

Even people who accept the simple logic and experience that most professional advisors can't beat the market are frequently seduced by the notion that a few can. Here's my take on this. Suppose someone comes up to me, takes a coin out of their pocket, says "I'm going to toss the coin 10 times and make it land heads every time", and then does so. What's my reaction? Well, there are three possibilities. They might just have been very lucky. They might be cheating (a two-headed coin, or a second concealed coin). Or they might be exhibiting a rare and difficult skill, the ability to toss a coin which in fact only rotates a specific few times in the air and lands predictably. Analogously, if you look at the 5 best-performing advisors over the last 5 or 10 or 15 years, then they might just have been lucky (*some* 5 people must be best), they might be exhibiting a rare and difficult skill of actually being able to consistently beat the market (as does Warren Buffett, in the opinion of many people) or they might be another Bernie Madoff. Whether or not a few advisors will in the future be able to consistently beat the market is an interesting intellectual question *but it doesn't matter to you* – you can't reliably pick one of these few out of the pack, any better than you can pick one of the few future-best-performing stocks out of the pack.

2.12 Wrap-up

I have only touched upon a corner of a large topic, but within this corner let me reiterate some key points.

1. Common sense says objects can be stationary or move slowly or move fast or move very fast, and that there should be no theoretical limit to speed – but physics says in fact you can't go faster than the speed of light. And that's a very non-obvious fact. Similarly, we know there are risk-free investments with low return; by taking a little risk (risk here equals short-term fluctuations) we can get higher long-term reward. Common sense

says this risk-reward trade-off spectrum continues forever. But it doesn't. As a math fact, you can't get a higher long-term growth rate than you get from the "100% Kelly strategy". You're free to take more risk if you like excitement but you don't benefit from it.

2. Mathematics (section 2.10) not only confirms that diversification is good but also shows it is even better than you might intuitively expect.

3. Any mathematics one can do, involving the stock market or wider aspects of finance and risk, either assumes (as we have) that the future will be statistically like the past, or makes an explicit assumption of some ways in which it will be different. Now of course the rules of the game do change with time, but the consensus view of such trends is already reflected in current prices. To profit one would need to adopt some minority view of the future, and be correct. Surely the majority of people who try to do so get it wrong.

4. Popular opinion often says that stock market fluctuations are larger than they "should be", whereas mathematics says that the stock market has historically fluctuated *less* than would the fortune of a Kelly-optimizing speculator. In fact no-one knows how large the short-term fluctuations "should be", under either a "rational" or a "psychological" theory of the market, and a testable theoretical prediction relating market fluctuations to measurable aspects of the real economy would surely win you a Nobel prize or enable you to make a fortune (by speculating on volatility).

5. Geometric Brownian motion is a mathematically natural, and reasonably accurate, model for the short term fluctuations of stocks. After Bachelier pointed this out in 1900, the model was mostly underused until the 1970s, but subsequently overused and treated as more accurate than it really is.

6. Unlikely things are going to happen over your 50-year investment lifetime, and thinking in terms of Kelly rather than mean-variance encourages you to pay attention to small chances of dramatic losses.

7. To adapt Churchill's comment on democracy,

The EMH is the worst way of thinking about the stock market, except for all those other ways that have been tried from time to time.

Or as [John Bogle](#) put it⁵

Sometimes markets are efficient, sometimes they are not, and it is not possible to know which is which.

⁵letter to the Economist, November 2013.

2.13 Further reading

Poundstone's *Fortune's Formula* is a non-technical book on the Kelly criterion. Much of the book is entertaining episodic anecdotal history of characters like Shannon, Kelly, Thorp, Milken, Boesky and Long Term Capital Management. It has an interesting account of the dispute between the proponents of Kelly (math types) and economists led by Samuelson who viewed it as too risky even in the long run. His memorable slogan, in place of my "speed of light" analogy, is

100% Kelly strategy marks the boundary between aggressive and insane investing.

Aaron Brown's 2011 *Red-Blooded Risk* discusses financial trading and risk management as actually practiced over the last 35 years, in relation to underlying ideas from mathematical probability. In particular his Chapter 5, which everyone interested in the basic mathematics of finance should read, compares and contrasts modern portfolio theory with Kelly. His memorable quote is that using Kelly enables you to "get rich exponentially slowly".

Malkiel's classic *A Random Walk Down Wall Street* sets out in plain language the academic view that seeking to beat the market via fundamental analysis is a mug's game. Taleb's recent best-seller *The Black Swan* is hard to describe in a few words (here is [my lengthy review](#)) but one of its main points is that the geometric Brownian motion model, and formulas such as Black-Scholes based on it, ignore small chances of unforeseen events that might have substantial effect. Reinert and Rogoff's *This Time Is Different: Eight Centuries of Financial Folly* is a quantitative academic study, pointing out (to quote one reviewer) "the highly repetitive nature of financial crises resulted from a dangerous mix of hubris, euphoria and amnesia".

Chapter 3

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

These topics are staples of popular science style books on probability; in this lecture we dig a little below the surface.

3.1 The birthday problem and its relatives

The **birthday problem**— 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 3.2 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 team¹ 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

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

readily check this prediction in class in a minute or so (print out the 30 pages and distribute among students).

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 might 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 players²). 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 repeat this critique more forcefully in section 3.9. My own (unsuccessful) attempt to do better is recounted in section 3.5.

One can invent and solve a huge number of analogous math probability problems and I show a glimpse of such problems in section 3.2. 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

²who 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.

we see streaks longer than predicted in an oversimplified chance model is not telling us anything concrete about the world of sports.

3.2 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 approximation³.

If events A_1, A_2, \dots 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) $\text{Poisson}(\mu)$, so

$$\mathbb{P}(\text{none of the events occur}) \approx \exp\left(-\sum_i \mathbb{P}(A_i)\right). \quad (3.1)$$

Consider the birthday problem with 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 (3.1)

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

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

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

³My 1989 book *Probability Approximations via the Poisson Clumping Heuristic* consists of 100 examples of such calculations, within somewhat more complicated models.

For the uniform distribution over N categories this becomes

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

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

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

and then in the uniform case,

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

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

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

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.

3.3 Coincidences in everyday life

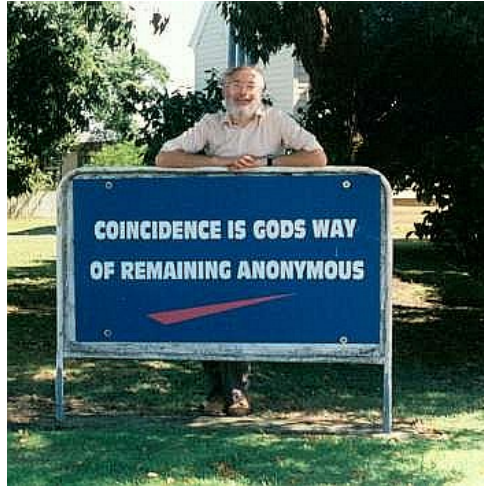
This lengthy discussion is mostly omitted in the lecture.

As Figure 3.1⁴ suggests, a long and continuing tradition outside mainstream science⁵ assigns spiritual or paranormal significance to coincidences,

⁴Photo found online; the gentleman is not me.

⁵e.g. Arthur Koesler *The Roots of Coincidence*, 1972.

Figure 3.1: Noticeboard outside a church.



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: here is [an example by G.J. Kirby](#) concerning the probability of meeting someone you know on a trip away from your home district, and not somewhere where either of you would usually be found. But such exercises 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.

Can we 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 a 1989 paper by Persi Diaconis and Fred Mosteller. Our “gazillions” explanation⁶, which they call the *law of truly large numbers* and which is also called *Littlewood's law*, 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

⁶Further comments will be given in section 3.8.

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 simplest 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 [Lincoln-Kennedy coincidences urban legend](#)); 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?

3.4 Coincidences in the news

Every time I teach the course I see relevant examples in current or recent news or in my email inbox that I can use. Here are two examples from the 2014 course.

Plane crash cluster. 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?

The relevant data is that over the last 20 years, crashes with substantial fatalities have occurred at rate 1/40 per day, so under the natural IID probability model the number N of crashes in a given 8 days has approximately Poisson(0.2) distribution, for which

$$\mathbb{P}(N = 3) \approx 0.2^3/6 \approx 1.33 \times 10^{-3}.$$

A calculation outlined here, which accounts for overlaps of 8-day intervals, shows that in the model such a cluster will occur “by chance” about once every 10 years. So this coincidence is not terribly unlikely.

This setting provides a concrete context for the section 3.3 general discussion. 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 could consider coincidences involving the same airline or the same region of the world or the same airplane model. Even if a coincidence within any one “specific type” were 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” typically rely on an analysis of the specifics of what did happen, but a meaningful analysis needs also to consider other types of coincidences that didn’t happen.

Assignment of court cases. 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.

A journalist emailed me the question:

What are the chances that one judge would pull these major cases when cases are supposedly assigned randomly?

In other words, is this just coincidence, or does it suggest maybe these cases were not assigned randomly? Note that we are not discussing the *merits of the judgments* – it would be nonsensical to model the judgements as random.

It turns out there are effectively⁷ 17.5 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. Guessing wildly, suppose there are 100 such groups-of-3. Then because, for each such group, there is the same 1/300 chance of all going to the same judge, then the chance that this happens for **some** group amongst the 100 groups is a little smaller than $100/300 = 1/3$, so would not be surprising.

Now of course the FDA-tobacco issue is unusually interesting. A more precise analysis would 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 (presumably not very large) number n , and the chance that some group "of interest to a journalist" were all assigned to the same judge (by pure chance) is around $n/300$. 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

and the next section provides a further illustration of this point.

3.5 Coincidences in Wikipedia

This section describes a project that we were unable to complete, but which remains interesting to me. The project was to examine 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) mentioned in section 3.3, while also allowing data to be gathered quickly and allowing independent replication by other people.

⁷some are part time.

Table 3.1: 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.

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

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 noticeable 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 material, except to find the name of the earlier coincident article.

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.

3.6 Classifying coincidences in everyday life

This is only briefly mentioned in lecture as a potential student project.

David Hand’s 2014 book *The Improbability Principle: Why Coincidences, Miracles, and Rare Events Happen Every Day* gives the standard rationalist explanation of coincidence: there are a vast number of possible events, so even if individual events are vastly unlikely, some such events will occur. Like other discussions of coincidences, it describes real-world events and anecdotes selected by the author in some unspecified way. In the spirit of our Lecture 1 examples, I would prefer to study examples obtained in some less subjective way. Fortunately there is a source of such examples. The [Cambridge Coincidences Collection](#) page invites readers to post their own coincidence stories. That site suggests typical types of coincidence, as follows.

- **Surprising repetitions:** for instance when you've had no 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?

From a somewhat different perspective Hand's book concludes with the invention of some "laws":

- The law of inevitability (the lottery case),
- the law of truly large numbers (vast number of possible coincidences),
- the law of selection ("surgeons bury their failures"),
- the law of the probability lever (inaccurate modeling of probabilities, as in the [Sally Clark children's SIDS case](#)),
- the law of near enough (we count near misses as hits).

A student project is to study real-world examples and devise a more systematic classification of types of everyday coincidence.

3.7 Near misses

Closely related to coincidences are a range of events that one might view as *near-misses*. That phrase originated in the setting of a physically 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 (xxx) section 8.7, 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 playing⁸ 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 typical student project is to 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 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.

3.8 What really has a 1 in a million chance?

This is fun to do in class, or in a Statistics Dept’s open day for the public, 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

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

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 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 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.

3.9 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.

3.10 Hot hands

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](#) discusses ongoing streaks, and it’s a good topic for student projects. The overwhelming conclusion of many statistical analyses is that in almost all sports the hot hand phenomenon is nonexistent or of negligible size. But as Amos Tversky once said

I’ve been in a thousand arguments over this topic. I’ve won them all, and I’ve convinced no one.

One analogous setting concerns cricket, in which there is a concept “getting your eye in” meaning that a batsman is more likely to be dismissed early in his innings, and this “cold hand” analog [does stand up to statistical analysis](#).

Chapter 4

Prediction markets, fair games and martingales

Prediction markets are in some ways analogous to stock markets but lead to simpler mathematics. This chapter overlaps substantially with a published write-up of the 2011 lecture¹.

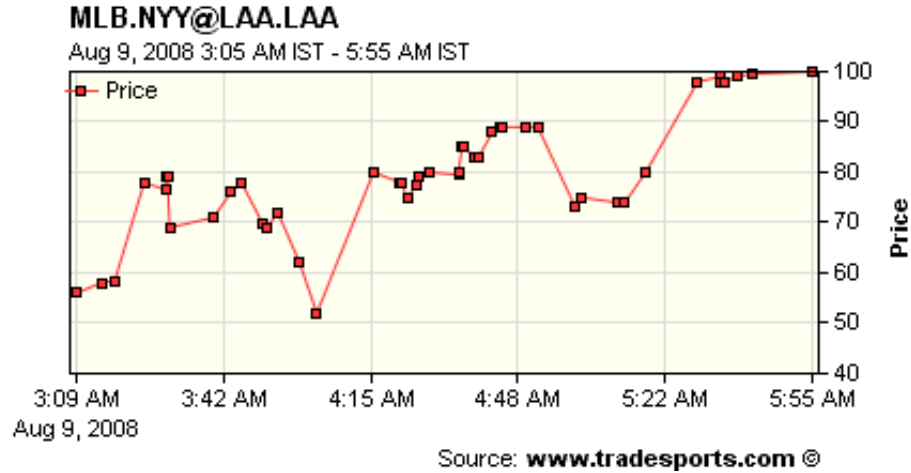
4.1 What is a prediction market?

As usual it is hard to better the following [Wikipedia description](#).

Prediction markets are speculative markets created for the purpose of making predictions. The current market prices can then be interpreted as predictions of the probability of the event People who buy low and sell high are rewarded for improving the market prediction, while those who buy high and sell low are punished for degrading the market prediction. Evidence so far suggests that prediction markets are at least as accurate as other institutions predicting the same events with a similar pool of participants.

¹ *Using prediction market data to illustrate undergraduate probability.* American Math Monthly 120 (2013) 583–393.

Figure 4.1: Times and prices of trades on a particular baseball match.



4.2 Some data: baseball and elections

First I describe some data collected from Tradesports² in 2008. Figure 4.1 refers to a baseball match between the New York Yankees and the Los Angeles Angels. There is a contract, which expires (at match end) at 100 if a specified team (in this case, the Angels) win, and expires at 0 if this team loses. The units are arbitrary; the non-arbitrary aspect is that one contract expiring at 100 is worth \$10. So to buy one contract at the opening “offer” price of 57 would cost \$5.70. You can bet on the other team by selling the contract. So you could sell one contract at the opening “bid” price of 56.

This setting differs from traditional gambling in that there isn’t a book-maker; you are trading with other participants. What you see on the web site is the bid and asked prices and quantities; you can either accept someone’s posted offer, or post your own offer. Tradesports made its money from a 4% fee on net winnings. So in the two cases above (buy or sell one contract at opening bid/offer prices), when the Angels won the match, your profit or loss would have been

$$[\text{buy}]: \text{profit of } (100 - 57)\% \times \$10 \times 96\% = \$4.13$$

$$[\text{sell}] \text{ loss of } (100 - 56)\% \times \$10 = \$4.40.$$

²Tradesports ceased business in October 2008. It was based in Ireland and showed IST time; the match time was really 7 - 10 pm in Los Angeles.

Figure 4.2: A long-running prediction market contract



Of course most trades involve a larger number of contracts, just as most stock market trades involve more than one share of the stock. During summer 2008, Tradesports emphasized three baseball games each day, and typically around 4,000 contracts were traded on each game.

Next is data from [Intrade](#), which was the largest political/economic prediction market in the world until being closed under U.S. regulatory pressure around the end of 2012. Figure 4.2 shows the prediction market price, over December 2010 - November 2012, for Obama to win the U.S. Presidential Election.

Prediction markets constitute one of the very few cases, outside of “games of chance based on artifacts with physical symmetry”, where one can make quantitative predictions without needing much input data from the real world. More precisely, even though there is usually no relevant mathematical theory to tell you what the current price should be, given the current price there *is* theory concerning future fluctuations. We will develop this theory and then return to data.

4.3 Fair games and martingales

Let me use the phrase *known pure chance* for a setting where probabilities and payoffs are explicitly known. Roulette is the iconic example, or blackjack using a specified strategy. Mathematical probability textbooks teach you how to calculate pretty much anything you want in such settings. As

much as possible I avoid repeating such standard material in these lectures, whose focus is on understanding the real world outside the casino, where we generally do not have reliable prior knowledge of probabilities.

Recalling material from section 2.2, in the known pure chance setting, a bet is called *favorable* to a player if that player's mean gain is positive, *unfavorable* if the mean gain is negative (i.e. a loss) and *fair* if the mean gain is zero. (Here *mean* is *mathematical expectation*). Note the word *fair* here has a different meaning from its everyday one. The rules of team sports are *fair* in the sense of being the same for both teams, so the more skillful team is more likely to win, so a bet at even odds is not a *fair* bet.

The iconic example of a fair game is to bet on a sequence of coin tosses, winning or losing one dollar each time depending on whether you predict correctly. Of course this is too boring for anyone to actually do! In the corresponding mathematical model (often called *simple symmetric random walk*) write X_n for your "fortune" (amount of money) after n tosses.

Martingales. Before stating the definition, let me distinguish between two settings. The first is exemplified by the simple symmetric random walk model above, and the second by prices in a prediction market.

- Within a specified math model, a sequence (X_n) either is or isn't a martingale, depending on whether it satisfies the definition below.
- When X_n denotes some real-world quantity at "time n ", we *could* model (X_n) as some unspecified martingale. In some contexts there are compelling reasons why this is reasonable; in some other contexts there are plausible but vague arguments why this is reasonable. In contexts where one has enough data, one can check empirically whether quantitative predictions from martingale theory are correct.

This parallels closely the situation with *stationary process* to be treated in Lecture 7. There we will give plausible but vague arguments why it is reasonable to model English text as some (unspecified) stationary process.

The formal definition of *martingale* is a process satisfying

$$\mathbb{E}(X_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_0 = x_0) = x_n, \quad \text{all } x_0, x_1, \dots, x_n.$$

This (and the Wikipedia [Martingale](#) page) are hard to appreciate if you haven't taken a course in mathematical probability, so let me try to explain in words, in the context of gambling. What makes a single bet *fair* is that, starting with a known fortune x_0 , you will get a random fortune X_1 , and

your gain $X_1 - x_0$ has mean zero, that is $\mathbb{E}X_1 = x_0$. The martingale property is saying that at each time the expectation of your next gain, conditional on what has happened so far, has mean zero. If the odds offered to you are fair, then however much you choose to bet, this will be true.

For any underlying fair game, if you bet repeatedly, choosing how much to bet each time using some arbitrary strategy of your choice, which you may change and which may depend on what has happened so far, then the progress of your fortune X_n is a martingale.

What makes *martingale* a useful concept for doing calculations is a result known formally as the **optional stopping theorem** but more informatively called the *conservation of fairness theorem*³. Ignoring some technical restrictions, this says

Conservation of fairness theorem. *In the context of betting on a fair game using an arbitrary strategy, however you choose the time T to leave the game, your final fortune X_T will be such that $\mathbb{E}X_T$ equals your initial fortune, that is the same as some single fair bet.*

To illustrate how this is useful, here is a

Basic Formula for Fair Games. *Start betting on a fair game with x_0 dollars and continue until your fortune reaches either 0 or a target value $B > x_0$, whichever comes first, and without making any bet whose success would cause you to overshoot B . Then the probability of reaching B is exactly x_0/B .*

To derive this formula, the mean value of your final fortune is pB , where p is the probability above, and by the conservation of fairness theorem this must equal x_0 .

4.4 Prediction markets and martingales: conceptual issues.

I don't emphasize this material in the lecture.

³This name is used in Ethier's *The Doctrine of Chances: Probabilistic Aspects of Gambling* book, for instance.

The axiomatic setup of mathematical probability assumes there are “true probabilities” for events, which depend on the “information” known. Within this setup one can consider, in a sports match setting,

$X_n =$ probability team A wins, given the information available at time n

and this is a martingale, as a mathematical theorem. How does this apply to an actual baseball game? One view – a common default view, I believe – is that the set-up is conceptually correct, but that in complex settings like baseball or elections there is just no automatic procedure to assign numerical values to these probabilities⁴. Another view is that in the axiomatic set-up, “information” is represented as follows.

There was some previously specified collection of events whose outcomes (happen or not) would be known at time n , and now we do know these outcomes.

But this is so far removed from the way we actually perceive and act upon information, in the context of baseball and elections, that (in this view) its conceptual utility is dubious.

Deferring such conceptual questions to Chapter ??, it is natural to view prediction market prices as “consensus probabilities”; a price of 63 on a contract for team A to win represents a “consensus opinion” that there is a 63% for A to win. The point, of course, is that a person confidently assessing the chance as higher than 63% would buy the contract as a perceived “favorable bet”, pushing the price higher; another person assessing the chance as lower than 63% would sell the contract as a perceived “favorable bet”, pushing the price lower; so the price we observe represents the balance of opinions. How such a process actually works in detail is hard to say without introducing an awful lot of *supposes*. But the approximate identifications

prices \leftrightarrow consensus probabilities \leftrightarrow probabilities within axiomatic setup suggest as a plausible hypothesis that

(H) prices in a prediction market should behave (that is, fluctuate in time) approximately as a martingale.

Rather than examine further the assumptions or logic that went into stating this hypothesis, let us consider how to study it empirically. The mathematics of martingales, which we introduced in the previous section and will continue in the next section, provides a wide variety of theoretical predictions. Are

⁴This is not a frequentist vs. Bayesian issue – both use the same mathematics.

they correct, in real prediction markets? This makes an interesting topic for student course projects, to gather data and test theory for oneself, or to read academic papers which have done so.

4.5 Theoretical predictions for the behavior of prediction markets

Given a historical database of prediction market price fluctuations until the contracts expire, the most obvious prediction to test is that, out of all contracts starting at price around 40, about 40% do in fact end with the event happening. For this purpose, “around 60” is the same as “around 40” by considering the opposite event, and we can make more efficient use of data by considering each contract that ever crosses 40 or 60, taking the first such crossing as our initial time. Such data generally agrees with the theory, except in the case of very small initial prices (see section 4.9).

A skeptic might argue “maybe the real probabilities were sometimes 50% and sometimes 30% and these just averaged out to the predicted 40%”. The slightly more elaborate calculations below give predictions which allow one to distinguish between “a wide spread around 40%” and “a small spread around 40%”, given enough data.

Write a for some price less than the current price x (maybe $a = 0$, maybe $0 < a < x$) and write b for some price more than the current price x (maybe $b = 100$, maybe $x < b < 100$). Write $p_x(a, b)$ for the probability that the price reaches b before reaching a . By the same argument as for the Basic Formula⁵

$$p_x(a, b) = \frac{x - a}{b - a}. \quad (4.1)$$

Formula (4.1) can be used to get various interesting formulas, and we will give two of them.

Crossings of an interval. Fix a price interval, say $[40, 60]$. If the price is ever in this interval, then there is some first time the price crosses 40 or 60 – suppose it crosses 40. Either it sometime later crosses 60, or it expires at 0 without crossing 60, and from formula (4.1) the chance it reaches 60 equals $2/3$. If it reaches 60, call this a first “crossing”. From 60, it may (with chance $2/3$) cross 40 again (a second “crossing”) or it may expire at 100 without

⁵More precisely, this would be exact if prices varied continuously; it’s only an approximation when prices can jump, but in the present context it’s usually a good approximation. However it ignores the possibility of sudden events having large impact.

crossing 40. So there is some random number $C \geq 0$ of crossings, and from the argument above this number has the (shifted Geometric) distribution

$$\mathbb{P}(C = i) = \frac{1}{3} \left(\frac{2}{3}\right)^i, \quad i = 0, 1, 2, \dots \quad (4.2)$$

Remember this is all under the assumption that the price reaches 40 or 60 sometime. I have a collection of charts for 103 baseball matches like that in Figure 1 earlier. Of these, 89 reached 40 or 60, and here is the data for the observed number of crossings of the interval $[40, 60]$.

	0	1	2	3	4	5	6	7+
observed	33	29	14	8	2	1	1	1
predicted	29.7	19.8	13.2	8.8	5.9	3.9	2.6	5.2

The discrepancy for large values can perhaps be explained by lack of continuity (near the end of a close game a single hit may have a substantial effect, and there may be no trades between hits).

Maximum and minimum prices. For a contract starting at price x , either

- (i) team A wins, the contract expires at 100, and there is some overall minimum price L_x such that $L_x \leq x$;
- (ii) or team A loses, the contract expires at 0, and there is some overall maximum price L_x such that $L_x \geq x$.

The formula for the distribution of L_x is

$$\mathbb{P}(L_x < a) = \frac{a(100 - x)}{100(100 - a)}, \quad 0 < a < x \quad (4.3)$$

$$\mathbb{P}(L_x > b) = \frac{x(100 - b)}{100b}, \quad x < b < 100. \quad (4.4)$$

Here's how to derive these formulas. With starting price x , consider the first time T that the price reaches a or b , whichever happens first. The chance b is reached first is $p = p_x(a, b)$, in which case the price at T is b , and with chance $1 - p$ the price is a . So the optional sampling theorem says

$$x = \mathbb{E}(\text{price at time } t) = pb + (1 - p)a$$

and solving this equation for p gives formula (4.1).

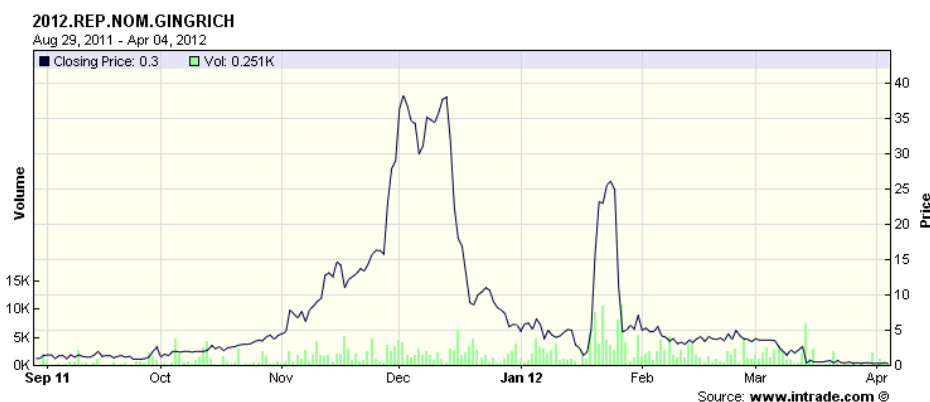
For $0 \leq a \leq x$,

$$\begin{aligned} \mathbb{P}(L_x < a, \text{ A wins}) &= \mathbb{P}(\text{price hits } a \text{ before } 100 | \text{initial price } 100) \times \mathbb{P}(\text{A wins} | \text{current price} = a) \\ &= (1 - p_x(a, 100)) \times \frac{a}{100} = \frac{100 - x}{100 - a} \times \frac{a}{100} \end{aligned}$$

which is formula (4.3); and formula (4.4) is derived similarly.

4.6 Were there improbably many candidates for the 2012 Republican nomination whose fortunes rose and fell?

Figure 4.3: Intrade price for Gingrich nomination.



In the race for the 2012 Republican Presidential nomination there were many candidates whose popularity rose and then fell noticeably – Donald Trump, Newt Gingrich (whose prediction market price is shown in Figure 4.3, Sarah Palin, Rick Perry, Michelle Bachmann, for instance. Many discussions of the race have shared the presumption that the number of such candidates was much larger than usual, and speculated on the reasons, e.g. an “anyone but Romney” sentiment. But is that presumption true?

We need to distinguish between two meanings. Opinion polls ask questions in a format “if you were voting tomorrow, who would you vote for?”. Mathematics says nothing about how much such opinions may fluctuate over a year-long campaign, just as mathematics says nothing about how much fashions in popular music may fluctuate. One could devise some statistic to measure these fluctuations and compare it empirically with the statistics from previous races, but one cannot compare it to any theoretical prediction.

On the other hand, the theoretical argument that every prediction market price should be a martingale is not affected by fashion or opinion poll results. So we can examine whether the prediction market prices in this particular race behaved differently from how theory says prediction market prices should behave, which would be an indication of some unusual aspect

of the 2012 race. Here is one interesting prediction of theory, which I call the **serious candidates principle**.⁶

Consider an upcoming election with several candidates, and a (prediction market) price for each candidate, and suppose initially all these prices are below b , for given $0 < b < 100$. Theory says that the mean number of candidates whose price ever exceeds b equals $100/b$.

Here is the mathematical argument, based on a hypothetical betting system. For each candidate, buy a contract on that candidate if and when their price reaches b . The total cost of these contracts is bN_b , where N_b is the random number of candidates whose price ever reaches level b . Exactly one candidate is elected, and your contract on that candidate earns you 100. So your gain is $100 - bN_b$. The conservation of fairness theorem says the expected gain equals zero, and the equation $\mathbb{E}[100 - bN_b] = 0$ rearranges to $\mathbb{E}N_b = 100/b$.

Some data. Table 4.1 shows the maximum (over time) Intrade prediction market price for each of the 16 leading candidates for the 2012 Republican Presidential Nomination.

Table 4.1: Maximum prediction market prices.

Romney	100	Perry	39	Gingrich	38	Palin	28
Pawlenty	25	Santorum	18	Huntsman	18	Bachmann	18
Huckabee	17	Daniels	14	Christie	10	Giuliani	10
Bush	9	Cain	9	Trump	8.7	Paul	8.5

These numbers might well suggest to a non-mathematician that the number of sometime-serious candidates was unusually large. But look at Table 4.2 below, which compares observed data with the mathematical prediction for “number of candidate with maximum price $\geq b$ ” for several values of b .

Table 4.2 indicates that the number of candidates whose fortunes rose and fell in this “probability of winning” sense was scarcely more than would be expected on mathematical grounds.

⁶As in previous formulas, the only assumption we need is that each candidate’s price is a continuous-path martingale. This corresponds to the idea of a “liquid market” with small spread between bid and ask prices, which is reasonably accurate for the election markets under consideration.

4.6. WERE THERE IMPROBABLY MANY CANDIDATES FOR THE 2012 REPUBLICAN NOMINATION

Table 4.2: Observed and expected numbers exceeding threshold prices.

	expected	observed
$b = 33.3$	3	3
$b = 20$	5	5
$b = 16.6$	6	9
$b = 12.5$	8	10

Two technical points. In Table 2 we used $100/b$ as “expected”, without considering whether some initial prices might have been greater than b . Data on initial prices is somewhat unreliable (because the contract may initially be thinly traded) but the only candidate whose initial price was clearly above 10 was Romney at about 23. Correcting for this would make the “expected” numbers slightly smaller for small b . Of course for a campaign where two candidates started with price 40 the “expected” numbers would be very different. Another important general point is that, for long-duration contracts, low prediction market prices overstate the true consensus probability because of the “covering your position” requirement. That is, even if you were certain an event would not happen, you might not be willing to sell a contract for 3 because your sure gain of 3 is offset by the opportunity or interest cost of the market requirement that you deposit 97 to cover a possible loss. Correcting for this effect would make the “expected” numbers in Table 2 larger than shown for small b .

A bottom line conclusion. To the extent that mathematics can say anything relevant, it says that the fundamental driving feature of the 2012 nominee campaign was that it started without any clear favorite. The subsequent fluctuations were then consistent with what theory predicts. In other words, even if it is actually true that the month-to-month fluctuations in opinion poll standings were greater than usual, we can see no sign that this unduly influenced the smart money being wagered on the prediction market.

Another check of theory and data. A mathematician familiar with martingale theory might look at the Figure 4.3 chart for Newt Gingrich and wonder if it shows too many fluctuations to be plausibly a martingale. For instance, the chart shows two separate downcrossings from 20 to 10, in December 2011 and in late January 2012. This mathematician has in mind

the *upcrossing inequality*⁷ which limits the likely number of such crossings. We can conduct another check of theory versus data by considering crossings. The relevant theory turns out to be:

Consider a price interval $0 < a < b < 100$, and consider an upcoming election with several candidates, and a (prediction market) price for each candidate, where initially all these prices are below b . Theory says that the expected total number of downcrossings of prices (sum the numbers for each candidate) over the interval $[a, b]$ equals $(100 - b)/(b - a)$.

To gather data for the interval $[10, 20]$, we need only look at the five candidates in Table 4.1 whose maximum price exceeded 20, and their numbers of downcrossings of $[10, 20]$ were:

Palin (2); Romney (0); Perry (1); Pawlenty (2); Gingrich (2).

So the observed total 7 is in fact close to the theoretical expectation of 8. To derive the formula quoted, we again consider a hypothetical betting system. For each candidate, buy a contract on that candidate if and when their price reaches b . If the price subsequently falls to a then sell; but buy again if the price reaches b , and continue. Exactly one of these contracts will expire at 100, and the others will be sold at price a , the number $D_{a,b}$ of these others being the number of downcrossings of $[a, b]$. So your gain is $(100 - b) - D_{a,b}(b - a)$. The conservation of fairness theorem says the expected gain equals zero, and the equation $\mathbb{E}[(100 - b) - D_{a,b}(b - a)] = 0$ rearranges to $\mathbb{E}[D_{a,b}] = (100 - b)/(b - a)$.

4.7 The halftime price principle.

Here we detour away from martingales to see a more elementary piece of theory which can be checked against data.

The halftime price principle. In a sports match between equally good teams, at halftime there is some (prediction market) price for the home team winning. This price varies from match to match, depending largely on the scoring in the first half of the match. Theory says its distribution should be approximately uniform on $[0, 100]$.

⁷See any textbook with a chapter on martingales.

To elaborate this principle we imagine a sport in which (like almost all team sports) the result is decided by point difference, and for simplicity imagine a sport like baseball or American football where there is a definite winner (ties are impossible or rare). Also for simplicity we assume the teams are equally good, in the sense that there is initially a 50% probability of the home team winning (that is, equally good after taking home field advantage into account). Write Z_1 for the point difference (points scored by home team, minus points scored by visiting team) in the first half, and Z_2 for the point difference in the second half.

A fairly realistic mathematical model of this scenario is to assume:

- (i) Z_1 and Z_2 are independent random variables, with the same distribution;
- (ii) their distribution is symmetric about zero; that is, their distribution function $F(z)$ satisfies $F(z) = 1 - F(-z)$.

For mathematical ease we add an unrealistic assumption (to be discussed later):

- (iii) the distribution is continuous.

Under these assumptions we can do a calculation, though we first recall the slightly sophisticated notation that treats conditional probabilities as random variables. For an event A and a random variable Y , the elementary notation for conditional probabilities

$$\mathbb{P}(A|Y = y) = g(y) \text{ for all } y$$

(the left side is always *some* function of y) can be rewritten as

$$\mathbb{P}(A|Y) = g(Y). \tag{4.5}$$

The following calculation exemplifies the usefulness of this notation.

The probability that the home team wins, given the first half point difference is z , is

$$\begin{aligned} \mathbb{P}(Z_1 + Z_2 > 0|Z_1 = z) &= \mathbb{P}(Z_2 > -z) \text{ by independence} \\ &= F(z) \text{ by symmetry} \end{aligned}$$

and therefore the price at halftime, which is the conditional probability of the home team winning, given the observed value of Z_1 , is

$$\mathbb{P}(Z_1 + Z_2 > 0|Z_1) = F(Z_1). \tag{4.6}$$

But as a textbook fact, for a continuous distribution it is always true that $F(Z_1)$ has uniform distribution on (0%, 100%).

That is the mathematical justification for the principle. One can think of various defects in the model, most obviously the fact that in real sports the points are integer-valued, but for reasons explained below, we suspect this does not make a huge difference, even in the worst case of a low-scoring sport like soccer.

A little data.

Errors using inadequate data are much less than those using no data at all. [Charles Babbage]

In the Figure 4.1 baseball match chart, the initial price was near 50 and the price at half-time (for baseball we simply used halfway through the match duration) was around 62.

In 30 baseball games from 2008 for which we have the prediction market prices as in Figure 4.1, and for which the initial price was around 50%, the prices (as percentages) halfway through the match were as follows:

07, 10, 12, 16, 23, 27, 31, 32, 33, 35, 36, 38, 40, 44, 46
50, 55, 57, 62, 65, 70, 70, 71, 73, 74, 74, 76, 79, 89, 93.

Figure 4.4 (left) compares the distribution function of this data to the (straight line) distribution function of the uniform distribution. The data appears roughly consistent with our halftime price principle.

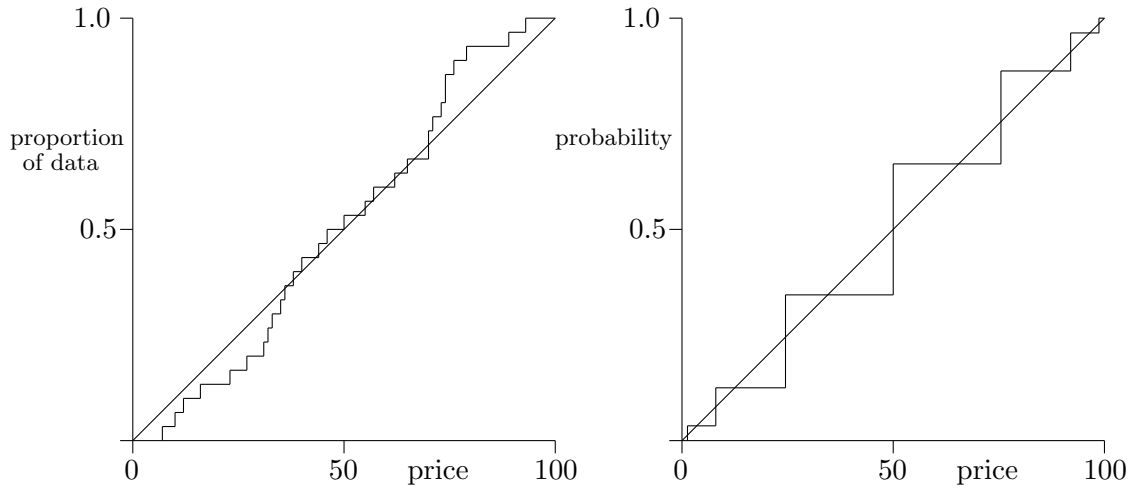
Caveat. The simplicity of the stated halftime price principle depends on the teams being equally good. For unequal teams the distribution of halftime price will depend on the distribution of the point differences Z_i as well as the initial price.

4.8 Other martingale calculations

One can sometimes do mathematical calculations which are not directly about gambling by considering hypothetical bets. This is sufficiently striking that in the example I will break my rule of not discussing unmotivated coin-tossing type problems.

Waiting times for patterns in coin-tossing. If you toss a fair coin repeatedly, how long before you see the pattern HTHT? It turns out the mean number of tosses equals 20.

Figure 4.4: The left diagram shows the empirical distribution function for the baseball data, compared with the uniform distribution. The right diagram shows the theoretical distribution function in the soccer model, again compared with the uniform distribution.



Here's the “gambling” argument we wish to spotlight (there are several other ways to solve this problem). Imagine a casino where you can bet at fair odds on the coin tosses. For each toss, say toss 7, imagine a gambler following the strategy

bet 1 that toss 7 is H

if win, bet all (now 2) that toss 8 is T

if win, bet all (now 4) that toss 9 is H

if win, bet all (now 8) that toss 10 is T

If win, walk away with your 16. If lose sometime, walk away with 0.

Recall that for each toss a different gambler starts this strategy. Look at what happens from the casino's viewpoint, up to and including the random toss (number S) on which the pattern HTHT is first completed. The casino has taken in 1 from each of S gamblers. It has paid out 16 to the gambler whose first bet was on toss $S - 3$. It has paid out 4 to the gambler whose first bet was on toss $S - 1$, but nothing to the other gamblers. So the casino's gain is $S - 20$. But by the Fair Game Principle, the mean gain must be zero.

4.9 Stock markets and prediction markets

Lecture 2 discussed long term stock market investment, as the prime real-world instance of “gambling on a favorable game”. Let me compare and contrast the two.

- A prediction market is conceptually simpler than a stock market because the final value corresponds to a definite event. Saying a stock index ends the year at 1321 says nothing other than it ends at 1321.
- A prediction market is mathematically simpler because we need no empirical data to make the theoretical predictions; for the analogous predictions in a stock market one needs an estimate of variance rate.
- Compared to stock markets, prediction markets are very thinly traded, suggesting they will be less efficient and less martingale-like. In particular, for a year-duration contract with a small price (say 5) on Intrade, the true probability might be much smaller, because to make the “sure thing” profit of 5 by betting against the event requires you to cover the contract and tie up 95 capital for a year.
- Standard economic theory asserts that long-term gains in a stock market will exceed long-term rewards in a non-risky investment, because investors’ risk-taking must be rewarded. In this picture, a stock market is a “positive sum game” benefiting both investors and corporations seeking capital; financial intermediaries and speculators earn their share of the gain by providing liquidity and convenient diversification for investors.
- In contrast a prediction market is a zero-sum game, in fact a slightly negative-sum game because of transaction costs.

The recent history of financial markets might provoke skepticism about “standard economic theory” in general, and in particular the *efficient market hypothesis* which appears in Lecture 2.

Prediction markets could in principle serve a non-gambling function via hedging real-world risks, but currently do not provide enough liquidity to do so.

4.10 Wrap-up

In lectures and public talks I used to choose some current event on the Intrade prediction market trading around 50, ask the audience for their

opinions, and buy or sell a few shares based on the majority opinion. And then track the price. The results of 9 such investments [can be found here](#).

The realization that

- (a) the general notion (as opposed to specific bets on specific games) of “results of bets at fair odds” has a precise formalization as *martingale*,
- (b) the way the current probability of a future event changes with time, as new information comes in, also behaves mathematically as a martingale

was one of the triumphs of twentieth century mathematical probability. Prediction markets provide the most concrete real-world instance of observable quantities (prices, here) that can be directly interpreted as probabilities. The mathematical theory of martingales can be used to make testable predictions about the behavior of prices in prediction markets.

On the other hand, the mathematical setup does not correspond well to our intuitive sense of how we actually make decisions under uncertainty. It is easy to make the assertion that the prediction market price obtained via different agents’ choices to act on their probability assessments gives a “consensus probability” that will behave as in the mathematical model of “information”. But hard to expand upon an argument for how this happens in detail.

4.11 Further reading

Broader ideas underlying prediction markets are often discussed under the phrase [The Wisdom of Crowds](#), from the title of a 2004 book by James Surowiecki. Though a cute title for a book, it is another example of an ill-chosen name for a concept. A prediction algorithm, whether based on software or wetware, does not constitute *wisdom*.

As I have already mentioned several times, these lectures do not focus on games of pure chance, which are treated in many existing works. Two books I recommend are Haigh’s *Taking Chances* at an elementary mathematical level, and Ethier’s *The Doctrine of Chances: Probabilistic Aspects of Gambling* for material going beyond typical textbook material.

A World of Chance: Betting on Religion, Games, Wall Street by Brenner, Brenner and Brown contains an interesting, mainly historical, account of the relation between the laws and culture surrounding gambling and the laws and culture surrounding business and market activity. *Probability with Martingales* by David Williams shows one can teach an advanced course on mathematical probability centered around martingales as a technical tool.

Chapter 5

Game Theory: Introducing Nash equilibria via some actual data

This lecture is rather more mathematical and more narrowly focussed than others. Berkeley has an entire course on Game Theory that some of my students take. Rather than give a 1-lecture overview, I jump quickly into the concept of Nash Equilibria and focus on an example where one can join a game in real time and also carry through some mathematical analysis. Here is a link to [an extended version of this write-up](#). Finally I talk briefly about the Least Unique Positive Integer game, which is fun to play in class.

There are many introductory textbooks on **Game Theory**, and I imagine all my students have heard of the subject, so let me start with only a quick bullet point overview.

1. Setting: players each separately choose from a menu of actions, and get a payoff depending (in a known way) on all players' actions.
2. Rock-paper-scissors illustrates why one should use randomized strategy, and why we assume a player's goal is to maximize their expected payoff. There is a complete theory of such two-person zero-sum games.
3. For other games, a fundamental concept is a **Nash equilibrium** strategy: one such that, if all other players play that strategy, then you cannot do better by choosing some other strategy. This concept can be motivated mathematically by the idea that, if players adjust their strategies in a selfish way to maximize their own payoff, and if the strategies converge

to some limit strategy from which a player cannot improve by further adjustment, then by definition the limit strategy is a Nash equilibrium.

4. More advanced theory is often devoted to settings where Nash equilibria are not optimal in some sense, as with **Prisoners' Dilemma**, and to understanding why human behavior is not always selfish.

This lecture will focus on point 3. We will briefly play, and then analyze, a specific game which, to a game theorist, fits exactly into the setting of point 3. The “learning-adjust” theory predicts that players who play repeatedly and play selfishly – being unable or unwilling to collaborate with other players – will tend to adjust their strategies to approximate the Nash equilibrium (which we now abbreviate to NE) strategy. Further discussion of NE can be found in many textbooks. Instead of general discussion, what I will do in this lecture is

- calculate the NE strategy in somewhat simplified versions of the real game;
- compare this with the data on what players actually do.

I do not seek to introduce and explain much standard game-theory terminology – for instance, the concept of NE refers, strictly speaking, to a *strategy profile*, that is a strategy for each player, but in our “symmetric over players” context we look only for NE strategy profiles in which each player uses the same strategy.

5.1 The specific game: Dice City Roller

The game is pogo.com’s *Dice City Roller (DCR)*.

In class, I start by spending a couple of minutes demonstrating the game by actually participating in it, in real time. In this write-up the written description of the DCR game is deferred to section 5.5; readers may wish to read it now, or go online and play it themselves, before reading further.

For our mathematical analysis, the following abstracted model of the game is sufficient, *with italicized comments on actual play*.

- There are M items of somewhat different known values, say $b_1 \geq b_2 \geq \dots b_M$ (*always $M = 5$, but the values vary between instances of the game*).
- There are N players (*N varies but 5 – 12 is typical*).

- A player can place a sealed bid for (only) one item, during a window of time (*20 seconds*).
- During the time window, players see how many bids have already been placed on each item, but do not see the bid amounts.

Of course when time expires each item is awarded to the highest bidder on that item. We assume players are seeking to maximize their expected gain. So a player has to decide three things; when to bid, which item to bid on, and how much to bid.

It turns out that without the time element (that is, if players make sealed bids without any information about other players' bids) the game above is completely analyzable, as regards Nash equilibria. This is the mathematical content of this lecture, and the results are broadly in line with intuition.

The time element makes the game more interesting, because various strategies suggest themselves: bid late on an item that few or no others have bid on, seeking to obtain it cheaply, or bid early on a valuable item to discourage others from bidding on it. Alas theoretical analysis seems intractable, at least at an undergraduate level.

We obtained data from 300 instances of the DCR game, and various statistical aspects of the data are shown in section 5.4. As mentioned above we do not have a precise formula for the NE strategy in the real game, but nevertheless we can formulate plausible approximations to the NE strategy. And the bottom line, discussed in section 5.4, is rather ambiguous. On one hand the “ordinary people” playing this game are not bidding in a way that is close to the NE strategy, but their deviations are not “foolish” in any specific way.

5.2 Analysis of the simplest case

We study the simplified version without the time window – each player just places a sealed bid without knowledge of other players' actions. We first study the simplest setting: 2 players, 2 items of values 1 and b , where $0 < b < 1$.

A player's strategy is a pair of functions (F_1, F_b) :

$$F_1(x) = \mathbb{P}(\text{bid an amount } \leq x \text{ on the first item}), \quad 0 \leq x \leq 1 \quad (5.1)$$

$$F_b(y) = \mathbb{P}(\text{bid an amount } \leq y \text{ on the second item}), \quad 0 \leq y \leq b \quad (5.2)$$

where

$$F_1(1) + F_b(b) = 1. \quad (5.3)$$

We can equivalently work with the associated densities

$$f_1(x) = F_1'(x), \quad f_b(y) = F_b'(y).$$

Suppose your opponent's strategy is some function (f_1, f_b) and your strategy is some function (g_1, g_b) . The formula for your expected gain is

$$\int_0^1 (1-x)g_1(x)[F_1(x) + F_b(b)] dx + \int_0^b (b-y)g_b(y)[F_b(y) + F_1(1)] dy. \quad (5.4)$$

We need the following fact, which is obvious when you sketch a diagram.

Given a payoff function $h(x) \geq 0$ with $h^* = \max_x h(x)$, consider the expected payoff $\int h(x)g(x)dx$ when we draw x from a probability density g which we may choose. Then we get the maximum expected payoff if and only if we choose g with the properties

$$h(x) = c \text{ for all } x \in \text{support}(g)$$

$$h(x) \leq c \text{ for all } x \notin \text{support}(g)$$

for some c (which is in fact h^*).

Now our expected gain (5.4) is of this form, thinking of (g_1, g_b) as a single probability density function. Applying the "obvious fact" above we deduce the following. Given your opponent's strategy (f_1, f_b) , your expected gain is maximized by choosing a strategy (g_1, g_b) satisfying, for some constant c

$$\begin{aligned} (1-x)[F_1(x) + F_b(b)] &= c \text{ on support}(g_1) \\ &\leq c \text{ off support}(g_1) \\ (b-y)[F_b(y) + F_1(1)] &= c \text{ on support}(g_b) \\ &\leq c \text{ off support}(g_b) \end{aligned}$$

Now the definition of (f_1, f_b) being a *Nash equilibrium* strategy is precisely the assertion that these relations hold for $(g_1, g_b) = (f_1, f_b)$. So now we have a set of relations for the NE strategy

$$(1-x)[F_1(x) + F_b(b)] = c \text{ on support}(f_1) \quad (5.5)$$

$$\leq c \text{ off support}(f_1) \quad (5.6)$$

$$(b-y)[F_b(y) + F_1(1)] = c \text{ on support}(f_b) \quad (5.7)$$

$$\leq c \text{ off support}(f_b) \quad (5.8)$$

with “boundary conditions”

$$F_1(0) = F_b(0) = 0; F_1(1) + F_b(b) = 1.$$

To digress for a moment, note that in any game we can do some similar argument to get equations for a NE. Most introductory game theory focusses on a discrete menu of actions – our example is continuous. Theory talks about existence and uniqueness of solutions, for general games. But we can just go ahead and solve these particular equations without using their game-theoretic origin (see [the extended version of this write-up](#) for details) and the result is quoted below, though we will see in section 5.3 that it’s easier to solve them using game-theoretic principles. The solutions are

$$F_1(x) = \frac{b}{1+b} \left(\frac{1}{1-x} - 1 \right) \text{ on } 0 \leq x \leq \frac{1}{1+b} \quad (5.9)$$

$$F_b(y) = \frac{1}{1+b} \left(\frac{b}{b-y} - 1 \right) \text{ on } 0 \leq y \leq \frac{b^2}{1+b}. \quad (5.10)$$

The corresponding densities are

$$f_1(x) = \frac{b}{1+b} (1-x)^{-2} \text{ on } 0 \leq x \leq \frac{1}{1+b} \quad (5.11)$$

$$f_b(y) = \frac{b}{1+b} (b-y)^{-2} \text{ on } 0 \leq y \leq \frac{b^2}{1+b} \quad (5.12)$$

and the expected gain for each player works out as

$$\mathbb{E}[\text{gain}] = \frac{b}{1+b}.$$

In the 2014 course I had each student “play this game once” by making a single bid, in the case $b = 1/2$, so we can now compare their bids to the NE distribution. The top two frames in Figure 5.1 compare the NE distribution functions F_1 and F_b at (5.9, 5.10) with the corresponding empirical distribution functions G_1 and G_b from the data. The bottom two frames in Figure 5.1 compare the NE expected gain from bidding different amounts with the corresponding empirical mean gain from the amounts bid by students. That is, a bid of 49 cents on the \$1 item had, when matched against a random other bid, mean gain of 29 cents, and this is represented by a point at (49, 29).

Here the data is not close to the NE. Students had some apparent intuition to bid around 50 cents on the \$1, and those who bid on the 50 cent items tended to overbid. But recall that the NE concept is motivated by the idea that, if players play repeatedly and adjust their strategies in a self-ish way, then strategies should typically converge to some NE. So it is not reasonable to expect NE behavior the first time a game is played.

In contrast, the actual DCR game is played repeatedly and so it is more meaningful to ask whether players’ strategies do in fact approximate the NE.

Figure 5.1: Class data compared with the NE.



The form of the NE exhibits one general principle about NE and another feature special to this model. The former is what I will call

The constant expected gain principle. If opponents play the NE strategy then any non-random choice of action you make in the support of the NE strategy will give you the same expected gain (which equals the expected gain if you play the random NE strategy), and any other choice will give you smaller expected gain.

This is true because the NE expected gain is an average gain over the different choices in its support; if these gains were not constant then one would be larger than the NE gain, contradicting the definition of NE. In our game, if you bid x on item 1, where x is in the support $0 \leq x \leq \frac{1}{1+b}$, then your chance of winning is (by calculation) $\frac{b}{1+b}(1-x)^{-1}$, so your expected gain is $(1-x) \times \frac{b}{1+b}(1-x)^{-1} = \frac{b}{1+b}$ as the constant expected gain principle says.

Now note that the gap between your maximum bid and the item's value is the same for both items;

$$1 - 1/(1+b) = b - b^2/(1+b) = b/(1+b).$$

This follows from the “constant expected gain” principle above; if you bid the maximum value in the support the you are certain to win the item, so your gain must be the same for both items. This fact – call it the **equal gap principle** – is special to the structure of this particular game, but is true for general numbers of players and items.

5.3 General numbers of players and items

Perhaps surprisingly, armed with the two principles above we can rather easily calculate explicitly the NE in the general case of $N \geq 2$ players and $M \geq 2$ items of values $b_1 \geq b_2 \geq \dots \geq b_M > 0$. The bottom line (with a side condition I'll explain) is the formula

$$\mathbb{E} (\text{gain to a player at NE}) = c = \left(\frac{M-1}{\sum_i b_i^{-1/(N-1)}} \right)^{N-1} \quad (5.13)$$

and the NE strategy is defined by the density functions

$$f_i(x) = \frac{M-1}{N-1} \frac{1}{\sum_j b_j^{-1/(N-1)}} (b_i - x)^{-N/(N-1)}, \quad 0 \leq x \leq b_i - c \quad (5.14)$$

for bids on prize i .

Here are the main steps in the calculation. Writing out the expression for the expected gain when you bid x_i on the i 'th item, the “constant expected gain” property says

$$(b_i - x) (1 - (F_i(x_i^*) - F_i(x)))^{N-1} = c, \quad 0 \leq x \leq x_i^* := b_i - c \quad (5.15)$$

where c = expected gain to a player at NE. Because a strategy is a probability distribution we have $\sum_i F_i(x_i^*) = 1$ and so

$$\sum_i (1 - F_i(x_i^*)) = M - 1.$$

Now using (5.15) with $x = 0$ we have

$$1 - F_i(x_i^*) = (c/b_i)^{1/(N-1)} \quad (5.16)$$

and so

$$\sum_i (c/b_i)^{1/(N-1)} = M - 1$$

identifying c .

The side condition. In the analysis above we implicitly assumed that the NE strategy included a bid on each item (*include* means “assigns non-zero probability to”). This may not be correct if, for instance, there was one prize with very small value. To take care of this issue, recall we order item values as $b_1 \geq b_2 \geq \dots \geq b_M > 0$. Inductively for $m = 2, 3, \dots, M - 1$ calculate the NE and the expected gain assuming we have only the first m items available. If the expected gain is greater than b_{m+1} then stop and use this NE strategy which does not include a bid on any of b_{m+1}, \dots, b_M . Otherwise continue to $m + 1$. However, we only need to do this procedure if the original formula (5.13) for expected gain is manifestly wrong, in giving a value greater than the smallest value b_M .

5.4 Comparing data from the DCR game with NE theory

Several complications arise when we seek to compare data from the DCR game with NE theory; we mention them briefly here and at greater length in [the extended version of this write-up](#).

5.4. COMPARING DATA FROM THE DCR GAME WITH NE THEORY 79

1. In the DCR game there is a minimum allowed bid on each item, but fortunately the NE analysis extends easily to this case.

2. In each auction there are 5 items, from a set of 12 different items, and the prizes, if you win a bid, are a random number of points, depending on which of the 12 items. In our mathematical analysis we will take the prizes to be the (non-random) expected value for each item. Our assumption is that players learn by experience the approximate value of these expected values (by observation one can calculate them exactly).

3. Strategic effects, mentioned earlier, resulting from the time window are not taking into account, except in a minor way indicated below.

4. The observable data in the DCR game is the number of bids, and the value of the winning bid, on each item – but we cannot see the values of losing bids. So, for a given pair (N, i) of (number of players, item), the data we have available is the empirical distribution of values of winning bids over auctions where there was at least one bid. This is plotted, for 4 of the items, as a distribution function G^* in Figure 5.2. We want to compare that to a “NE theory” distribution, and we obtain this by assuming that the amounts of bids follow the NE distribution ((5.14) modified for minimum allowed bid), but (to allow for strategic effects) we use the true empirical distribution for the *number* of bids. Then we can numerically calculate a “NE theory” distribution function for value of winning bid, and this is plotted as a distribution function G in Figure 5.2.

Figure 5.2 shows the comparison between data and NE theory. The labels “150 match” etc are our names for some of the items (explained in section 5.5), and this data is for $N = 8$ players.

One’s first reaction to the Figure 5.2 data is that the players’ bids are not very close to what NE theory would predict. One could imagine many reasons for this discrepancy. A typical player self-description is “age 63, retired nurse: interests church, crafts, grandkids”; on this basis we suppose the typical player is not a student of game theory, so might not consider the idea of conscious randomization. The fact that the winning bid is, in roughly a third of these cases, the minimum allowed bid is clearly a consequence of time-window strategy (making a last-second bid on an item no-one else has bid on) not taken into account in our theory, so the data might be closer to the true NE than to our approximate NE.

Figure 5.2: Comparison of winning bid distribution from data and from NE theory.

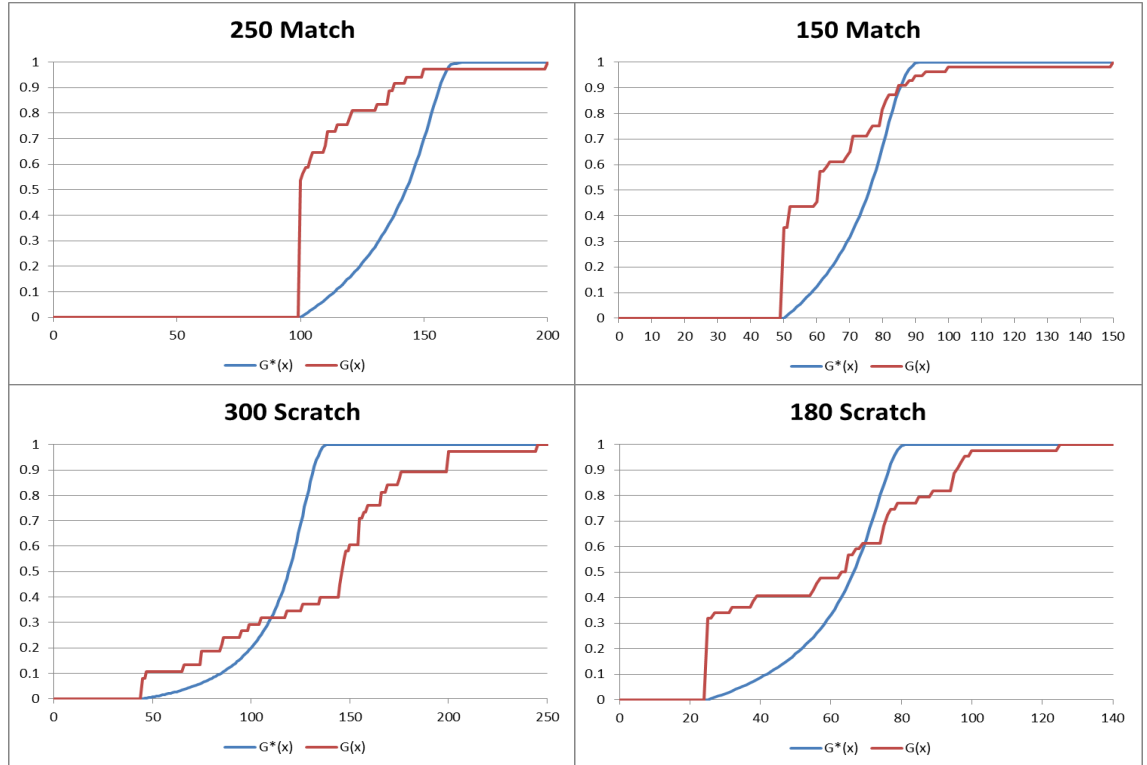


Figure 5.3: Screenshot of basic play of the underlying game in DCR.



5.5 Dice City Roller

The game *Dice City Roller* (DCR) motivating this article is found on pogo.com, a free online casual gaming website offering over 150 different games. At a typical time there may be about 10 different active “rooms” each containing typically having 5 - 15 competing players – other rooms with 1 or 2 players do not concern us. The underlying game is illustrated by the screenshot in Figure 5.3 (details below not relevant to our mathematics, until further notice). An instance of the game consists of 12 repetitions of the following “turn”. The player is shown five rolled dice, allocates them onto “cards” to fill out specified combinations (over several turns); when a card is completed the player earns points and a new card is offered.

For instance, in Figure 5.3 the rolled dice show 1, 6, 2, 4, 1. The player could place a 1 on the Full House, place a 2 and 6 on the Straight and place a 1 on the 3 Of A Kind, to complete 3 cards, placing the remaining 4 on one of the other cards. The player has 15 seconds to decide upon and execute these placements.

This underlying game is more subtle than it may appear, because there is a bonus for completing several cards on the same turn, so a simple greedy

Figure 5.4: Screenshot of auction in progress.



scheme for filling cards is not optimal. However this activity is not game-theoretic, because there is no interaction between players – one simply seeks to maximize one’s score, that is one’s total number of points at the end of the game.

What is relevant to us is the “auction” version which adds the following step, 3 times during the game. Players are allowed to use some of their points to bid for one of 5 prizes, a prize being the chance to earn extra points. The bidding proceeds as described earlier:

During a 20 second time window, players see how many bids have already been placed on each item, but do not see the bid amounts.

The screenshot in Figure 5.4 shows a situation 5 seconds before the window closes. Three players have placed bids, on different cards – these numbers of bids are shown in the disc at the cards’ bottom left corner. Three other players had not yet placed bids. In the 5 seconds remaining after the screenshot, it happened that two players bid on the top right card (with 0 earlier bids) and one player bid on the bottom left card (with 1 earlier bid).

In each auction there are 5 cards, from a set of 12 different cards. The prizes, if you win a bid, are a random number of points. As mentioned earlier, in our

mathematical analysis we took the prizes to be the (non-random) expected value for each card.

More details. As seen in Figure 5.4, each card shows the minimum bid allowed, the maximum possible prize and the “type” which is mostly *match* or *scratch*: our names like “150 match” refer to type and maximum prize. In “150 match” there are 6 covered numbers, and the winning bidder uncovers each until finding two equal numbers, and that number becomes the prize. One can learn that the 6 covered numbers are 50, 100 and 150, with two copies of each. So the prize is equally likely to be 50, 100, 150, with expectation 100. In a “scratch” card there are also 6 covered numbers; except that one is a bomb; the player uncovers numbers until reaching the bomb, and the prize is the sum of the values uncovered. For such a “scratch” card the maximum prize is the sum of the 5 numbers and the expected value is **exactly** half of this maximum. Learning all these numerical values requires careful observation, and we suspect typical players do not explicitly know these expected values. In particular, for “match” cards the expected value is always **more than** half of the maximum prize shown on the card. A player unaware of this distinction is liable to underbid on the “match” cards or overbid on the “scratch” card, which appears to be happening in the Figure 5.2 data.

5.6 The Least Unique Positive Integer game

This is another game for which we expect a unique NE, and which is easy and quick to play in a lecture class, needing only pen and paper.

Least Unique Positive Integer game. Each of N players chooses a number from $1, 2, 3, \dots$. The winner is the person who chooses the smallest number that no-one else chooses.

There might be no winner, but this is very unlikely except for very small N .

Here a player’s strategy is a probability distribution $\mathbf{p} = (p_1, p_2, \dots)$, where p_i is the probability of choosing i . Let me outline the easy approximate analysis of the NE, for reasonably large N .

We expect the support of the NE to be $1 \leq i \leq K$ for some K depending on N . If other players use \mathbf{p} then

$$X_i = \text{number others choosing } i \approx \text{Poisson}(\lambda_i = (N - 1)p_i)$$

The “constant expected gain” principle says that, whatever your choice of i in the support $[1, K]$, your chance of winning is $c \approx N^{-1}$. Choosing i , you win if no-one else chooses i and there is no unique chooser of any $j < i$, giving approximate equations

$$\mathbb{P}(X_i = 0, X_j \neq 1 \forall j < i) = 1/N, \quad 1 \leq i \leq K.$$

For the left side there is a (complicated) formula in terms of \mathbf{p} , which can be solved numerically. In particular, for $i = 1$ we see

$$\exp(-(N-1)p_1) \approx 1/N$$

and so

$$p_1 \approx \frac{\log N}{N-1}.$$

Then p_i decreases, slowly at first – the reader can find graphics in the paper cited below. Playing this in class with around 30 students, the winning number is typically in the 3 - 7 range, consistent with NE theory.

This game was played on a large scale (around 50,000 players) for money prizes in Sweden, 7 times in 2007. Analysis in [this 2010 paper *Testing game theory in the field: Swedish LUPI lottery games*](#). shows data roughly consistent with NE theory. However, both theory and actual results show the winning number is likely to be at most 7,000. Knowing that, one could buy (or form a coalition of players to buy) a ticket with each number in the 1 - 7000 range and almost guarantee a win. For this reason (presumably) the game was stopped after 7 weeks. Note that a “buy all the tickets” strategy is rarely advantageous in a usual lotto game because all winning combinations share the prize; what is different about this game is the *unique* winner.

5.7 Further reading

Game theory is an appealing mathematical topic, and there are perhaps a hundred books giving introductory accounts in different styles. Styles using minimal mathematics range from “popular science” (Len Fisher’s *Rock, Paper, Scissors: Game Theory in Everyday Life*) to airport bookstore Business section bestseller (Dixit - Nalebuff *Thinking Strategically: The Competitive Edge in Business, Politics, and Everyday Life*). A wide-ranging account with a modicum of mathematics is provided in *The Complete Idiots Guide to Game Theory* while careful rigorous expositions at a lower division mathematical level can be found in Saul Stahl’s *A Gentle Introduction to Game Theory* or Philip Straffin’s *Game Theory and Strategy* or the recent e-book

Game Theory Through Examples by Erich Prisner. A representative of the numerous textbooks aimed at students of Economics is Robert Gibbons's *Game Theory for Applied Economists*, and an erudite overview from that viewpoint is given in David Kreps's *Game Theory and Economic Modelling*. But such books either contain no real data, or occasionally quote data obtained by others.

Chapter 6

Short and medium term predictions and risks in global politics and economics.

I have no crystal ball for making accurate predictions, but we can look at how well others have done. To get started, what does “prediction” mean and how do we assess its accuracy?

In this lecture I am not thinking of “routine” issues – predicting election results or macroeconomic indicators in a particular country a few months ahead – but of more substantial or unique geopolitical issues. Things that are not just continuations of current trends. For instance, 5 years ago as I write¹, few people imagined that Russia would annex Crimea or that Scotland would almost become independent, and 10 years ago even fewer would have predicted that, according to this week’s *Economist* (explicitly) and to most people I see in everyday life (implicitly), the defining artifact of our age is the smartphone.

It may be tempting to just throw up one’s hands and say such matters are completely unpredictable, but that is hardly helpful. Where is the line between “predictable” and “completely unpredictable”, and what do these concepts actually mean? The latter will be our first topic. Because there is no magic formula that will give reliable predictions for unique events, this lecture’s focus is on how to assess how good other people’s past predictions

¹March 2015

have been, and in particular we will look at

- The Good Judgment Project
- The annual World Economic Forum Global Risks Survey.

6.1 Some conceptual issues.

Over the last 20 years I have often read assertions of the form²

nobody predicted the peaceful ending of Soviet control of Eastern Europe (1989) and the subsequent breakup of the Soviet Union (1991).

But what exactly does that mean? A scholarly analysis of literature in the International Relations discipline was given in a 1993 paper by Gaddis. What's relevant to this lecture is his underlying premise

for a theory of International Relations to be regarded as successful, it should have been able to predict (in say 1985) that the end of the Cold War (before say 1995) was *likely* (page 18, edited).

Such an “unlikely events don't happen” attitude strikes me as very strange. To me it's self-evident that, in most cases of future uncertainty, instead of saying “this will or will not happen” one should think of alternative outcomes and assign probabilities. I happen to have a 1985 book (Dunnigan – Bay *A Quick and Dirty Guide to War, 1st edition*) which actually does this (list alternative outcomes and assign probabilities) for 15 potential future conflicts in different parts of the world. On the topic of the Cold War in Europe, their assessed probabilities for 1985-1995 were

65% status quo
25% internal revolts in Eastern Europe lead to decrease in Soviet control
5% military attack by Soviet Union on West Germany
5% Soviet Union falls apart for internal reasons

²Tracking down such assertions illustrates the difficulties of searching for pre-internet material. A quick search finds the Wikipedia page [Predictions of the dissolution of the Soviet Union](#) but no opposite page! And those predictions were of the style “it's a bad system that can't last forever” rather than any testable prediction.

and their phrase “the empire crumbles” for that final alternative proved rather accurate. Surely anyone else who seriously considered possibilities in 1985 would also assign some small probability to “the empire crumbles”. Living through and subsequently reading the actual history of this time period, my view (unprovable, of course) is that the outcome we saw really was *a priori* unlikely. Whenever we speculate about the future, we need to remember that unlikely events do sometimes happen!

Terminology: predict vs forecast. It is important to distinguish between between the two activities

- asserting “X will happen”
- listing some alternative possible future events and assigning probabilities to these alternatives.

English language is unhelpful, because the word *predict* invariably carries the former meaning, and *forecast* usually does: there is no standard word or phrase for the latter activity. Because the phrase *weather forecast* has some probability associations (“the chance of rain tomorrow is”), in this lecture I will use the word *forecast* for the second activity, in contrast to *predict* for the first activity. Thus “prediction markets” (the subject of Lecture 4) provide consensus *forecasts*, that is probabilities of the future event in question.

To me it seems self-evident that, in thinking about future uncertainties, one should think in terms of forecasts rather than predictions. My speculations on why this is not done more widely are deferred to Chapter ??.

6.2 The annual WEF Global Risks assessment

Where can we find some recent past forecasts, in order to try to assess their accuracy in retrospect? Here is the most interesting source that I know. At the end of each January you may see news reports involving the **World Economic Forum** (WEF) annual meeting in Davos. The WEF meeting itself is often criticized from various ideological perspectives, a debate I have no wish to enter. But somewhat paradoxically, by having no official status it is beholden to no-one and therefore is able to solicit background briefings written by a more extensive international and interdisciplinary group of experts than would be obtained by any governmental or academic sponsor. In

particular, each year since 2006 it has produced a “global risks” report³. A summary of each report is provided by a graphic indicating perceived likelihood and seriousness of about 30 major risks which might affect substantial parts of the world. Figures 6.1 - 6.3 show these graphics for the years 2007, 2011 and 2014 in similar (but with intriguing detailed differences – see point 4 below) format: the horizontal axis shows relative probabilities and the vertical scale shows relative economic effects.

Some details regarding these graphics. 1. Each phrase on the graphic is expanded to a sentence in the report. For instance “asset price collapse” is expanded to

A collapse of real and financial asset prices leading to the destruction of wealth, deleveraging, reduced household spending and impaired aggregate demand.

2. The reports are presumably written over the quarter before the January meeting, so the “2007” report would be based on knowledge and opinions from Fall 2006.

3. The future time period involved is not specified very consistently, but is apparently 5-10 years.

4. The labeling on the axes has changed over the years, as follows.

- In 2007 there are numerical probabilities (called “likelihood”) and numerical dollar economic effects.
- In 2010 the former is called “perceived likelihood” on a qualitative (unlikely to very likely) scale; the effects are still in dollars and called “perceived impact”.
- In 2014 respondents were asked to assess both likelihood and impact on a 1-7 scale with only verbal descriptions (very unlikely to very likely; low impact to high impact) of their meaning.

This backoff from quantitative forecasts is puzzling to me – perhaps it was intended to emphasize that the forecasts are based on subjective opinions of surveyed respondents, or perhaps as part of a legal disclaimer of liability for errors?

³Current and recent ones should be [available via this site](#).

Figure 6.1: Global Risks Perception 2007

The 23 Core Global Risks: Likelihood with **Severity by Economic Loss**

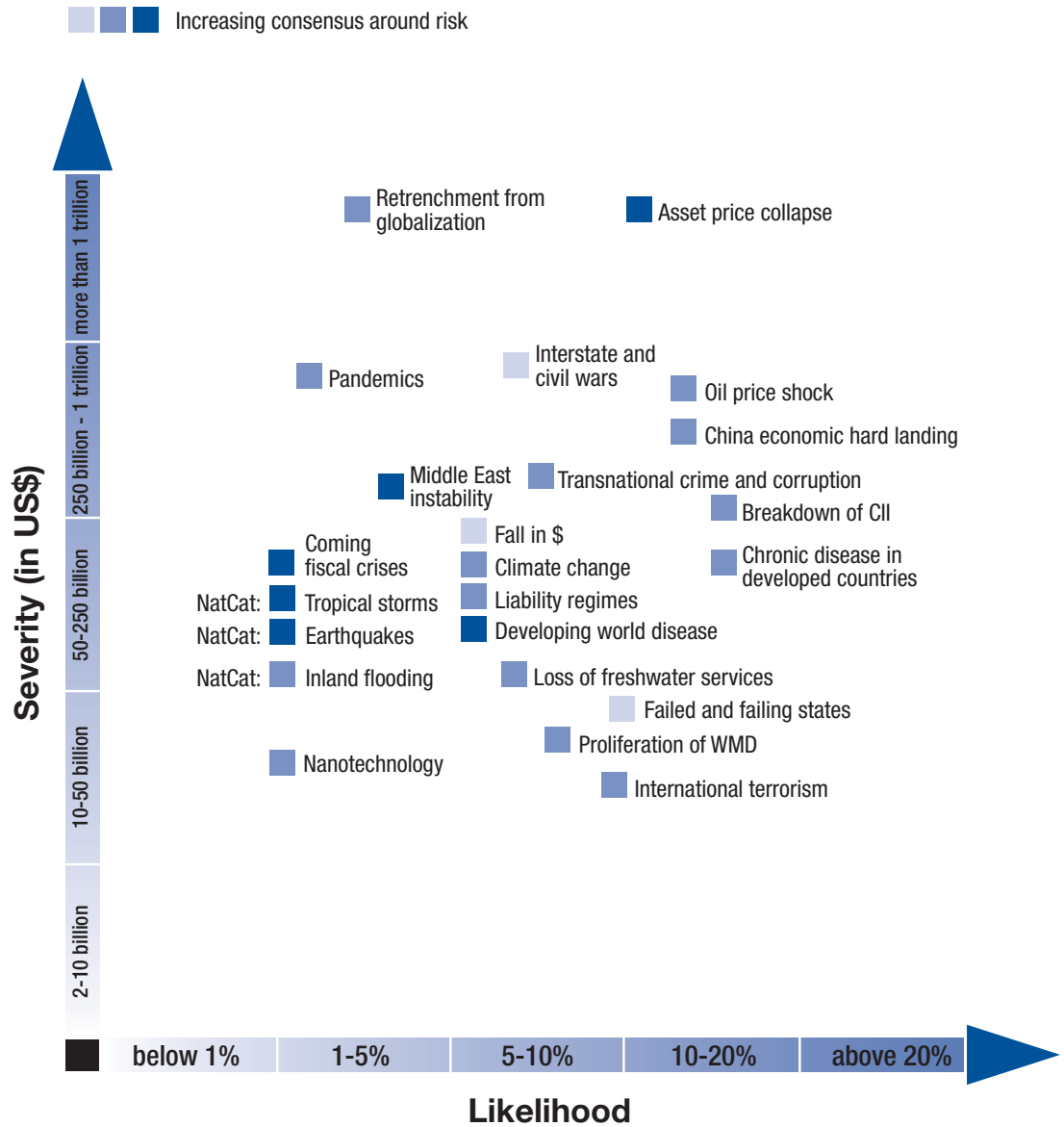


Figure 6.2: Global Risks Perception 2011

Figure 1 | Global Risks Landscape 2011:
Perception data from the World Economic Forum's Global Risks Survey

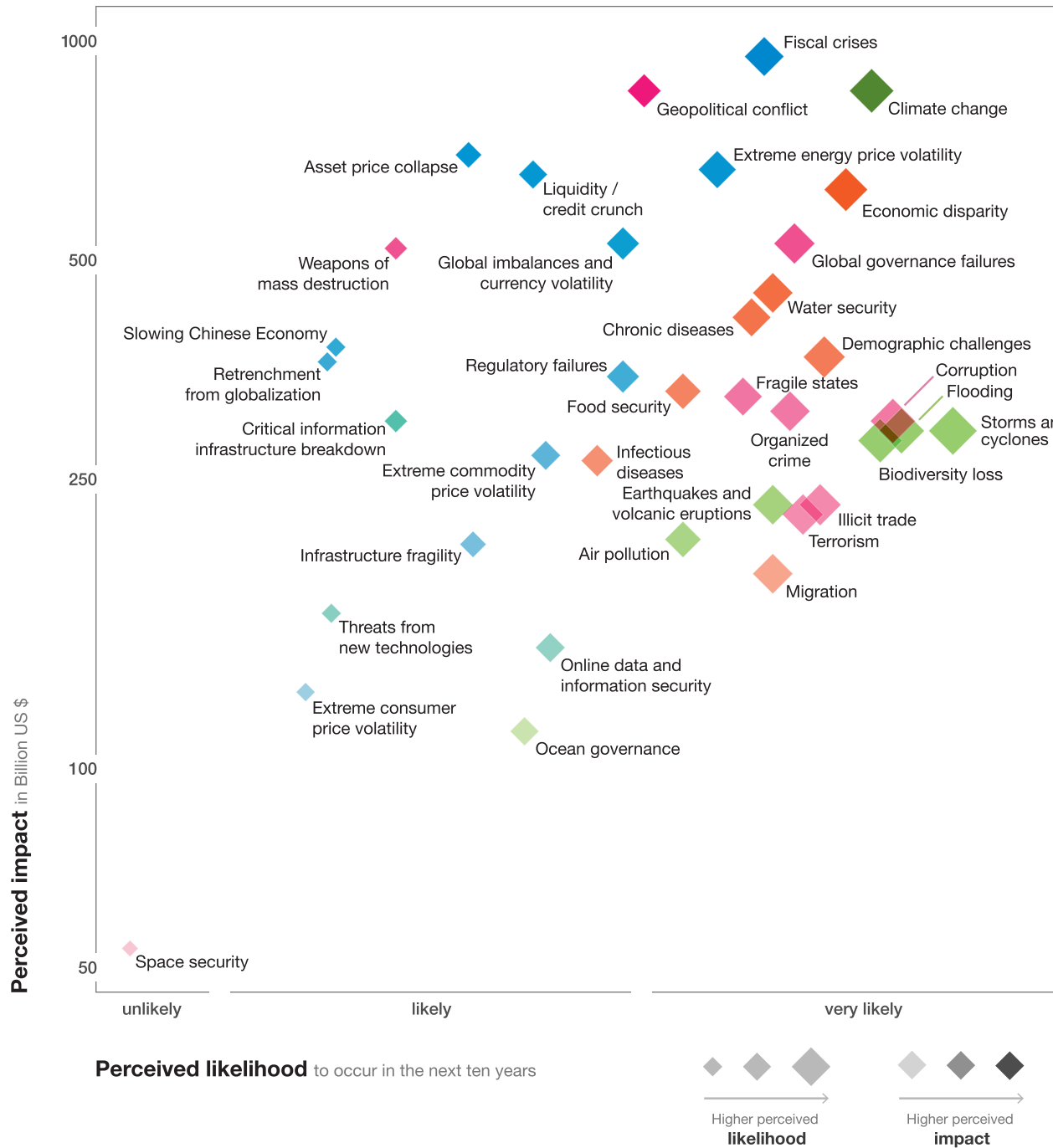
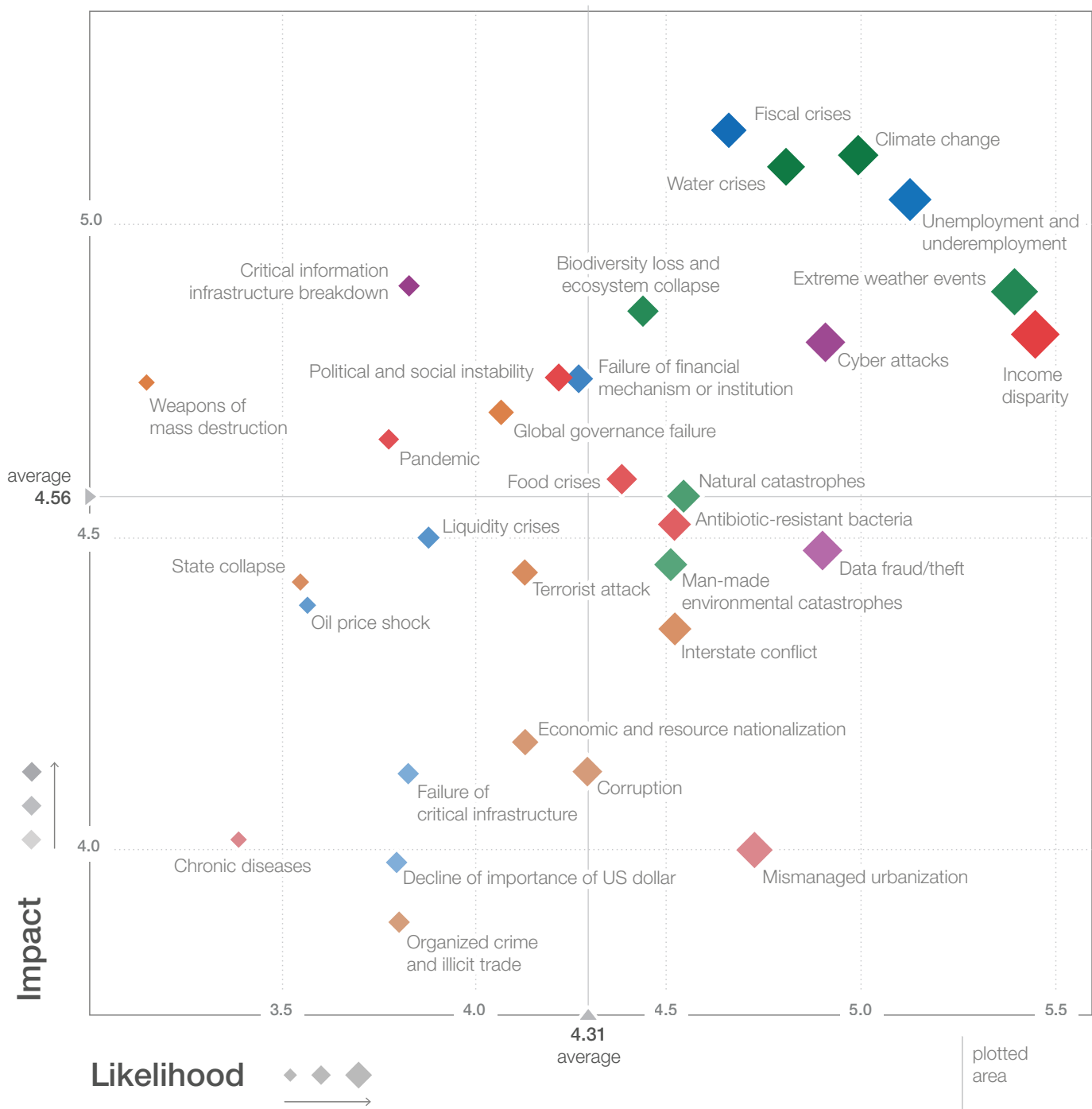


Figure 6.3: Global Risks Perception 2014

Figure 1.1: The Global Risks Landscape 2014



Part 1
Part 2
Part 3

7.0

In class I show the graphic from about 3 years ago and talk briefly about which of those risks have subsequently manifested. Then I ask the students what risks do they think are most prominent in the latest report, and then I show the latest year graphic. Below are a few such remarks.

Over the period I have given this lecture, the natural major “nobody foresaw . . .” event is surely the **Financial crisis of 2007-08**.⁴ A cynical view of retrospective analysis of this event is that commentators say either “no-one saw it coming” or “I saw it coming”, depending on whether they can exhibit evidence of the latter! Is such cynicism justified? The 2007 WEF Global Risks report was compiled in late 2006, at which time there were concerns about the worldwide boom in house prices, and some concerns about U.S. subprime mortgages, but nothing dramatic had happened in other markets. In that report we see (Figure 6.1) that the most serious risk (in a rough likelihood times severity sense) was “asset price collapse”, defined earlier, and this is almost precisely what subsequently happened. So although the severity of the subsequent **Great Recession** was undoubtedly underestimated, the oft-repeated “nobody saw it coming” view of the financial crisis is just plain wrong, in that it was widely viewed as a substantial risk.

Looking at the 2014 report and comparing with subsequent events during 2014, the two major relevant events are surely

- (i) **the annexation of Crimea by Russia** together with the subsequent **Russian military intervention in Ukraine**
- (ii) the emergence of **ISIL** as a military force controlling territory in Iraq and Syria.

These would fall into the risk categories *interstate conflict* and *state collapse*. Such risks were assessed (January 2014) as having low and medium (respectively) likelihood and medium impact. Unsurprisingly, the January 2015 report has reassessed *interstate conflict* as essentially the most serious of all the risks, while *state collapse* is assigned increased likelihood but not increased impact.

Is there a bottom line? While it is interesting to continue discussing particular examples – did this particular risk eventuate? – such political/economic history is not the subject of these lectures. The examples above underscore the difficulty in placing events on some predictable–unpredictable spectrum; one can recognize the general possibility of *state collapse* or *interstate conflict* without specifying particular states. Also recall that the WEF

⁴Remember this is ancient history to my current students.

reports are addressing total likelihood and impact over the next 5-10 years, rather than trying to predict specific events over the coming 12 months. For many reasons it is hard to make any quantitative assessment of the accuracy of these past risk assessments, so we will turn to a setting designed to allow such quantitative assessment.

A final philosophical point is that history pays much more attention to events that did happen than to events that didn't happen, but to make any inference about causality – does action *A* typically cause a war? – you need to know how often *A* happened but a war did *not* result. These reports provide a rare instance where one can check for events that didn't happen as easily as those that did.

6.3 The Good Judgment Project

In the [Good Judgment Project](#)⁵, non-expert individuals in teams are asked to assess the probabilities of selected future events best described as “near-term geopolitics and economics” with explicit deadlines. For instance, four of the 64 questions open as I write⁶ are

- Will there be a lethal confrontation between Chinese and Indian national military forces before 1 June 2015?
- Before 1 June 2015, will SWIFT restrict any Russian banks from accessing its services?
- Will negotiations on the Transatlantic Trade and Investment Partnership (TTIP) be completed before 10 June 2015?
- Will a unity government be formed in Libya before 1 June 2015?

More details about the Project will be provided as needed. To the reader who feels it is ridiculous to pose such questions to non-experts we would reply: do you feel that trial by jury is ridiculous? In both cases the point is to listen to evidence and to expert opinion and then deliberate before giving an answer.

We emphasize that contestants are **not** asked to give a Yes/No prediction, but instead are asked to give a numerical probability. We also emphasize that we are not going to propose any mathematical way for deciding on such probabilities. Instead, in this section we focus on how to measure, after

⁵Currently scheduled to end in mid-2015. I was an ordinary participant in 2014-5.

⁶March 2015

outcomes are known, the relative or absolute accuracy of such probability assessments made by others. Mathematically inclined readers will find the following discussion rather obvious and may skip to the next section. My point is to emphasize the distinction between *predicting* and *forecasting*, as defined in section 6.1.

Consider for a moment a scenario where two people, A and B, are asked to *predict* the outcome of each of 100 events. Eventually we know all the actual outcomes – suppose A gets 80 correct, and B gets 70 correct. There is no great subtlety in interpreting this data; either A is genuinely better than B at predicting the kind of events under study, or one person was unusually lucky or unlucky. In this lecture we consider the other scenario, where A and B are asked to give a *forecast* probability for each event. Now our data is of the form

event	A's forecast	B's forecast	occurs?
...
63	0.7	0.8	yes
64	0.5	0.6	no
...

Here it is less obvious what to do with this data – which person is better at assessing probabilities, and how good are they in absolute terms? To analyze such data, a basic method is to assign a score to each forecast, given by a formula involving the assessed probability p and the actual outcome. A mathematically natural choice of formula is

$$\begin{aligned} \text{score} &= (1 - p)^2 \text{ if event occurs} \\ &= p^2 \text{ if not.} \end{aligned} \tag{6.1}$$

As in golf, you are trying to get a low score. For instance if you forecast $p = 0.8$ then your score will be 0.04 if the event occurs but will be 0.64 if it does not occur.

This particular scoring formula has two nice features. Suppose you actually believe the probability is q . What p should you announce as your forecast? Under your belief, your mean score (by the rules of elementary mathematical probability) equals $q(1 - p)^2 + (1 - q)p^2$ and a line of algebra shows this can be rewritten as

$$(p - q)^2 + q(1 - q). \tag{6.2}$$

Because you seek to minimize the score, you should announce $p = q$, your honest belief – with this scoring rule you cannot “game the system” by being dishonest in that way.

Now write q for the true probability of the event occurring (recall we are dealing with future real-world events for which the true value q is unknown), and write p for the probability that you forecast. Then your (true) mean score, by exactly the same calculation, is also given by (6.2). The term $(p - q)^2$ is the “squared error” in your assessment of the probability. When contestants A and B forecasts the same event as probabilities p_A and p_B , (6.2) implies that the mean difference between their scores equals the difference between their squared errors. When A and B assess probabilities of the same long sequence of events, we can calculate their average (over events) scores s_A and s_B . We cannot know the corresponding mean-squared-errors $\text{MSE}(A)$ and $\text{MSE}(B)$, defined as the average (over events) of the squared errors $(p_A - q)^2$ and $(p_B - q)^2$, because we do not know the true probabilities q . But (6.2) implies that

$$s_A - s_B \text{ is a sample estimate of } \text{MSE}(A) - \text{MSE}(B) \quad (6.3)$$

in the law of large numbers sense, that as the number of events gets larger and larger, the difference between $s_A - s_B$ and $\text{MSE}(A) - \text{MSE}(B)$ gets smaller and smaller. In the golf analogy, 4 fewer strokes on one round is not convincing evidence that one player is better than another, but an average of 4 fewer over many rounds is.

The setting discussed above is called a *prediction tournament*⁷, and the Good Judgment Project is an example of a prediction tournament. The bottom line of the discussion above is

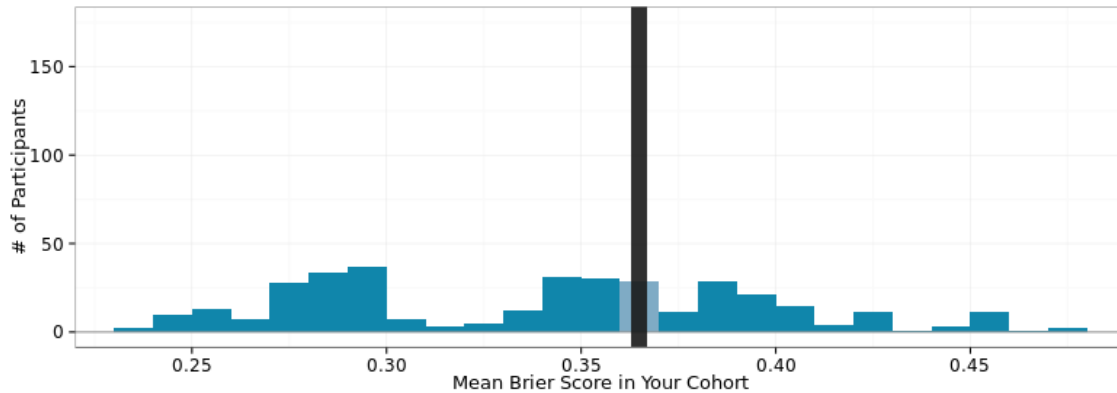
in a prediction tournament, the *differences* in scores indicate *differences* in forecasting skill, but one cannot assess the quality of forecasts in absolute terms.

In the golf analogy, it’s like having no given “par” score for a course.

Scores in the Good Judgment Project. Figure 6.4 shows a histogram of scores of individuals in the 2013-14 season. Interpretation of the numerical values is rather infeasible, for reasons explained later. The season scores were based on 144 questions, and a back-of-an-envelope calculation gives the SE due to intrinsic randomness of outcomes as around 0.02, which is much smaller than the spread observed in the histogram. The key conclusion is that there is wide variability between players – as in golf, some people are just much better than others at forecasting these geopolitical events.

⁷This name is unfortunately inconsistent with my usage of *prediction* and *forecast*.

Figure 6.4: Some scores in the Good Judgment Project (ignore the black bar).



6.4 More about the GJP

The Good Judgment Project (GJP) has roots in Philip Tetlock’s work, described in his 2005 book *Expert Political Judgment*, which may be best known for its conclusion that the “expert” forecasters he studied were often hard-pressed to do better than the proverbial dart-throwing chimp. Tetlock and colleagues believe that forecasting tournaments are the best way to compare forecasting ability; and that participants can improve their forecasting skills through a combination of training and practice, with frequent feedback on their accuracy. Combining training and practice with what GJP’s research suggests is a stable trait of forecasting skill seems to produce the phenomenon that GJP calls “superforecasters”. These have been so accurate that, according to a recent report by Washington Post columnist David Ignatius, they even outperformed the forecasts of intelligence analysts who have access to classified information. *Extracts from the (public) Project blog, with minor edits.*

The book is well worth reading, with undergraduate-level mathematical statistics arising from serious conceptual issues. Just for fun, I quote his categorization of excuses that experts make when their predictions turn out

wrong. I have changed some of his titles.

1: Implicit conditions not satisfied. For instance, you predict that implementing a certain policy will have good results; if not then you say the policy must have been implemented badly.

2: Exogeneous shocks. Nobody could have expected⁸.

3: Close call counterfactual. I was almost right.

4: Just off on timing. The war lasted a bit longer than the question deadline.

5: Politics is unpredictable, anyway. So my mistake wasn't really a mistake.

6: I made the right mistake. An error in the other direction would have been more serious.

7. Unlikely events sometimes happen.

Details of GJP scoring. A major aspect of the GJP is that, instead of making a single forecast on a question, participants update their forecasts over time as new information is acquired; the “score” on a question is then the average (over days) of the scores for each day’s forecast. Scoring actually uses **Brier score**, which for yes/no questions is twice the score **6.1** but which extends to the minority of questions with more than two alternatives. Teams consist of 8-15 people; with typically 50+ questions. Typically 3-4 people make forecasts on a given question; an individual who does not forecast is given the team median score. All these effects make it hard to interpret the numbers in Figure **6.4**.

Finally there are two technical issues caused by the participants (and organizers!) not paying attention to fine details of scoring. First, if no team member makes a forecast then a 0.5 probability is imputed; because at opening most event probabilities seem not close to 0.5 a team that delays making a first forecast (e.g. because of no hard news) is penalized. Second, if the event occurs before the deadline then the question is closed and the scores are averaged over days until closing. But this allows one (in theory) to “game the system”. If you believe an event will either occur quickly or not at all, then⁹ (if that belief is correct) it is advantageous to you to overstate the probability of the event at the start of the time window So some of the variability in scores might be due to such scoring artifacts rather than actual forecasting skill.

⁸ *the Spanish Inquisition*

⁹This is an exercise for the mathematically-inclined reader.

6.5 The cost of errors in assessing probabilities

Here we digress somewhat to a general issue, surprisingly not much discussed in either textbook or popular accounts. Suppose you believe an event has probability $q = 42\%$ whereas it really has probability $p = 57\%$. What is your error? Well, the *size* of your error could be measured in several ways, most simply by the difference $|p - q| = 15\%$. But what is the *cost* of such an error? This is a very vague question, and clearly the answer is very context-dependent. In our prediction tournament context we used cost $(p - q)^2$ as being mathematically convenient. Is there some justification other than mathematical convenience?

In this section I describe two contexts in which this cost function arises naturally, at least as a first-order approximation for small $|p - q|$. I am implicitly not considering “highly asymmetric” cases where probabilities are close to 0 or 1, or where the consequences of occurrence and non-occurrence are dramatically different.

The simplest decision problem. We consider a very simple model of a decision under uncertainty, which we could view as a bet against Nature, an opponent who is indifferent to our actions and wishes.

Model. An event F will occur with unknown probability p . You have a choice of action A, which you would take if you knew F would occur, or action B, which you would take if you knew F would not occur. So we suppose there is a payoff table

- (action A): payoff = a if F occurs, payoff = b if F does not occur
- (action B): payoff = c if F occurs, payoff = d if F does not occur

where $a > c$ and $d > b$. (If payoffs are random we can just take their expectations. We assume the classical setting of linear utility, not being risk-averse). Now we calculate the mean payoffs

- (action A): mean payoff = $pa + (1 - p)b$
- (action B): mean payoff = $pc + (1 - p)d$

There is a critical value p_{crit} where these mean payoffs are equal, and this is the solution of

$$\frac{p_{\text{crit}}}{1 - p_{\text{crit}}} = \frac{d - b}{a - c}.$$

If we knew p our best strategy is

do action A if $p > p_{\text{crit}}$, do action B if $p < p_{\text{crit}}$.

Instead all we have is our guess p_{guess} , so we use this strategy but based on p_{guess} instead of p .

What is the cost of not knowing p ? If p_{guess} and p are on the same side of p_{crit} then we take the optimal action and there is zero cost; if they are on opposite sides we take the sub-optimal action and the cost is

$$|p - p_{\text{crit}}|z \text{ where } z = a - b - c + d > 0. \quad (6.4)$$

Now consider what happens in many repeated different games of this type. Assume the different payoffs are all of order 1 and are independent (over games) of the probabilities, and hence p_{crit} is independent of p and p_{guess} . Then the proportion of times that p_{crit} happens to be in the interval between p and p_{guess} should be of order $|p - p_{\text{guess}}|$, assuming the latter is small; and when this occurs the mean cost is also, by (6.4), of order $|p - p_{\text{guess}}|$. So in this particular “decision under uncertainty” context the cost of errors is indeed of order $(p - p_{\text{guess}})^2$.

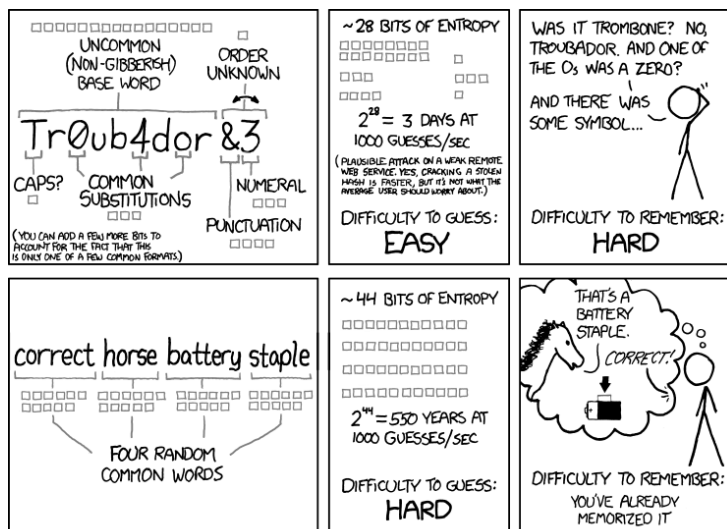
Gambling at favorable odds. Suppose someone offers you a bet (in either direction) at fair odds based on their belief the probability of the event is p , and suppose you know the probability is really q , where $|p - q|$ is small. You can now make a favorable bet, with profit of order $|p - q|$ per unit bet. And the Kelly criterion from section 2.6 tells you that, presented with repeated bets of this type, the proportion of your fortune to bet each time is also order $|p - q|$. So the rate of growth of your fortune is order $(p - q)^2$, representing the “cost” of your counterpart’s error.

Chapter 7

Coding and entropy

Note. At Berkeley, information theory is taught in a graduate course but not an undergraduate one, so I assume my students have not seen any of this material. The final section summary should be comprehensible even if all the math is skipped.

Figure 7.1: xkcd.com/936



THROUGH 20 YEARS OF EFFORT, WE'VE SUCCESSFULLY TRAINED EVERYONE TO USE PASSWORDS THAT ARE HARD FOR HUMANS TO REMEMBER, BUT EASY FOR COMPUTERS TO GUESS.

7.1 Introduction

In an earlier survey I asked students to write down a common five-letter English word. I start this lecture by showing the xkcd cartoon in Figure 7.1 and then demonstrate the cartoon’s essential truth by using a [password strength checker](#) to assess the strengths of

- a concatenation of 4 of the students’ words
- a volunteer student’s password.

Invariably the former is judged “strong” or “very strong” and the latter “weak” or “medium”.

This lecture introduces a few topics from a big field known misleadingly as *Information Theory* – see the “further reading” section 7.10. Each of the words *coding* and *entropy* have rather specific meanings in this lecture, so I first must explain these meanings.

7.2 Entropy as a measure of unpredictability

For a probability distribution over numbers – Binomial or Poisson, Normal or Exponential – the mean or standard distribution are examples of “statistics” – numbers that provide partial information about the distribution. Consider instead a probability distribution over an arbitrary finite set S . Simple concrete examples we have in mind for S are

- Relative frequencies of given names (Table 7.1)¹.
- Relative frequencies of letters in the English language (Figure 7.2)

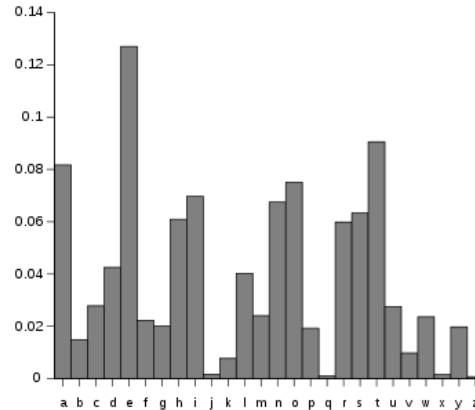
Table 7.1: 2013 U.S. births given names.

Rank	Male name	Percent of total males	Female name	Percent of total females
1	Noah	0.9043%	Sophia	1.1039%
2	Liam	0.8999%	Emma	1.0888%
3	Jacob	0.8986%	Olivia	0.9562%
4	Mason	0.8793%	Isabella	0.9161%
5	William	0.8246%	Ava	0.7924%
6	Ethan	0.8062%	Mia	0.6844%

- Relative frequencies of words in the English language.

¹The extensive such data from [the Social security site](#) is an interesting source for student projects.

Figure 7.2: Relative frequencies of letters in the English language (from Wikipedia)



(iv) Relative frequencies of phrases in the English language².

For a probability distribution $\mathbf{p} = (p_s, s \in S)$ on such sets S it does not make sense to talk about *mean* or *standard deviation*. But it does make sense to devise statistics that involve only the *unordered* set of values $\{p_s\}$, and the particular statistic relevant to this lecture is

$$\mathcal{E}(\mathbf{p}) := - \sum_s p_s \log p_s$$

which is called the *entropy* of the probability distribution \mathbf{p} . This terminology is confusing, partly because “entropy” is often used for what is properly called *entropy rate* (section 7.4), and partly because of the only indirectly related notion of *entropy* in statistical physics.

A basic fact is that the uniform distribution on an n -element set has entropy $= \log n$ whereas the “degenerate” distribution concentrated at a single element has entropy zero. The entropy statistic serves to place a distribution on the spectrum from degenerate to uniform; entropy is described as “amount of randomness” but for our purposes is better regarded as a measure of *unpredictability*. Note that many other statistics serve the same general purpose, as discussed further in Lecture xxx under the phrase *diversity statistic*.

²See the [Google Books Ngram Viewer](#), which has various interesting uses. To see usage of *data* as singular or plural, compare frequencies of “the data is” and “the data are”.

A good way to interpret the numerical value of $\mathcal{E}(\mathbf{p})$ is via the “effective number” N_{eff} – the number³ such that the uniform distribution on N_{eff} elements has the same statistic. See section xxx for an illustration concerning the changes over time in the diversity of given names (Table 7.1).

Entropy in physics. The reader has likely seen a statement of the **second law of thermodynamics** in a verbal form such as

the total entropy of any isolated thermodynamic system increases over time, approaching a maximum value

and the informal description of entropy as a measure of *disorder*. When expressed in mathematical terms one can indeed see connections between this physics formulation of *entropy* and our definition of $\mathcal{E}(\mathbf{p})$, but this connection is not particularly helpful for an introductory treatment of the topic of this lecture.

7.3 Coding, compression and encryption

The word *coding* nowadays primarily means “writing computer code” but here we are concerned with representing data in some convenient form. A simple example is the original ASCII scheme (section 7.6) for representing letters and typewriter symbols in binary. In choosing how to code a particular type of data there are several issues one might consider.

- *Compression*: coding to make a text shorter

is useful both in data storage and in data transmission, because there is some “cost” both to storage space and transmission time.

- *Encryption*: coding for secrecy

is familiar from old spy novels and from modern concerns about security of information sent over the internet. These differ in an obvious way. Compressing files on your computer will produce, say, a `.zip` file, and the algorithms for compressing and decompressing are public. Encryption algorithms in widespread use are commonly like **public-key cryptography** in that the logical form of the algorithms for encryption and decryption are public, but a private key (like a password) is required to actually perform decryption. In contrast, intelligence agencies presumably use algorithms whose

³The solution of $\mathcal{E}(\mathbf{p}) = \log N_{\text{eff}}$, typically not actually an integer.

form is secret. For concreteness, in this lecture I talk in terms of coding English language text, but the issues are the same for any kind of data.

A third issue I will not discuss is

- robustness under errors in data transmission: **error-correcting code**

Intuitively there seems no particular connection between encryption and compression – if anything, they seem opposites, involving secrecy and openness. But a consequence of the mathematical theory outlined in this lecture is that

(*) finding good codes for encryption is the same as finding good codes for compression.

Here is a verbal argument for (*). A code or cipher transforms *plaintext* into *ciphertext*. The simplest **substitution cipher** transforms each letter into another letter. Such codes – often featured as puzzles in magazines – are easy to break using the fact that different letters and letter-pairs occur in English (and other natural languages) with different frequencies. A more abstract viewpoint is that there are $26!$ possible “codebooks” but that, given a moderately long ciphertext, only one codebook corresponds to a meaningful plaintext message.

Now imagine a hypothetical language in which *every* string of letters like QHSKUUC ... had a meaning. In such a language, a substitution cipher would be unbreakable, because an adversary seeing the ciphertext would know only that it came from of $26!$ possible plaintexts, and if all these are meaningful then there would be no way to pick out the true plaintext. Even though the context of secrecy would give hints about the general nature of a message – say it has military significance, and only one in a million messages has military significance – that still leaves $10^{-6} \times 26!$ possible plaintexts.

Returning to English language plaintext, let us think about what makes a *compression* code good. It is intuitively clear that for an ideal coding we want each possible sequence of ciphertext to arise from some meaningful plaintext (otherwise we are wasting an opportunity); and it is also intuitively plausible that we want the possible ciphertexts to be approximately equally likely (this is the key issue that the mathematics deals with).

Suppose there are 2^{1000} possible messages, and we’re equally likely to want to communicate each of them. Then an ideal code would encode each as a different 1000-bit (binary digit) string, and this could be a public algorithm for encoding and decoding. Now consider a substitution code based on the 32 word “alphabet” of 5-bit strings. Then we could encrypt a message by

- (i) apply the public algorithm to get a 1000-bit string;
- (ii) then use the substitution code, separately on each 5-bit block.

An adversary would know we had used one of the $32!$ possible codebooks and hence know that the message was one of a certain set of $32!$ plaintext messages. But, by the “approximately equally likely” part of the ideal coding scheme, these would be approximately equally likely, and again the adversary has no practical way to pick out the true plaintext.

Conclusion: given a good public code for compression, one can easily convert it to a good code for encryption.

7.4 The asymptotic equipartition property

We now jump into math theory to state a non-elementary result, and accompany it with some discussion. The basis of the mathematical theory is that we model the source of plaintext as random “characters” X_1, X_2, X_3, \dots in some “alphabet”. It is important to note that we do *not* model them as independent (even though I use independence as the simplest case for mathematical calculation later) since real English plaintext obviously lacks independence. Instead we model the sequence (X_i) as a *stationary process*, which implies that there is some probability that three consecutive characters are CHE, but this probability does not depend on position in the sequence, and we don’t make any assumptions about what the probability is.

To say the setup more carefully, for any sequence of characters (x_1, \dots, x_n) there is a *likelihood*

$$\ell(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n).$$

The *stationarity* assumption is that for each time t and each sequence (x_1, \dots, x_n)

$$\mathbb{P}(X_{t+1} = x_1, \dots, X_{t+n} = x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n). \quad (7.1)$$

Consider the *empirical likelihood*

$$L_n = \ell(X_1, \dots, X_n)$$

which is the prior chance of seeing the sequence that actually turned up. The central result (non-elementary; I teach it in a graduate course as the Shannon-McMillan-Breiman theorem) is

The asymptotic equipartition property (AEP) . For a stationary ergodic⁴ source, there is a number $\mathcal{E}nt$, called the *entropy rate* of the source, such that for large n , with high probability

$$-\log_2 L_n \approx n \times \mathcal{E}nt.$$

The rest of this section is the mathematical discussion of the theorem that I say in class. I'm not going to attempt to translate it for the general reader, who should skip to the next section to see the relevance to coding. It is conventional to use base 2 logarithms in this context, to fit nicely with the idea of coding into bits.

For n tosses of a hypothetical biased coin with $\mathbb{P}(H) = 2/3, \mathbb{P}(T) = 1/3$, the *most likely* sequence is $HHHHHH \dots HHH$, which has likelihood $(2/3)^n$, but a *typical* sequence will have about $2n/3$ H's and about $n/3$ T's, and such a sequence has likelihood $\approx (2/3)^{2n/3}(1/3)^{n/3}$. So

$$\log_2 L_n \approx n\left(\frac{2}{3} \log_2 \frac{2}{3} + \frac{1}{3} \log_2 \frac{1}{3}\right).$$

Note in particular that log-likelihood behaves differently from the behavior of sums, where the CLT implies that a “typical value” of a sum is close to the most likely individual value.

Recall that the *entropy* of a probability distribution $\mathbf{q} = (q_j)$ is defined as the number

$$\mathcal{E}(\mathbf{q}) = -\sum_j q_j \log_2 q_j. \quad (7.2)$$

The AEP provides one of the nicer motivations for the definition, as follows. If the sequence (X_i) is IID with marginal distribution (p_a) then for $\mathbf{x} = (x_1, \dots, x_n)$ we have

$$\ell(\mathbf{x}) = \prod_a p_a^{n_a(\mathbf{x})}$$

where $n_a(\mathbf{x})$ is the number of appearances of a in \mathbf{x} . Because $n_a(X_1, \dots, X_n) \approx np_a$ we find

$$L_n \approx \prod_a p_a^{np_a}$$

$$-\log_2 L_n \approx n \left(-\sum_a p_a \log_2 p_a \right).$$

⁴The formal **definition of ergodic** is hard to understand; basically we exclude a source that flips a coin to choose between “all English” and “all Russian”.

So the AEP identifies the *entropy rate* of the IID sequence with the *entropy* $\mathcal{E} = -\sum_a p_a \log_2 p_a$ of the marginal distributions X .

Let me mention three technical facts.

Fact 1. (easy). For a 1-1 function C (that is, a code that can be decoded precisely), the distributions of a random item X and the coded item $C(X)$ have equal entropy.

Fact 2. (easy). Amongst probability distributions on an alphabet of size B , entropy is maximized by the uniform distribution, whose entropy is $\log_2 B$. So for any distribution on binary strings of length m , the entropy is at most $\log_2 2^m = m$.

Fact 3. (less easy). Think of a string (X_1, \dots, X_n) as a single random object. It has some entropy \mathcal{E}_k . In the setting of the AEP,

$$k^{-1} \mathcal{E}_k \rightarrow \mathcal{E}nt \text{ as } k \rightarrow \infty.$$

Finally a conceptual comment. Identifying the entropy rate of an IID sequence with the entropy of its marginal distribution indicates that *entropy* is the relevant summary statistic for the non-uniformness of a distribution when we are in some kind of *multiplicative* context. This is loosely analogous to the topic of Lecture 2, the Kelly criterion, which is tied to “multiplicative” investment.

7.5 Entropy rate and minimum code length

Here we will outline in words the statement and proof of the fundamental result in the whole field. The case of an IID source (recall section 2.2) is **Shannon’s source coding theorem** from 1948. The “approximation” is as $n \rightarrow \infty$.

A string of length n from a source with entropy rate $\mathcal{E}nt$ can be coded as a binary string of length $\approx n \times \mathcal{E}nt$ but not of shorter length.

More briefly, the optimal coding rate is $\mathcal{E}nt$ bits per letter.

Why not shorter? Think of the entire message (X_1, \dots, X_n) as a single random object. The AEP says the entropy of its distribution is approximately $n \times \mathcal{E}nt$. Suppose we can code it as a binary string (Y_1, \dots, Y_m) of some length m . By Fact 1, the entropy of the distribution of (Y_1, \dots, Y_m) also $\approx n \times \mathcal{E}nt$, whereas by Fact 2 the entropy is at most m . Thus m is approximately $\geq n \times \mathcal{E}nt$ as asserted.

How to code this short. We give an easy to describe but completely impractical scheme. Saying that a typical plaintext string has chance about 1 in a million implies there must be around 1 million such strings (if more then the total probability would be > 1 ; if less then with some non-negligible chance a string has likelihood not near 1 in a million). So the AEP implies that a typical length- n string is one of the set of about $2^{n \times \mathcal{E}nt}$ strings which have likelihood about $2^{-n \times \mathcal{E}nt}$ (and this is the origin of the phrase *asymptotic equipartition property*). So in principle we could devise a codebook which first lists all these strings as integers $1, 2, \dots, 2^{n \times \mathcal{E}nt}$, and then the compressed message is just the binary expansion of this integer, whose length is $\log_2 2^{n \times \mathcal{E}nt} = n \times \mathcal{E}nt$. So a typical message can be compressed to length about $n \times \mathcal{E}nt$; atypical messages (which could be coded in some non-efficient way) don't affect the limit assertion.

The second argument is really exploiting a loophole in the statement. Viewing the procedure as transmission, we imagine that transmitter and receiver are using some codebook, but we placed no restriction on the size of the codebook, and the code described above uses a ridiculously large and impractical codebook,

The classical way to get more practical codes is by fixing some small k and coding blocks of length k , Thus requires a codebook of size A^k , where A is the underlying alphabet size. However, making an optimal codebook of this type requires knowing the frequencies of blocks that will be produced by the source. Rather than explain further, we shall jump (after a brief historical digression) to more modern codes that don't assume such knowledge..

7.6 Morse code and ASCII

Invented around 1840, **Morse code** codes each letter and numeral as a sequence of dots and dashes: for instance T is $-$ and Z is $--\bullet$. Logically this is like coding into a *three*-letter alphabet, because one also needs to indicate (by a pause) the spaces between letters. As is intuitively natural, common letters (like T) are coded as short sequences and uncommon letters (like Z) are coded as longer sequences. Given frequencies of letters, there is a theoretical optimal way (**Huffman coding**) to implement such a *variable length* code, and this has the same intuitive feature. But it's important to note that Huffman coding is optimal only amongst codes applied to individual letters, and depends on known fixed frequencies for letters.

Developed in the 1960s, **ASCII** codes letters, numerals and other symbols

into 128 7-bit strings: for instance T is 101 0100 and Z is 101 1010. At first sight it may seem surprising that ASCII, and its current extension [unicode](#), don't use variable length codes as did Morse code. But the modern idea is that with any kind of original data one can first digitize into binary in some simple way, and then compress later if needed.

7.7 Lempel-Ziv algorithms

In the 1970s it was realized that with computing power you don't need a fixed codebook at all – there are schemes that are (asymptotically) optimal for any source. Such schemes are known as Lempel-Ziv style⁵ algorithms, though the specific version described below, chosen as easy to describe, is not the textbook form.

Suppose we want to transmit the message

010110111010|011001000

and that we have transmitted the part up to |, and this has been decoded by the receiver. We will next code some initial segment of the subsequent text 011001000 To do this, first find the longest initial segment that has appeared in the already-transmitted text. In this example it is 0110 which appeared in the position shown.

010110111010|011001000

Writing n for the position of the current (first not transmitted) bit, let $n - k$ be the position of the start of the closest previous appearance of this segment, and ℓ for the length of the segment. In the example, $(k, \ell) = (10, 4)$. We transmit the pair (k, ℓ) ; the receiver knows where to look to find the desired segment and append it to the previously decoded text. Now we just repeat the procedure:

0101101110100110|01000

the next maximal segment is 0100 and we transmit this as $(7, 4)$.

How efficient is this scheme? We argue informally as follows. When we're a long way into the text – position n say – we will be transmitting segments of some typical length $\ell = \ell(n)$ which grows with n (in fact it grows as order $\log n$ but that isn't needed for this argument). By the AEP the likelihood

⁵The current [Wikipedia article](#) is not so helpful for the general reader.

of a particular typical such segment is about $2^{-\ell \times \mathcal{E}nt}$ and so the distance k we need to look back to find the same segment is order $2^{+\ell \times \mathcal{E}nt}$. So to transmit the pair (k, ℓ) we need $\log_2 \ell + \log_2 k \approx \ell \times \mathcal{E}nt$ bits. Because this is transmitting ℓ letters of the text, we are transmitting at rate $\mathcal{E}nt$ bits per letter, which is the optimal rate.

7.8 Checking for yourself

On my Mac I can use the Unix `compress` command, which implements one version of the Lempel-Ziv algorithm. A simple theoretical prediction is that if you take a long piece of text, split it into two halves of equal uncompressed length, and compress each half separately, then the two compressed halves will be approximately the same length. It takes only a few minutes to check an example. I used a text of *Don Quixote*, in English translation, downloaded from Project Gutenberg.

Table 7.2: Bytes in Don Quixote

	uncompressed	compressed
first half	1109963	444456
second half	1109901	451336
whole	2219864	895223

The prediction works pretty well. Further predictions can be made based on the notion that the algorithm incurs some “start-up cost” before the coding becomes efficient, implying

- The compressed size of a complete text will be shorter than the sums of compressed sizes of its parts. (We see this in the example above, though the difference is very small).
- For a text broken into pieces of different sizes, the compression ratio for the pieces will be roughly constant but also will tend to decrease slightly as size increases.

To illustrate the latter, I used the L^AT_EX text of the Grinsted-Snell textbook *Introduction to Probability*.

Table 7.3: Bytes in Grinsted-Snell

chapter	uncompressed	compressed	ratio
1	101082	46029	.465
2	73966	32130	.434
3	139490	61571	.441
4	123784	53962	.436
5	100155	43076	.430
6	134256	57577	.429
7	39975	18021	.451
8	39955	18759	.470
9	90019	39853	.443
10	79560	35058	.441
11	166626	69181	.415
12	56463	25299	.448

7.9 ...but English text is not random

So one could just demonstrate that compression algorithms work in practice on natural English text, and stop. But this doesn't address a conceptual issue.

(B) If you designed a vehicle to work well as an airplane, you wouldn't expect it to work well as a submarine. So why do algorithms, designed to work well on random data, in fact work well in the completely opposite realm of meaningful English language?

A standard explanation goes as follows. Do we expect that the frequency of any common word (e.g. "the") in the second half of a book should be about the same as in the first half? Such "stabilization of frequencies" seems plausible – we are not looking at meaning, just syntax, which doesn't change through the book. This idea of "the rules are not changing" suggests the analogy between written text and a deterministic physical system. An iconic mental picture of the latter is "frictionless billiard balls" which, once set in motion, continue bouncing off each other and the table sides forever. For certain kinds of such physical systems, *ergodic theory* predicts "stabilization of frequencies" – e.g. the proportion of time a ball spends near a corner should be about the same in the first hour as in the second hour. One can

introduce randomness into the story by taking, for the physical system, a random time as “time 0”, or a random page as “page 0” in a text, and then counting time relative to this start. And the notion of “stabilization of frequencies” turns out to be mathematically equivalent to saying that by a special choice of a random initial state (e.g. what we would see at a time chosen at random from a very long time interval) one sees a stationary random process in the sense (7.1). Granted this as a model for English text, we get both “stabilization of frequencies” and the theory for coding that we described earlier, as mathematical consequences.

What is unsatisfactory about that explanation? Well, we are asked to accept, in this particular setting of writing text, the analogy between conscious decisions and a physical system. But it is hard to think of another setting where conscious decisions of a single individual can reasonably be modeled probabilistically, so it begs the question of what is so special about writing text.

7.10 Wrap-up and further reading

For the topic of this lecture

- There is extensive mathematical theory, and algorithms based on the theory are used widely.
- Some consequences of theory are readily checkable.
- The use of probability is conceptually subtle. We don’t think of speech or writing as random in everyday life, not does it fit naturally into neat philosophical categories like “intrinsic randomness” or “opinion/lack of knowledge randomness”.
- But there is no explanation of *why* algorithms work except via a model of randomness.

In Lecture 4 we saw a context (prediction markets and strategies for fair games) where one can make numerical predictions without needing a very specific model but only assuming a structural property (martingale). This lecture shows the same for the context of data compression, the structural property being stationarity. A third such context is spatial networks (section 9.4 later), the structural property being scale-invariance.

Further reading. This lecture’s topic grew from a 1948 Shannon paper with the title “a mathematical theory of communication” and the broad academic field subsequently acquired the name *Information Theory*. This is an unfortunate name – the thought-provoking book *Information: A Very Short Introduction* by Floridi gives one view of how this field fits into the much bigger picture of what “information” really is. The Wikipedia article [Information Theory](#) outlines the scope of this academic field, and the Cover - Thomas textbook *Elements of information theory* is a standard first mathematical treatment.

Chapter 8

From physical randomness to the local uniformity principle

The presumption that observed quantitative data should typically follow some smooth distribution, absent some specific reason otherwise, is widely used but rarely discussed, and indeed has no standard name. I will call it the local uniformity principle, and illustrate contexts where it is implicitly used.

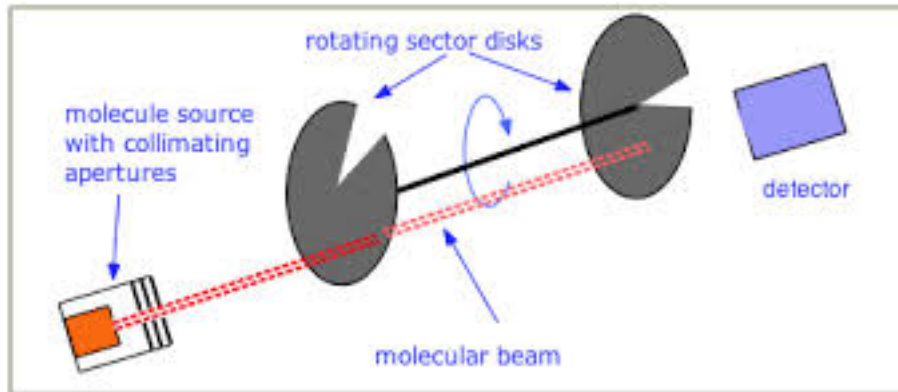
8.1 A glance at physical randomness

Although it is more mathematical than is appropriate for these lectures, I cannot resist advertising a little known book, *The Physics of Chance* by Charles Ruhla, which I once reviewed as follows.

It takes a selection of standard topics but treats them in a serious, careful and well written way, via a "horizontal integration" of math theory, its meaning within physics and its experimental verification. Topics include measurement error, the Maxwell velocity distribution for an ideal gas, Boltzmann's statistical physics, deterministic chaos illustrated by a compass needle undergoing forced oscillations, a detailed account of the quantum theory of interference and an "inseparable photons" experiment.

One of my desiderata for a lecture topic was that *the mathematics leads to some theoretical quantitative prediction that my students can test by gathering fresh data*. So because I don't expect them to be able to do physics experiments, the uses of probability within physics are downplayed in these

Figure 8.1: Experimental equipment to check the Maxwell velocity distribution.



lectures. In particular, in talking about the course I say that the three-dimensional-Gaussian velocity distribution for an ideal gas predicted by Maxwell is a paradigm example of an established uncontroversial theory. but one that my students cannot check. So let me digress to show how experimental physicists have actually verified the prediction, via an oversimplified explanation illustrated in Figure 8.1. The apparatus is within an evacuated space. There is a detector which can measure how many molecules hit it¹, and a container of gas molecules with an aperture aligned toward the detector. In between are two spinning discs with small sectors cut out. For given rotation speeds of the discs, there is essentially only one small interval of velocities at which molecules (moving in a straight line at constant speed because they are in a near-vacuum) will reach the detector, and therefore by varying disc rotation speeds one can measure the relative frequencies of different speeds of molecules.

¹In relative terms, not a count of molecules.

8.2 Dart throws as a simple example of physical randomness

I use “throwing a dart at a target” as a paradigm example of one kind of “physical” randomness. Figure 8.2 shows the result of 100 throws² at a dartboard – the kind of experiment you could easily repeat yourself. To show scale, we draw an imaginary playing card centered at the center of the board. Textbooks are apt (cf. section 1.4) to give examples of the kind “suppose the throws land uniformly (or as bivariate Normal)” but neither supposition is remotely accurate.

The following comments are rather trite in the context of darts, but provide starting points for discussion within other contexts, as I will try to show in the rest of the lecture.

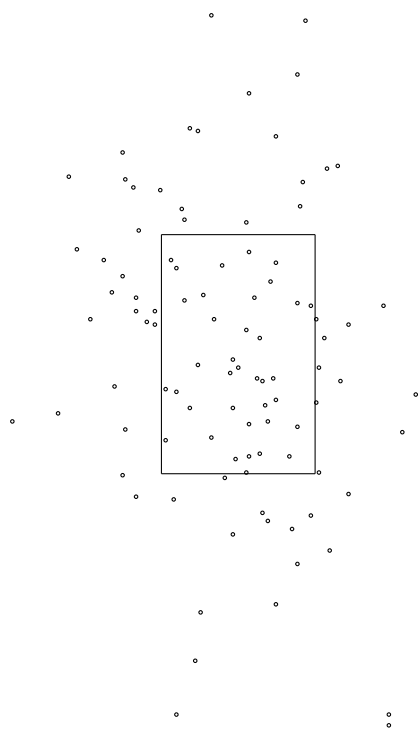
Darts are not dice. As the next point says, the result of dart throws are a combination of skill and chance, rather than the “pure chance” of a die roll, and so the chances vary between individual throwers, unlike dice. And the “physical symmetry” of a die or analogous gaming artifact allows one to assert that different outcomes are equally likely, which has no counterpart for dart throws.

Darts provide a vivid and quantifiable instance of the luck-skill combination. Whether one can quantify the relative contributions of skill and luck to observed success, in some particular field of human endeavor or for some individual person, is one of the most intriguing aspects of probability in the real world³. For darts one can quantify skill as, for instance, mean-square deviation from target point, and presumably there is strong correlation between that statistic and success at traditional games based on dart throws. In many professional team sports there are statistics for individual player (e.g. pitcher or quarterback) performance, which again are presumably correlated with the player’s contribution to team success. But moving on to entrepreneurs or movie stars, it is hard to know what “skill” statistic can be measured, to compare with some quantitative measure of success.

²By a student Beau La Mont, who was aiming at the center of the board. One missed the board.

³But I haven’t managed to write a satisfactory lecture on this topic.

Figure 8.2: 99 dart throws, centered on a $2.25'' \times 3.5''$ playing card.



It seems perfectly reasonable to model dart throws via density functions. A dart throw is a textbook example of a random point in two dimensions whose distribution is assumed to be described by some density function. What this means, in words, is that for any two points z and z' that are sufficiently close, the chance of the dart landing very close to z is almost the same as the chance of landing very close to z' . This seems so reasonable that it is rarely commented upon. A main point of this lecture is that in fact this kind of assumption, which is made in many contexts and which I will call the *local uniformity principle*, has more consequences than one might imagine. Let me contrast with contexts where one assumes *independence*, in which case we are fully aware that independence is an assumption (maybe justifiable on intuitive or empirical grounds) and that the reliability of any mathematical consequence we might derive is linked to the reliability of the assumption.

It seems perfectly reasonable to model dart throws as IID. One can think of several reasons the IID model might not be *accurate*. The first few throws may involve learning adjustments, and eventually the thrower might get tired or bored. But aside from such specific effects, the IID model seems conceptually *reasonable*. Now freshman textbooks seem to leave the impression (amongst non-professional statisticians) that any list of data⁴ can be regarded as IID samples from some probability distribution – if you do a textbook test of significance or confidence interval, then you are modeling them as IID. But consider for instance the areas of the 50 U.S. States, or the unemployment rates in each State next year, or the change in unemployment rate between this year and next year. To me there is no reason to think these are independent (in the sense of probability theory) between States. I will return to this issue in section 8.6.

8.3 County fair

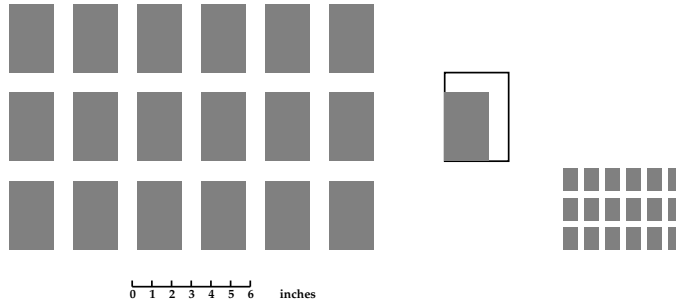
This section treats a concrete setting which will seem very special, but in the next two sections I will show that the abstract idea applies more broadly.

There is a fairground game in which playing cards⁵ are stuck to a large board in a regular pattern, with space between cards. See Figure 8.3. You

⁴Implicitly, the same attribute of different “individuals”.

⁵The game is also played with balloons as targets. The balloons are squashed together without empty spaces, making the game appear easier. But you soon learn that to burst a balloon, the dart needs to hit it head-on; a hit on the side just bounces off.

Figure 8.3: Playing cards on a wall. The playing cards on the left are “bridge size” 2.25 by 3.5 inches, with spacing 1 inch between rows. The wall is much larger than shown, with hundreds of cards attached. In the center is the “basic unit” of the repeating pattern. On the right is the pattern shrunk by a factor of 3.



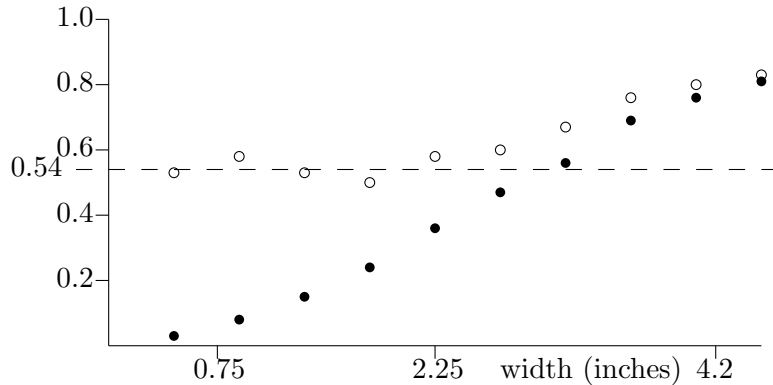
pay your dollar, get three darts, and if you can throw the darts and make them stick into three different cards then you win a small prize.

One could conduct time-consuming experiments with throwers of different skills and different patterns. But – for the point I want to make – I can be lazier and work with the previous data set of 100 throws by Beau, and see what would have happened with differently scaled cards. 36 throws would have hit a normal sized card as target, so we estimate the probability as 0.36. The \bullet in Figure 8.4 show how this probability increases with the card size. As one expects, this probability is near zero for a postage stamp size and near one for a paperback size. If we repeated the experiment with a different person we would confidently expect a curve of \bullet which was qualitatively similar but shifted left or right according to skill at darts.

Returning to the fairground game, we imagine scaling the pattern (as on the right of Figure 8.3). For normal size cards, Beau would have 58 hits (that is, 36 on the aimed-at card and 22 fortuitously on a different card) and these probabilities are shown as \circ in Figure 8.4. As we explain in a moment, without looking at data we can make a theoretical prediction that, regardless of skill level, when the “pattern repeat distance” becomes small the probability of hitting some card will become about 0.54. And the data shows this is indeed true for Beau on scales smaller than a playing card.

What does theory say about this example? The key point underlying the theory is that there is a *regular repeating pattern* on the wallboard, consisting of repeats of the *basic unit* in the center of Figure 8.3; the basic

Figure 8.4: In the setting of Figure 8.2, Beau’s estimated probability of hitting a specific card \bullet and the probability of hitting some card \circ , as a function of width. A small postage stamp has width about 0.75, a playing card 2.25, and a cheap paperback book 4.2.



unit is a rectangle of board, partly occupied by a card. Since the space between cards is 1 inch, this rectangle has size 3.25 by 4.5 inches. So the proportion of the area of the basic unit which is occupied by the card equals $(2.25 \times 3.5)/(3.25 \times 4.5) = 54\%$. Because the pattern just repeats the basic unit, this means that a proportion 54% of the wallboard is covered by cards. And this proportion is unchanged by shrinking cards and spaces together. So a dart hitting a region of the board, without propensity to hit or to miss cards, should have a 54% chance to hit a card. So the underlying theory is that, when the cards are small relative to the variability of our throws, we have little chance of hitting the particular aimed-at card, and instead our hit is essentially like hitting a purely random point. I will restate this theory idea in the next section as the *fine-grain principle*.

8.4 The physics of coin-tossing and the fine-grain principle

At first sight the notion of “regular repeated pattern” may seem special to the example above, but let me show that it occurs somewhat more broadly.

The physics of coin-tossing. Why do we think that a tossed coin should land Heads with probability 1/2? Well, the usual argument by symmetry goes something like this.

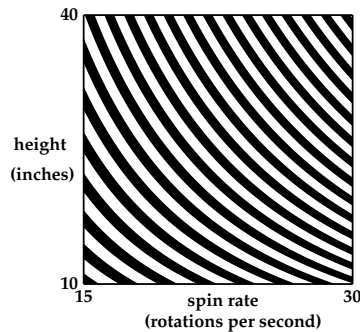
- There is some chance, p say, of landing Heads.
- By symmetry, there is the *same* chance p of landing Tails.
- Neglecting implausible possibilities (landing on edge, being eaten by passing bird, ...) these are the only possible outcomes.
- Since some outcome must happen, i.e. has probability 1, it must be true that $p + p = 1$.
- So $p = 1/2$.

Like most people, I find this argument (and the corresponding argument for dice, roulette etc) convincing. But such an argument by logic doesn't give much insight into where physically the number $1/2$ comes from. After a moment's thought, in this case the physics is actually quite simple, if we over-simplify matters a little. Suppose you toss a coin straight up, that it spins end-over-end relative to a horizontal axis, and that you catch the coin at the same height as you tossed it. Then the coin leaves your hand with some vertical speed v and some spin rate of r rotations per second. And there's no randomness – it either lands Heads for sure, or lands Tails for sure, depending on the values of v and r via a certain formula. Now we can't see v or r , but we can see the height h that the coin rises before starting to fall. Figure 8.5 shows the result of the coin toss in terms of h and r , for a certain interval of values.

At the instant we toss the coin, we are at some point in the *phase space* illustrated in the figure, and this point determines whether the coin lands Heads or Tails. I don't have any honest data for the points in phase space determined by an actual series of tosses. But if you practice tossing a coin 24 inches high, you will find it very difficult to be more accurate than 24 ± 3 inches, and so we may envisage a series of tosses as creating a collection of points in phase space scattered in some unstructured fashion in the spirit of Figure 8.3. The symmetry of the coin is reflected in the fact that the bands determining Heads or Tails have equal width; 50% of the phase space determines Heads. A machine can make tosses in such a consistent way that the spread in phase space was small compared to the width of the bands, but a person cannot. A person tossing a coin is like a person throwing darts at stamp-sized cards – without any bias toward any particular band, we have 50% chance to hit a point in phase space which determines Heads.

How can we abstract this idea to other settings? A mixture of peanuts and cashews is coarse-grained, in that you can pick and choose an individual

Figure 8.5: Phase space for coin tossing. The shaded bands are where an initially Heads-up coin will land Heads, as determined by the height and rotation rate of the toss. Each band indicates a specific number of rotations, from 7 to 27 over the region shown. The formula underlying Figure 4 is as follows. The height h and time-in-air t are determined by $h = v^2/(2g)$ and $v = gt/2$ where $g = 32$ feet per sec². So $t = \sqrt{8h/g}$. If the coin starts Heads-up, then it lands Heads after n rotations if $n - \frac{1}{4} < rt < n + \frac{1}{4}$. So the curves in the figure are the curves $r\sqrt{8h/g} = n \pm \frac{1}{4}$.



nut, while a mixture of salt and pepper is fine-grained, in that you can't avoid picking a mixture. This provides a nice metaphor.

The fine-grain principle. Many instances of physical randomness can be regarded as outcomes of deterministic processes with uncertain initial conditions: the randomness comes only from the initial uncertainty. A particular outcome corresponds to the initial conditions being in some particular subset of phase space, which we visualize as a collection of clumps such as the rectangles in Figure 8.3 or the curved bands in Figure 8.5. If the subset of phase space has a certain kind of regularity – that the local proportion of phase space that lies in the subset is approximately the same proportion p regardless of global position within phase space – then we can confidently predict that the probability of the outcome is about p , provided the spread of the distribution of the initial point is at least somewhat large relative to the distance between clumps. The numerical value p comes from the *pattern* in phase space, not from any details of the uncertainty in initial conditions.

The fine-grain principle is one of those good news/bad news deals. As

a conceptual idea it's very nice; throwing a die⁶ to roll on a table is a much more complicated deterministic process than coin-tossing, but one can imagine a high-dimensional phase space which is divided into six regions in some complicated way analogous to Figures 8.3 and 8.5. As a concrete tool it's terrible, because for die-throwing (or just about any real-world example more complicated than the two we've discussed) one can't actually work out what is the pattern in phase space. The actual reason we all believe that a die lands 5 with probability $1/6$ is the argument by symmetry (and a supposition that other people have checked it empirically). We can't honestly "do the physics" to confirm this via the fine-grain principle, but we have a world-view (i.e. a belief) that it would be confirmed if we could⁷

To digress, there is also a "proof by economics". If the fine-grain principle were false for die-throwing, then there would be a way of throwing a die so that the landing probabilities were non-uniform – *and some gambler would already have discovered this.*

8.5 The smooth density idealization for data and Benford's law

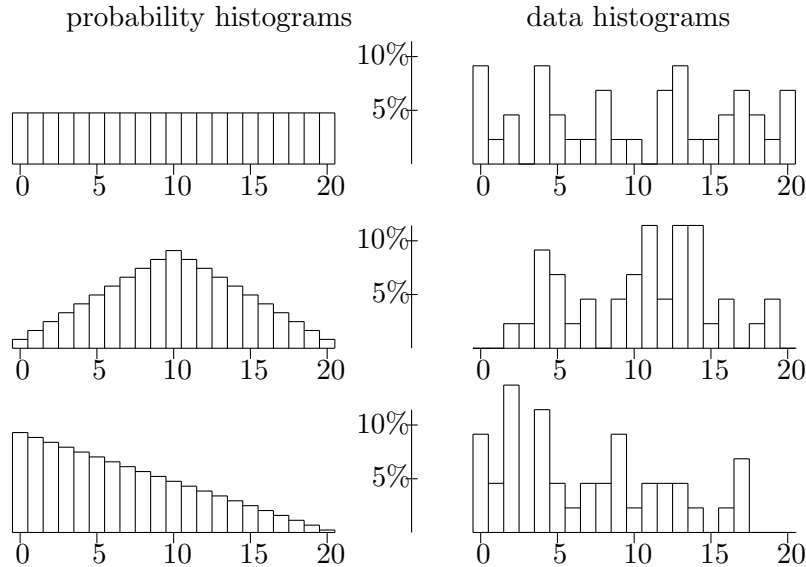
We turn to an idea that is much more broadly applicable, though less dramatic in its consequences. We first illustrate the idea with artificial data, then with real data. Consider the simplest kind of data-set, a list of numbers, say 40 numbers between 0 and 20; to have a convenient language think of these as exam scores for 40 students. Figure 8.6 illustrates several such artificial data sets. 40 scores are picked IID according to the probability distribution on the left, and the empirical histogram is shown on the right.

In the second case the model is deliberately biasing likely values toward the center, and in the third case toward the left, and this non-uniformity is readily visible in the data. But the aspect I want to emphasize here is a *similarity* between all three cases. Each model is "smooth" in the sense that probabilities do not change much from one possible score to the next, whereas in each case the data is "locally irregular". That is, in the first model we expect each score to come up "on average" about twice, but the exact number of times is often 1 or 3 (rather than 2) and sometimes 0 or 4; the observed frequencies of successive scores switch unpredictably between

⁶I humorously tell students to write on the blackboard 100 times: the singular of *dice* is *die*.

⁷Analogous to the world-view that coincidences happen no more often than is explainable by chance.

Figure 8.6: Each right side shows a data histogram for 40 random picks from the probability histogram on the left side.



these values. The average frequency varies between models and between different parts of the range of scores, but always has this local irregularity.

Moving quickly on to real data, Figure 8.7 is a histogram of actual scores for a class of 71 students.

With large data sets authors often accompany the real data histogram with a smoothed curve – in technical language, an estimated *density function* – such as Figure 8.8 below.

Let me invent a name for what we are doing here.

The smooth density idealization. Many statisticians implicitly believe that it is helpful to associate with actual data (such as in Figure 8.7) some theoretical smooth histogram (such as in Figure 8.8). Textbooks and practitioners rarely discuss what the smooth curve is intended to represent. In a case where it is reasonable to regard the data as IID random samples from some unknown smooth distribution, then of course we can regard the curve as an estimate of the unknown density function; there is mathematical theory on how to do the estimate, and computer packages will draw the estimated density function curve for you.

Figure 8.7: Scores for a class of 71 students. The maximum possible score was 92.

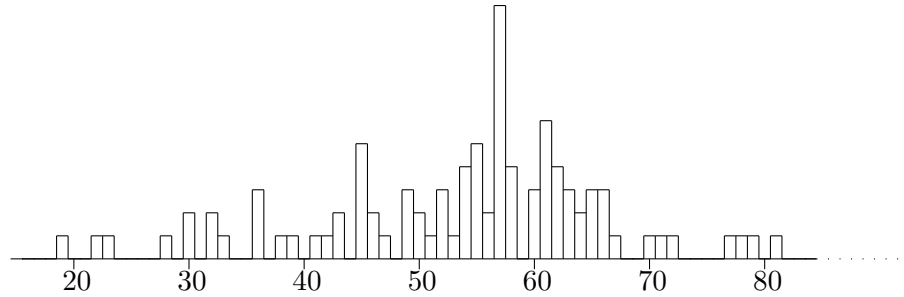
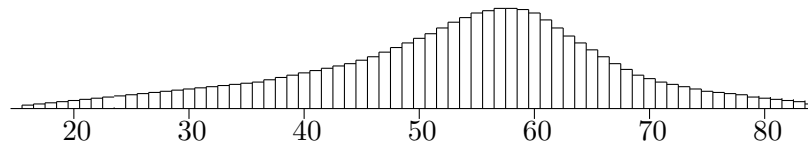


Figure 8.8: A possible theoretical histogram one might associate with the Figure 8.7 data.



But data like exam scores, the areas of the 50 States, or the wine-case data in Figure 8.9 later, are in fact not random samples of anything. In such cases – surely the majority of cases where density estimates are used – I find it hard to articulate exactly what the associated smooth density is intended to mean. In the particular context of exam scores, one might say it represents what one might expect the scores to be based on some qualitative information. That is, there is no reason to think that a score of 47 will be substantially more or less frequent than a score of 51 – so the theoretical histogram will be *smooth* – but good reason to believe that either of those scores will appear more frequently than a score of 2 or 89 – so it’s not uniform over the entire range. But such interpretations seem too context dependent.

Let me emphasize that in the smooth density idealization we are appealing to some vague notion of “associated with” – the conceptual question (which I cannot answer) is to say more clearly what is meant. This is distinct from the precise, thought stronger and harder to justify, notion of “IID samples from”. I will revisit this kind of conceptual issue in section 8.6.

Checking predictions of the smooth density idealization. Instead of worrying *a priori* about whether it is reasonable to model data as being associated with some unknown smooth distribution, we could ask whether observed data is *consistent* with this assumption. To do this we need to find some quantitative theoretical predictions for such data, which (to be useful) should not depend on the unknown distribution. Let me give two simple examples of such quantitative predictions, described in the context of our exam data. A course project is to repeat this analysis for other data sets.

Prediction 1: Coincidences and near misses. For a given pair of students, they might get the same score (call this a coincidence) or they might get consecutive scores (call this a near-miss). So we can count the number of coincidences (that is, the number of *pairs* of students with the same score) and we can count the number of near-misses. The theoretical prediction is

if a data set is associated with a smooth distribution, then the number of near-misses will be about twice the number of coincidences.

In the data of Figure 8.7 there are 49 pairs representing coincidences and 86 pairs representing near-misses. So the prediction works pretty well.

The math argument. If student A scores (say) 45 then (by supposition of smooth distribution) the chances of student B scoring 44 or 45 or 46 are approximately equal, so the chance of a near miss is approximately twice the chance of a coincidence.

Prediction 2: Least significant digit. For a score of 57 the “most significant” first digit is 5 and the “least significant” second digit is 7. Looking at the most significant digit of the Figure 6 data (as one might do for assigning letter grades) we clearly are going to see substantial non-uniformity, and indeed we do

first digit	1	2	3	4	5	6	7	8
frequency	1	3	10	15	18	17	6	1

If (with less motivation) we look at the second digit, there is a theoretical prediction:

if a data set is associated with a smooth distribution, then the distribution of least significant digits will be approximately uniform.

In this data set it is:

second digit	0	1	2	3	4	5	6	7	8	9
frequency	6	8	10	6	4	11	10	7	3	6

The math argument. For 3 we add the frequencies of ...43, 53, 63, ... and for 4 we add the frequencies of ...44, 54, 64, ...; by supposition the probabilities being added are approximately equal.

The small print. As above, the supposition of a smooth theoretical distribution must be plausible. And obviously if all data values are within 7 of each other the prediction can't be correct, so we need a condition of the form "the *spread* of data is not small relative to 10". Measures of spread are a textbook topic; let's use the interquartile range (the difference between the 25th percentile and the 75th percentile), which in this data set is $62 - 43 = 19$. Rather arbitrarily, let's say the prediction should be used only when

$$\text{the interquartile range is more than 15.} \quad (8.1)$$

Benford's law. Figure 8.9 shows a histogram⁸ for a data-set of the total production (number of cases) of each of 337 wines reviewed by Wine Spectator magazine in December 2000 (this data is not claimed to be representative of all wine production). So the data is a list like 517, 5300, 1490, ...; the minimum was 30 cases and the maximum⁹ was 229,165 cases. Note the log scale on the horizontal axis, used to fit such widely varying data onto one figure.

For obvious economic reasons, few wines have production levels of 1 case or 1 million cases, and one might identify less obvious practical reasons for other features of the data. It may seem surprising that such data could be used to illustrate any general principle, but it can. Look at the first digit of each number in the data – so that 45 or 4,624 or 45,000 are each counted as "4". There is a theoretical prediction, called **Benford's law**, for the frequencies of the 9 possible first digits in data like this. Table 8.1 shows the predictions, derived from a formula written later.

The surprise is that theory doesn't predict equal frequencies, but instead predicts that 1 should appear much more often than 9 as a first digit. Figure 8.10 shows that for our wine-case data the prediction is fairly good – certainly much better than the "equal frequency" prediction which would imply a flat histogram.

⁸Here one needs to be careful about the precise definition of a histogram, which represents data by *area*.

⁹"Medium-bodied and a bit rustic, but a good everyday quaff".

Figure 8.9: The wine-case data.

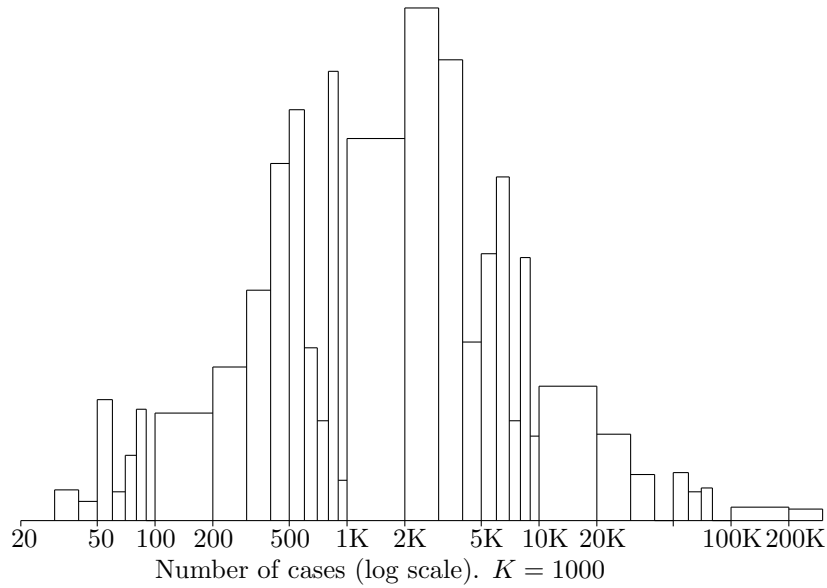
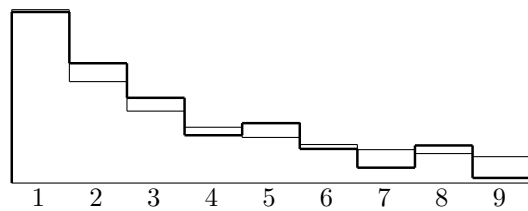


Table 8.1: What Benford's law predicts.

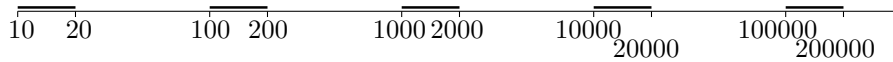
first digit	1	2	3	4	5	6	7	8	9
frequency	.3010	.1761	.1249	.0969	.0792	.0669	.0580	.0512	.0458

Figure 8.10: Benford's law and the wine-case data. The thick lines show the histogram for first digit in the wine-case data; the thin lines are the histogram of frequencies predicted by Benford's law, Table 8.1.

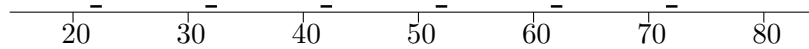


Though striking and memorable, upon a little reflection one realizes that Benford’s law is just a lightly disguised instance of an ideas discussed earlier in this lecture.

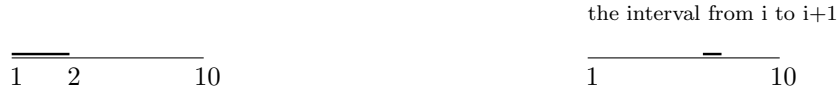
The math argument. Look again at Figure 8.9. What parts of the horizontal axis correspond to case numbers with first digit 1? That’s easy to picture.



Compare with the exam score data, where we look at a given *second* digit, say 2:



Both pictures show a “repeating pattern” on the one-dimensional line, analogous to the two-dimensional repeating pattern in Figure 2 (playing cards on the wall), whose “basic unit” is shown on the left below.



In each case, the smooth density idealization implies that the proportion of data in the marked subset will be approximately the density of the marked subset, that is the proportion of the length of the “basic unit” that is in the subset. Since we’re working on a log scale, the length of this line from 1 to 10 is $\log 10$, which (interpreting “log” as “log to base 10”) equals 1; and the part of the line from 1 to 2 has length $\log 2$, which works out to be 0.301. Similarly, for each possible first digit i there is a repeated unit, as shown on the right in the diagram above, and we get the formula

$$\text{predicted frequency of } i \text{ as first digit} = \log(i + 1) - \log i$$

which gave the numbers in Table 8.1.

The small print. To a mathematician, the Benford prediction is equivalent to the second significant digit prediction¹⁰, applied to $\log(\text{data})$. The implicit assumptions are smoothness (on the log scale) and sufficient spread, which (copying (8.1) but undoing the log transformation) becomes

¹⁰i.e. the least significant digit prediction applied to 2-digit numbers, as in our exam scores data.

$$(75\text{th percentile})/(25\text{th percentile}) \geq 30 \approx 10^{1.5} .$$

So the key requirement for the plausible applicability of Benford’s law is that the data be widely varying – that it not be too uncommon to find two numbers in the data where one number is more than 30 times the other.

Checking Benford’s law. Checking the accuracy of the Benford prediction on real data, and in particular checking the implicit prediction that (for similar size data sets) it will be more accurate for more spread-out distribution, is a natural course project [written up here as an illustration of course projects](#).

8.6 Which of 4 foundational principles do you believe?

Let me first clarify the distinction used, in the context of dart throws in section 8.2, between observed data “following some smooth density f ” and “being IID samples from density f ”. The latter is a very precise quantitative assertion; the former is the more qualitative assertion that the observed proportion of data values between a and b is “approximately” $\int_a^b f(x)dx$, without seeking to quantify what “approximately” means.

What do you think about the following 4 general default assumptions?

1. Assume different outcomes are equally likely, unless there is a reason to believe otherwise.
2. Assume different events are independent, unless there is a reason to believe otherwise.
3. Assume data¹¹ comes from a smooth density unless there is a reason to believe otherwise.
4. Assume data arises as IID samples from a smooth density unless there is a reason to believe otherwise.

Now one reaction to the first two assertions is “but almost always there *is* a reason to believe otherwise”. In other words, they are formally true but essentially vacuous. I would not argue with that view, but would prefer to use the opposite default:

¹¹Implicitly, the same attribute of different “individuals”.

- Do not assume different outcomes are equally likely, unless there is a reason to believe so.
- Do not assume different events are independent, unless there is a reason to believe so.

What about the third and fourth assertions? Let me use the exam data in Figure 8.7 as a paradigm example. My own view is that in such examples, assertion 3 is reasonable – the estimated smooth distribution in Figure 8.8 means something, even if it’s hard to say what it means – but assertion 4 is not *a priori* sensible – where is the asserted independence coming from?

Anyway, my point is not to set forth some argument for my personal views, but rather to state that I find it strange that these questions are not discussed more seriously in statistics textbooks, because tests of significance or their Bayesian counterparts depend on IID or analogous specific models. Let me continue with two more focused points.

Testing for IID. Testing whether data might plausibly have arisen as IID samples from a *specified* distribution or family of distributions is a staple¹² of classical mathematical statistics. Testing whether it might plausibly have arisen as IID samples from some *unspecified* smooth distribution is a less standard topic, but it is not hard to devise several schemes. One involves testing the local irregularity of observed frequencies. For the exam data, write f_i for the frequency of score i . Then “smoothness” says that on average f_i should be close to $(f_{i+1} + f_{i-1})/2$, and the squared difference $[f_i - (f_{i+1} + f_{i-1})/2]^2$ is a measure of local irregularity. This suggests looking at the statistic

$$S = \sum_i [f_i - (f_{i+1} + f_{i-1})/2]^2.$$

As outlined below, theory says that for IID samples from a smooth distribution, S should be around $3n/2$ when n is the number of observations.

The math argument. Smoothness says that the probabilities p_i satisfy $p_i \approx (p_{i+1} + p_{i-1})/2$; “IID random” says that the random variables f_i should be approximately Poisson distributed with mean np_i . This gives

$$E[f_i - (f_{i+1} + f_{i-1})/2]^2 \approx \text{var}[f_i - (f_{i+1} + f_{i-1})/2] \approx n[p_i + (p_{i+1} + p_{i-1})/4] \approx np_i \times 3/2$$

as the first-order approximation.

It is a course project to investigate S and other test statistics, for interesting observational data sets such as those mentioned next.

¹²Kolmogorov-Smirnov test

On modeling observational data as IID. As our freshman statistics textbook¹³ puts it,

Chance models are now used in many fields. Usually, the models only assert that some things behave like the tickets drawn at random from a box.

It is a pervasive custom in some fields to model observational data as IID from some unknown distribution, and try to infer something about that distribution. To illustrate, I copy below brief descriptions of 24 data sets used in [this 2009 Clauset-Shalizi-Newman paper](#). Let me emphasize that the paper itself is a careful and valuable analysis of how closely such data-sets follows some power law distribution; my point is that seeking to quantify conclusions via any calculation based on an IID assumption is hard to justify¹⁴. The point is that independence does not mean “I can’t think of any direct connection” but is a precise assertion about numerical equality of conditional probabilities, and it is hard to argue this is a plausible assertion in contexts where (as in most examples below) there is no clear notion of what “probability” refers to.

1. The frequency of occurrence of unique words in the novel *Moby Dick*.
2. The degrees (i.e., numbers of distinct interaction partners) of proteins in the partially known protein-interaction network of the yeast *Saccharomyces cerevisiae*.
3. The degrees of metabolites in the metabolic network of the bacterium *Escherichia coli*.
4. The degrees of nodes in the partially known network representation of the Internet at the level of autonomous systems for May 2006.
5. The number of calls received by customers of AT&T’s long distance telephone service in the United States during a single day.
6. The intensity of wars from 1816–1980 measured as the number of battle deaths per 10 000 of the combined populations of the warring nations.
7. The severity of terrorist attacks worldwide from February 1968 to June 2006, measured as the number of deaths directly resulting.
8. The number of bytes of data received as the result of individual web (HTTP) requests from computer users at a large research laboratory during a 24-hour period in June 1996.

¹³Freedman-Pisani-Purves-Adhikari *Statistics*.

¹⁴This is not a frequentist vs Bayesian issue. I have no quarrel with thinking of unknown quantities as random, just with assuming independence or a Bayesian analog without any reason.

9. The number of species per genus of mammals.
10. The numbers of sightings of birds of different species in the North American Breeding Bird Survey for 2003.
11. The numbers of customers affected in electrical blackouts in the United States between 1984 and 2002.
12. The numbers of copies of bestselling books sold in the United States during the period 1895 to 1965.
13. The human populations of US cities in the 2000 US Census.
14. The sizes of email address books of computer users at a large university.
15. The sizes in acres of wildfires occurring on US federal land between 1986 and 1996.
16. Peak gamma-ray intensity of solar flares between 1980 and 1989.
17. The intensities of earthquakes occurring in California between 1910 and 1992, measured as the maximum amplitude of motion during the quake.
18. The numbers of adherents of religious denominations, bodies, and sects, as compiled and published on the web site.
19. The frequencies of occurrence of US family names in the 1990 US Census.
20. The aggregate net worth in US dollars of the richest individuals in the United States in October 2003.
21. The number of citations received between publication and June 1997 by scientific papers published in 1981 and listed in the Science Citation Index.
22. The number of academic papers authored or coauthored by mathematicians listed in the American Mathematical Society's MathSciNet database.
23. The number of "hits" received by web sites from customers of the America Online Internet service in a single day.
24. The number of links to web sites found in a 1997 web crawl of about 200 million web pages.

8.7 The local uniformity principle and asteroid near-misses

I used the phrase *fine-grain principle* in the context of deterministic processes with (slightly) random initial conditions, and the phrase *smooth density idealization* in the context of fairly general observational data from human society. Trying to relate these very different contexts seems unreasonably speculative, but the mathematics is similar, so let me invent a third and

much vaguer phrase *local uniformity principle* to denote what we are doing if we *assume*, in any context, that experimental or observational quantities at some level are random with some locally smooth distribution. Here is another example.

99942 Apophis is a 350-meter long asteroid which is confidently predicted to pass Earth just below the altitude (35,000 km) of geosynchronous satellites (which provide your satellite TV) on Friday, April 13, 2029.


This fact prompts discussion of the chances of an asteroid collision with Earth. The actual spatial density of different sized asteroids at different points in the solar system is of course an empirical issue. But it's perfectly reasonable to make the "local uniformity" assumption that the chance of an asteroid being at one point near the Earth's orbit is not substantially different from the chance of it being at a different point a quarter million miles away. Then mathematics, and the empirical fact that the ratio (radius of Moon orbit)/(radius of Earth) is approximately 60, shows

Amongst asteroids which pass closer to Earth than the Moon's orbit, about one in 3,600 ($= 60^2$) will hit Earth.

Hypothetically, if astronomers could and did detect all asteroids of diameter greater than 50 meters passing within the Moon's orbit for a period of years, and found there were on average 3.6 per year, then one could infer that such an asteroid would hit the Earth about once every 1,000 years on average¹⁵.

8.8 Wrap-up and further reading

I occasionally indulge in a quixotic quest to retire "die" as the icon for randomness, because

- dice are greatly overused, both as a verbal metaphor and as a visual image – even Wikipedia uses the graphic 
- dice are simply unrepresentative of the way we really do encounter chance in the real world.

As section 8.2 suggests, "dart throws" are my best suggestion for a replacement.

The big picture I have tried to explain in this lecture is undoubtedly widely known to academics at some vague level. But the only attempt I

¹⁵This numerical conclusion is a typical estimate in the literature but not obtained in this hypothetical way.

know at a systematic exposition is a 2003 monograph *Bigger Than Chaos* by Streven, which I find rather impenetrable. Similarly, the analysis of coin-tossing has undoubtedly been rediscovered many times – most cited is a 1986 paper by Keller.

Real coin tossing. In a fascinating 2007 paper by Diaconis - Holmes - Montgomery, high-speed photography of coin tosses shows that in fact the axis of rotation is typically not horizontal and that the axis precesses during the flight. This leads to a theoretical prediction that a tossed coin should land *the same way up as it was thrown* with probability about 51%. To detect this effect, i.e. a difference from 50%, one would need about 40,000 tosses. Apparently the only actual experiment was done by two of my students, and the title says it all: 40,000 coin tosses yield ambiguous evidence for dynamical bias.

In fact two examples have long been known where the “argument by symmetry” gives substantially wrong answers. One concerns spinning a coin vigorously on its edge and waiting for it to fall; a typical U.S. penny is noticeably biased toward Tails. The second concerns holding a wine cork horizontally and dropping it onto a table with a hard surface. One would think that (like a tossed coin finishing upright on its edge) it is very unlikely for the cork to bounce and finish in upright position. But in fact, starting about 1.5 cork lengths above the table, it is not so unlikely – try it!¹⁶

Fine-grain principle. Though the principle is well-understood, there is no standard phrase: *fine-grain principle* is my coinage, though the phrase fine-grain is used literally and metaphorically in several areas of science. Finding precise mathematical formulations that capture actual usage seems difficult; neither long-run arguments (ergodic theory) nor asymptotics based on imagining we can shrink the pattern¹⁷ really engage the issue, which is more akin to studying the number of card shuffles required to mix a deck (Lecture ??).

¹⁶A synthetic cork works best.

¹⁷von Plato (1983) *The method of arbitrary functions*, not open access.

Chapter 9

A glimpse at research: spatial networks over random points

Students find it difficult to envisage what research in mathematical probability consists of. This lecture is my attempt to illustrate, using the part of my own recent research that is least technical and most amenable to description via graphics. It is based on [this overview paper](#) and [this overview paper](#), both with undergraduate co-authors who did simulations and graphics.

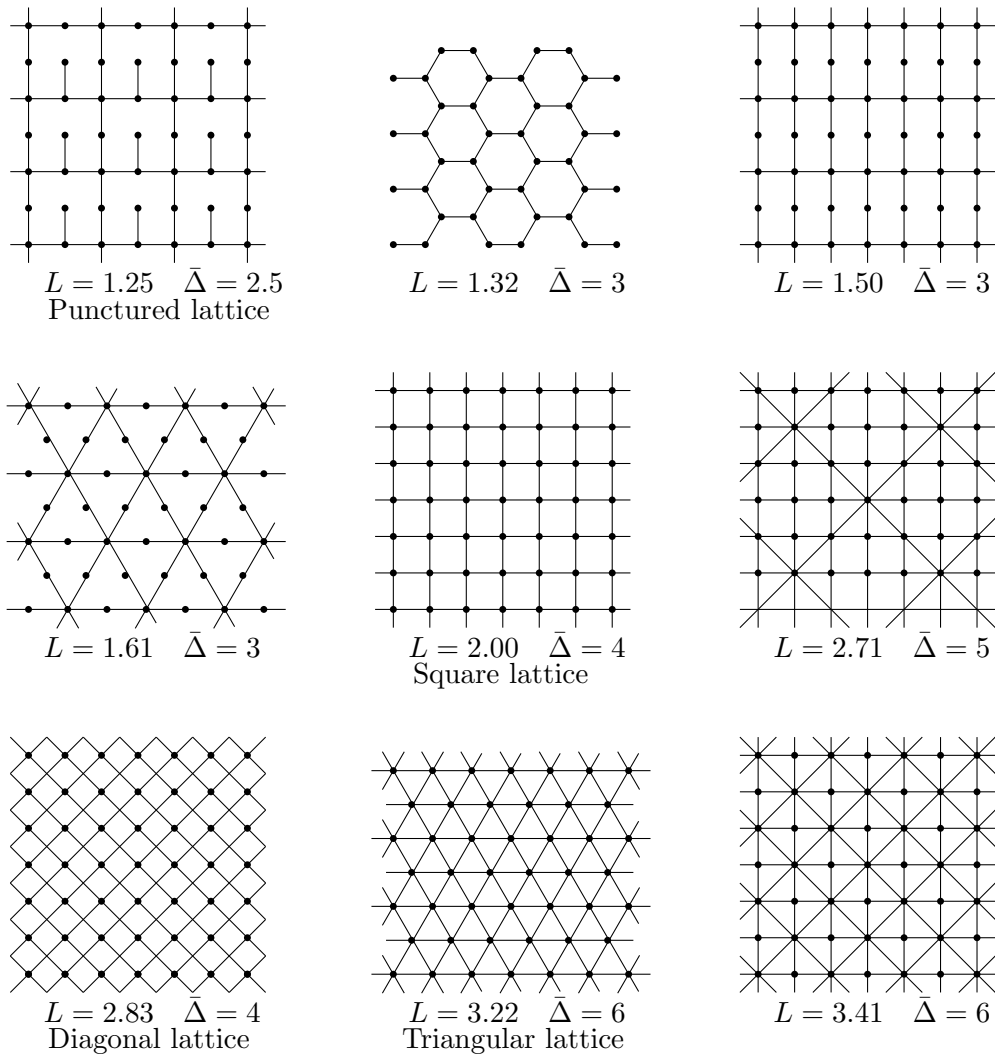
9.1 Regular networks

By a [spatial network](#) ¹ I mean one whose edges are physically situated in two-dimensional space. This contrasts with the broader field of [complex networks](#), focussing mainly on networks such as WWW links or social networks, for which the edges are abstract or the geographical locations are relatively unimportant. In particular let us think of road networks, for which extensive data is readily available.

Figure 9.1 illustrates hypothetical “regular” or “structured” networks, and mathematicians have implicitly studied (via Euclidean geometry) such networks for many centuries. In that figure there are only two different patterns of vertices, but different ways of linking vertices by edges. For comparison purposes it is natural to scale so the density of vertices equals 1 per unit area. For each pattern the figure states the average degree (number

¹The Wikipedia page is rather incoherent.

Figure 9.1: Variant square, triangular and hexagonal lattices. Drawn so that the density of cities is the same in each diagram, and ordered by value of L .



of edges per vertex) $\bar{\Delta}$ and the average edge-length per unit area L .

Consider a map showing the major roads linking cities. Of course cities are not arranged in a regular pattern, so the map will not look much like those in Figure 9.1. Let us instead imagine that city positions are random points, in these sense of a Poisson process of rate 1 per unit area. Are there mathematically natural ways to define links between cities – “roads” – so that the network looks somewhat like an actual road network?

9.2 Networks on random points

A first issue is that we want the network to be connected. The first “local” rules one might invent – link each point to the k closest points, or to all points within distance d – do not always give connected networks. Figure 9.2 (left) shows the **relative neighborhood network** defined by

put a link between cities A and B if there is no other city C such that the distances from C to A and to B are each smaller than the distance from A to B

(here *distance* is straight-line distance). Informally, the relative neighborhood network is the sparsest (fewest links) network that can be defined by a simple local rule and is always connected².

This idea of defining a network by a “no other city” rule can be extended to define the one-parameter family of **beta-skeleton** networks. One member of this family, shown in Figure 9.2 (right) is the **Gabriel** network, defined by

put a link between cities A and B if there is no other city C within the circle on which A and B are diametrically opposite.

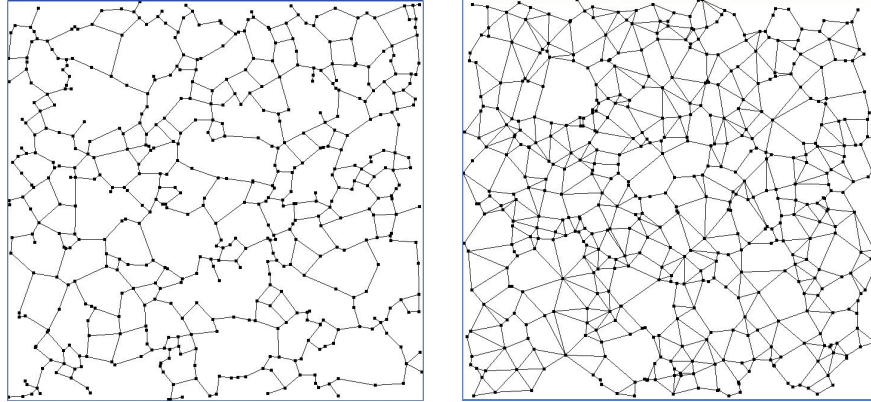
The networks of Figure 9.2 do not resemble maps of modern road networks, though I imagine that they might resemble a map of roads or footpaths between medieval villages³. It turns out to be surprisingly hard to devise simple mathematical models yielding networks that visually resemble modern road networks. Indeed I do not know any such models, though the following two sections could be viewed as indirect approaches toward that goal.

Let me first show some mathematical results concerning the networks above, for random points of mean density 1. It turns out there are simple formulas for the mean degree $\bar{\Delta}$ and the mean length-per-unit-area L , as

²because it contain the minimum spanning tree.

³A student project is to find such maps to obtain some actual data.

Figure 9.2: The relative neighborhood network and the Gabriel network on random points.



follows. Write c for the area of the excluded region in the definition of the Gabriel network or the relative neighborhood network when cities A and B are distance 1 apart. Then

$$L = \frac{\pi^{3/2}}{4c^{3/2}} \quad (9.1)$$

$$\bar{\Delta} = \frac{\pi}{c}. \quad (9.2)$$

For the Gabriel network we immediately see $c = \pi/4$, and for the relative neighborhood network a brief calculation shows $c = \frac{2\pi}{3} - \frac{\sqrt{3}}{4}$. What is fascinating is that for the Gabriel network we therefore find

$$\bar{\Delta} = 4, \quad L = 2$$

which exactly coincides with the values for the square lattice. I do not know of any explanation to suggest this is more than mere coincidence, but it does suggest treating the Gabriel network as the random-point analog of the square lattice.

Calculation for (9.1, 9.2). For readers familiar with calculations involving the Poisson point process, here is the essence of the calculation. Take a typical city at position x_0 . For a city x at distance s the probability that (x_0, x) is an edge equals $\exp(-cs^2)$ and so

$$\text{mean-degree} = \int_{\mathbb{R}^2} \exp(-c\|x - x_0\|^2) dx = \int_0^\infty \exp(-cs^2) 2\pi s ds$$

$$L = \frac{1}{2} \int_{\mathbb{R}^2} \|x - x_0\| \exp(-c\|x - x_0\|^2) dx = \int_0^\infty s \exp(-cs^2) 2\pi s ds.$$

Evaluating the integrals gives (9.2,9.1).

In studying road networks a major feature of interest is (shortest) route-length. Write $\ell(i, j)$ for the route-length (length of shortest path) between cities i and j in a given network, and $d(i, j)$ for Euclidean distance between the cities. So $\ell(i, j) \geq d(i, j)$, and we write

$$r(i, j) = \frac{\ell(i, j)}{d(i, j)} - 1$$

so that “ $r(i, j) = 0.2$ ” means that route-length is 20% longer than straight line distance. It is not practical to find useful explicit formulas for route-lengths in models, but easy to find numerical values by simulation. Let us consider

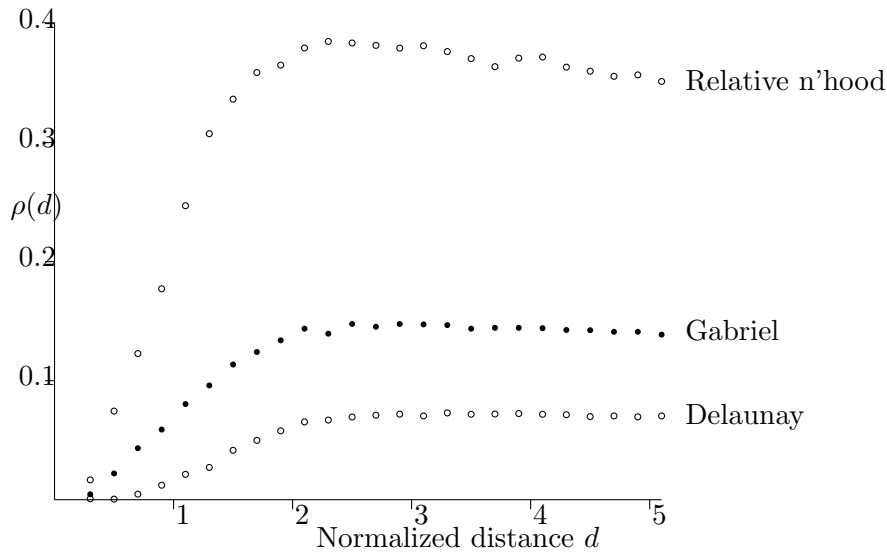
$$\rho(d) := \text{mean value of } r(i, j) \text{ over city-pairs with } d(i, j) \approx d \quad (9.3)$$

This function is shown in Figure 9.3 for the two networks above and for a denser network, the **Delaunay triangulation**.

Now obviously the values of $\rho(d)$ will tend to be smaller for denser networks. But for other reasonable model of connected networks on random points the function $\rho(d)$ has the same characteristic shape as in Figure 9.3), attaining its maximum between 2 and 3 and slowly decreasing thereafter. This characteristic shape – exhibited also in real-world networks (see Figure 9.6 later) – has a common-sense interpretation. Any efficient network will tend to place roads directly between unusually close city-pairs, implying that $\rho(d)$ should be small for $d < 1$. For large d the presence of multiple alternate routes helps prevent $\rho(d)$ from growing. At distance 2 – 3 from a typical city i there will be about $\pi 3^2 - \pi 2^2 \approx 16$ other cities j . For some of these j there will be cities k near the straight line from i to j , so the network designer can create roads from i to k to j . The difficulty arises where there is no such intermediate city k : including a direct road (x_i, x_j) would increase L , but not including it would increase $\rho(d)$ for $2 < d < 3$. So when a network designer is trying to minimize $\rho(d)$ for given L , the difficult values of d are around $[2, 3]$.

A final comment is that Figure 9.3 suggests there are limits $\rho(\infty) = \lim_{d \rightarrow \infty} \rho(d)$ in those three models, and indeed **more sophisticated math theory** shows that the limit exists in any reasonable network model over random points.

Figure 9.3: The function $\rho(d)$ for three theoretical networks on random cities. Irregularities are Monte Carlo random variation.



9.3 An optimality criterion for road networks

One way to look at real-world intercity road networks with a mathematical perspective would be to compare the actual network to some hypothetical “optimal” network connecting the cities. But what optimality criteria should we use?

A main goal of a road network is to provide short routes. Recall that “ $r(i, j) = 0.2$ ” means that route-length is 20% longer than straight line distance. With n cities we get $\binom{n}{2}$ such numbers $r(i, j)$; what is a reasonable way to combine these into a single statistic R , which measures how effective the network is in providing short routes? Two natural possibilities⁴ are

$$\begin{aligned} R_{\max} &:= \max_{j \neq i} r(i, j) \\ R_{\text{ave}} &:= \text{ave}_{(i,j)} r(i, j) \end{aligned} \tag{9.4}$$

where $\text{ave}_{(i,j)}$ denotes average over all distinct pairs (i, j) . However, being an “extremal” statistic R_{\max} seems unsatisfactory as a descriptor of real world

⁴The statistic R_{\max} has been studied in the context of the design of **geometric spanner networks** where it is called the *stretch*.

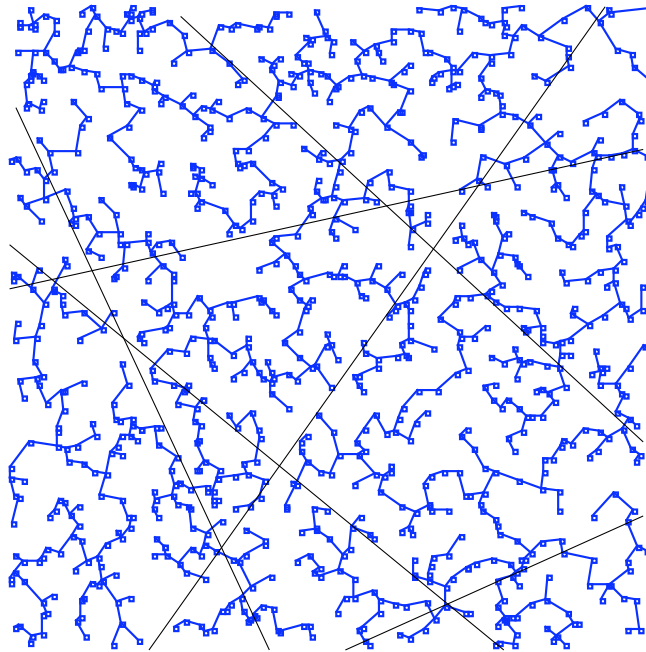
networks – for instance, it seems unreasonable to characterize the U.K. rail network as inefficient simply because there is no very direct route between Oxford and Cambridge.

The statistic R_{ave} has a more subtle drawback, which is a nice illustration of the pitfalls of the usual theory methodology of comparing statistics in the asymptotic ($n = \text{number of cities} \rightarrow \infty$) regime. Consider a network consisting of

- the minimum-length connected network – that is the **Steiner tree** – on given cities;
- and a superimposed sparse collection of randomly oriented lines (a *Poisson line process*).

See Figure 9.4.

Figure 9.4: An artificial network



By choosing the density of lines to be sufficiently low, one can make the normalized network length be arbitrarily close to the minimum needed for connectivity. But it is easy to show that one can construct such networks

so that $R_{\text{ave}} \rightarrow 0$ as $n \rightarrow \infty$. Of course no-one would build a road network looking like Figure 9.4 to link cities, because there are many pairs of nearby cities with only very indirect routes between them. The disadvantage of R_{ave} as a descriptive statistic is that (for large n) most city-pairs are far apart, so the fact that a given network has a small value of R_{ave} says nothing about route-lengths between *nearby* cities.

One can avoid this pitfall by using a statistic R which is intermediate between R_{ave} and R_{max} , defined by

$$R := \max_{0 \leq d < \infty} \rho(d). \quad (9.5)$$

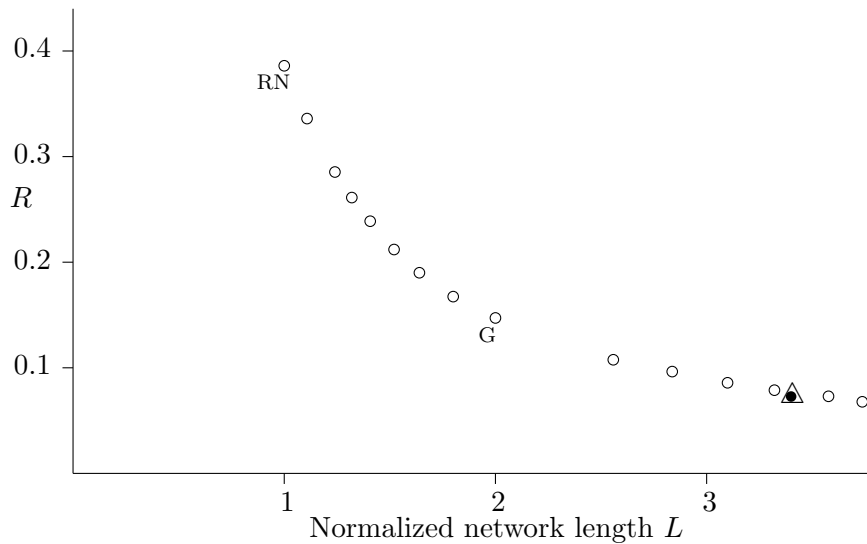
for the function ρ discussed in the previous section. In words, $R = 0.2$ means that on every scale of distance, route-lengths are on average at most 20% longer than straight line distance. For a mathematical model network on many random points R_{ave} is essentially $\rho(\infty)$. We see in the Figure 9.3 hypothetical networks, and (we suspect) in typical real-world networks, that R is only slightly larger than R_{ave} .

As a conclusion, the criterion we propose for “optimality” of a network is that the network has minimum R for its given length. We could use this in the context of n real-world cities – compare R and total length for the real-world network with the minimum value of R over hypothetical networks linking the cities with the same total length. Or in our model of random points as a rate-1 Poisson process on the plane, we could ask what are the minimum R and the corresponding minimizing network each given value of $L = \text{length-per-unit-area}$. I cannot answer these questions, because I do not have effective algorithms for finding the optimal networks. Regarding the latter question, Figure 9.5 shows the $R - L$ trade-off for the β -skeleton family. Seeking to improve on that family has been a student project, but only very small improvements have been obtained, so it is possible that the β -skeleton family are in fact close to optimal in our sense.

9.4 Scale-invariant random networks

The networks described already, based on “local rules”, are frankly unrealistic for intercity road networks, which in practice arise from centralized planning of a “backbone” of long-distance and fairly straight major roads. I don’t know a good simple model for networks of this type, which would involve some explicit use of roads at different levels of a major road - minor road spectrum. Instead I will outline a quite different approach in which

Figure 9.5: The length-per-unit-area L and the route length efficiency statistic R for certain networks on random points. The \circ show the beta-skeleton family, with RN the relative neighborhood network and G the Gabriel network; and \triangle shows the Delaunay triangulation.



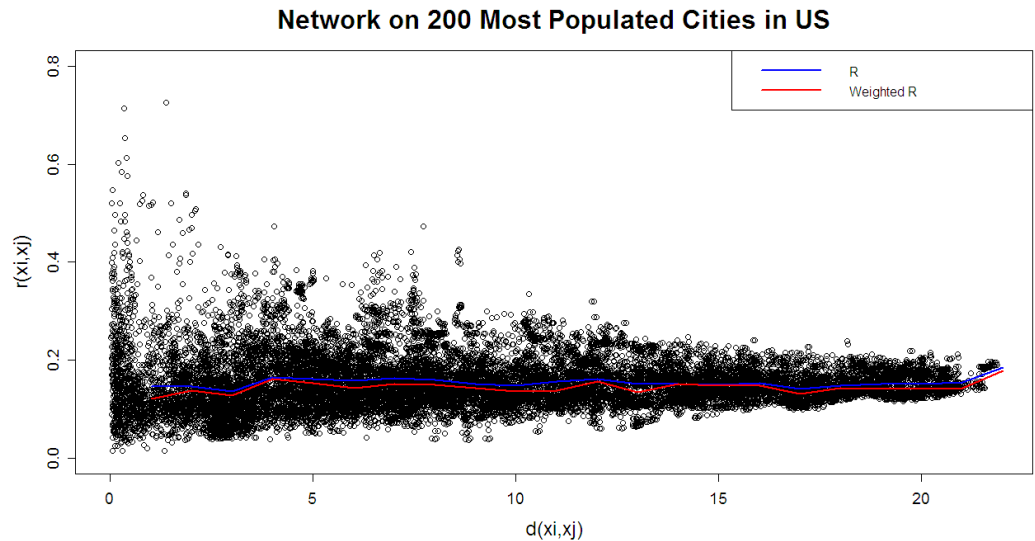
such a spectrum emerges without being explicitly assumed. We will seek to model the entire road network⁵.

Figure 9.6 shows a scatter diagram of the values $r(i, j)$ of relative excess route lengths for routes between the 200 largest U.S. cities. As mentioned before, the “line of averages” is roughly constant, consistent with our earlier discussion of $\rho(d)$. It also shows a “law of large numbers” effect, that the spread in values of $r(i, j)$ decreases as the distance increases. However, major city centers are not “typical” places with respect to the road network; if we looked at the corresponding scatter diagram for the routes between 200 random addresses, we would expect slightly larger averages and also less rapid decrease of spread.

A *scale-invariant* network is (very informally) one with the property

⁵Here is a fascinating map showing (only) **every road in the U.S.**.

Figure 9.6: Scatter diagram of relative excess route lengths $r(i, j)$ between each pair from the 200 largest U.S. cities. The horizontal scale is distance, normalized so that there is on average one city per unit area. The two lines show unweighted and population-weighted average excess, as a function of normalized distance. Each average is around 18% at all distances.



(*) when you look at the map of the network within a window, the statistical properties of what you see do not depend on the size of the window.

Conceptually, what we are doing here is analogous to the Lecture 7 discussion of modeling English text as a stationary random process. Both English text and roads are consciously planned rather than “random” in the everyday sense, but this does not preclude statistical regularity. And in both cases we do not posit some explicit “toy model” but instead consider the family of models that satisfy a particular requirement.

Defining property (*) more precisely is too technical for this lecture, but various concrete consequences are easier to understand. In a scatter diagram like Figure 9.6 for random addresses in a scale-invariant network, the distribution of $r(i, j)$ -values would not change at all with distance d , and

the averages $\rho(d)$ would be constant:

$$\rho(d) = R, \quad 0 < d < \infty, \quad \text{for some constant } R. \quad (9.6)$$

Although scale-invariance is at best only roughly realistic (for instance it would imply there are arbitrarily small roads leading everywhere), a hypothetical such network has some interesting mathematical properties. For instance, one can quantify where a particular piece of road lies in the major road - minor road spectrum by considering whether it forms part of some long distance route; precisely, say its importance is the largest d such that the given piece of road forms part of the route between some two addresses both at distance $\geq d$ from that given piece. Now consider

$$A(d) = \text{average length-per-unit-area of roads of importance } \geq d.$$

Then scale-invariance implies

$$A(d) = A/d, \quad 0 < d < \infty, \quad \text{for some constant } A \quad (9.7)$$

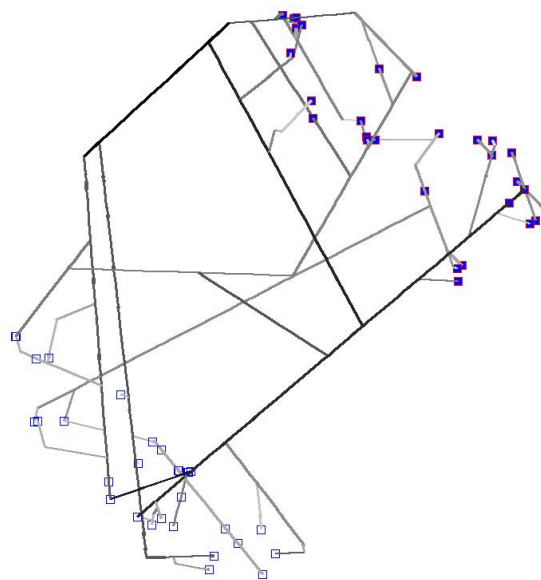
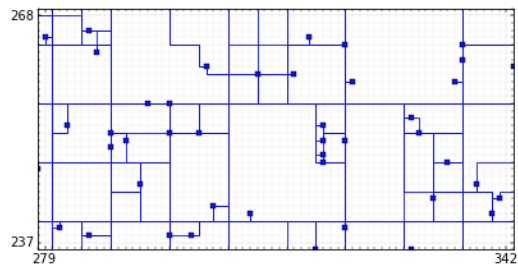
so we automatically have existence of roads over the entire major road - minor road spectrum. Now suppose instead we sample random points of density λ per unit area and consider the length-per-unit-area $\ell(\lambda)$ of the subnetwork consisting of routes between these points. For a scale-invariant network we would have

$$\ell(\lambda) = \lambda^{1/2}\ell, \quad 0 < \lambda < \infty, \quad \text{for some constant } \ell. \quad (9.8)$$

So one could test whether a real-world network is approximately scale-invariant by checking whether the three scaling relations (9.6) - (9.8) are approximately true over appropriate ranges.

Probability models of networks with the scale-invariance property do exist but are difficult to define and then difficult to draw satisfactorily. Figure 9.7 (top) shows the subnetwork of routes between randomly sampled points in a network consisting of north-south and east-west roads, and Figure 9.7 (bottom) shows the subnetwork of routes between pairs (\circ, \bullet) in a network consisting of randomly oriented roads.

Figure 9.7: A spanning subnetwork



Chapter 10

The other lectures

10.1 Psychology of Probability: Predictable Irrationality

There is a huge body of academic research, building on 1970s work of psychologists Daniel Kahneman and Amos Tversky, on how people make decisions under uncertainty, and in particular decisions involving risks and rewards. Nowadays this topic forms part of broader disciplines such as **cognitive science** and **behavioral economics**. Conclusions are mostly based on experimental or survey data, typically obtained from the most convenient source of subjects (undergraduates attending psychology courses) and based on hypothetical questions – *what would you choose if presented with these choices?* In this regard it is quite different from our Lecture 1; here we are prompting people to think about chance, whereas there we were seeking to understand in what contexts people (unprompted) perceive chance as playing a role in their lives.

In teaching the course in 2011 and earlier I sampled topics from many sources, in particular the comprehensive 2004 book *Cognition and Chance: The psychology of probabilistic reasoning* by Raymond Nickerson. That work touches upon many different topics, and gives around 1000 references, so it's an invaluable resource for seeing the big picture of what scholars have thought about, and for leads into the research literature, and for student projects (try this experiment on your friends!). Then the publication of Kahneman's *Thinking, Fast and Slow* showed, unsurprisingly, that he can write about this material infinitely better than I can. So in 2014 I merely cribbed from his book. Chapters 10 and 13–21 are the most relevant to my course, though I encourage students to read the entire book.

This topic is fun to do in class because I can try experiments on my students. In the 2014 course I asked them

(a) It was estimated that in 2013 there were around 1,400 billionaires in the world. Their combined wealth, as a percentage of all the wealth (excluding government assets) in the world, was estimated as roughly

1.5% 4.5% 13.5% 40.5%

(b) I think the chance my answer to (a) is correct is%

The figures are from Piketty's *Capital in the Twenty-First Century* who gives the estimate 1.5%. The student answers were

response	number students	average guess P(correct)
1.5%	5	54%
4.5%	3	37%
13.5%	12	36%
40.5%	14	64%

One can regard this as an instance of *anchoring*, because I placed the correct answer at one extreme of the possible range of answers. It is also a dramatic illustration of *overconfidence* in that the people most confident in their opinion were in fact the least accurate.

10.2 Science fiction meets science

This is a fun lecture to present. I discuss three related topics.

The Fermi Paradox.

The Universe is very big and very old; given there is a human technological civilization on Earth, why don't we see evidence of technologically advanced extraterrestrial civilizations?

Devising possible explanations is an interesting exercise, as organized logic. A top-down organization might start with the alternatives

- they (almost) never arise.
- they don't last long in a form we would recognize, so none are currently close enough to detect.
- they do currently exist but we can't detect them for some reason.

50 more detailed possible explanations are given in the non-technical book *If the Universe Is Teeming with Aliens ... Where Is Everybody?* by Stephen Webb, and these are mostly copied to the [Wikipedia Fermi Paradox page](#). This topic is perhaps the best illustration of the Mark Twain quote

There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment of fact.

The Great Filter. This is a very speculative line of thought, due to Robin Hanson. Consider the product

$$Npq \tag{10.1}$$

where

- N is the number of Earth-like (loosely, and at formation) planets in the galaxy
- p is the chance that, on such a planet, an intelligent species at a technological level comparable to ours will arise at some time
- q is the chance that such a species would survive in such a way as to be observable (via communication or exploration) to other galactic species for an appreciable length of time.

The point is that Npq represents the number of *other* intelligent species we expect to observe in the galaxy. Because we don't observe any, we conclude *prima facie* (treating absence of evidence as evidence of absence) that it cannot be true that $Npq \gg 1$. Since it would be a bizarre coincidence if $Npq \approx 1$, we should conclude that $Npq \ll 1$ and so humans are most likely to be the only technological species in the galaxy.

Now consider the following argument.

Human beings did not create the Universe or direct the course of evolution, so N and p are not our responsibility. But q , as applied to us (i.e. will our species leave its mark on the galaxy?) is presumably under our control. Viewing q very roughly as the chance that a hypothetical technological species arising across the galaxy 25 million years in the future would then be able to observe the then-current or previous existence of humans, being told that $q = 10^{-6}$ would be rather depressing. Depressing, because of the ways this might come about, for instance if humans soon become extinct, or change and cease to interact with

the macroscopic physical universe. Knowing $q = 10^{-6}$ would be knowing that something like this is *almost certain* to happen. Now having decided that Npq is small, implying pq is very small, the only way to avoid the depressing possibility of q being very small is to for p to be very small.

This argument leads to the counter-intuitive conclusion that

$$\text{we should } \textit{want} \textit{ } p \textit{ to be very small} \quad (10.2)$$

where the sense of *want* is, as above, “to be consistent with humanity surviving long enough to have at least a tiny chance of leaving its mark on the galaxy”.

The [Wikipedia Great Filter article](#) is not very helpful; [my own paper on the topic](#) gives a mathematical development, but I do not emphasize that in the lecture.

Global Catastrophic Risks. Thinking about this involves an issue of time-scale. We know the world (as a human technological society) has changed in the last 100 years, and a default is to assume some comparable amount of change in the next 100 years. We can’t imagine 1 million years ahead (which was relevant to the Fermi paradox). So let’s fix on 500 years.

Question: How might it happen that in 500 years there might be no recognizable “human technological civilization”?

The 2008 book *Global Catastrophic Risks* contains 15 chapters analyzing particular risks. In class I ask students to suggest such risks, then show chapter titles, and then discuss a few of their risks.

10.3 Ranking and rating

The 2012 book *Who’s #1?: The Science of Rating and Ranking* by Langville-Meyer describes a range of rating methods based on undergraduate linear algebra. That is not my cup of tea, but the general topic is a natural one for my course. In the 2014 course I asked my grad student Dan Lanoue to give the lecture, and his overview slide was

- Two general rating methodologies: Elo and PageRank,
- Two very specific rating methods for the NFL: DVOA and EPA,

- Compare these methods and what that tells us about the science of ratings.
- I've chosen the NFL as a recurring example I find the most interesting. In particular the NFL is a nice balance between the Moneyball “science” of baseball and the “Wild West” of college football.

I plan to develop a more probability-oriented treatment for the next course.

10.4 Risk to Individuals: Perception and Reality

In this lecture I talk about 4 related topics. First I ask students to rank the seriousness of various specific risks.

Data about risks. The concepts of [micromort](#) and [microlife](#) and numerical estimates. Bad statistical comparisons abound, e.g. [this Economist article](#) – for recreational activities, better to assess risk “per afternoon engaged in activity”. Then I refer to the 2002 Ropeik - Gray book *Risk. A practical guide for deciding what's really safe and what's really dangerous in the world around you*. This has 6-8 page chapters on each of 48 specific risks, written in a fixed format and summarized by “likelihood” and “consequences” of each risk. A natural student project is to seek data and to write a report on another risk in the same style.

Psychological factors which can make a risk seem more threatening or less threatening than it really is. Based on the 2010 Ropeik book *How Risky is it Really?*. Also cite a discussion on Quora under the title [What are some odd-sounding but completely rational ways we might live our lives if we paid attention to the true probability of good or bad things happening to us?](#)

Presentation of statistical risk data to the public , based in part on the online article [2845 ways to spin the risk](#) showing how data on risks “can be “spun” to look bigger or smaller by changing the words used, the way the numbers are expressed, and the particular graphics chosen”.

Economic and public policy aspects of risk , starting from [an article by Trudy Ann Cameron](#) on the unwise choice of phrase *statistical value of life*.

For wrap-up:

Purveyors of risk information usually claim that their rational and scientific assessments help individuals choose safer or more beneficial courses of action. Perhaps so, but in the process they also confront risk consumers with an ever proliferating array of private risk situations. Attempting to mitigate one risk, such as the risk of breast cancer, forces an encounter with a new risk, such as the risk of radiation from a mammogram. The island of safety and security that risk analysis promises to deliver never comes into view because each risk decision only delivers us to ever more numerous and vexing risk assessments still. *Jason Puskar*

10.5 Tipping points and phase transitions

Here I distinguish between

- a *tipping point*, where a system subjected to an external force suddenly changes behavior
- a *phase transition*, where the equilibrium distribution of a system changes qualitatively as a parameter varies.

Usage of the phrase *tipping point* **increased dramatically** after the 2000 publication of Malcolm Gladwell's bestselling book of that name, and it's fun to **search for recent metaphorical usages** to see what writers think it means.

The lecture contains a little math content – the simplest probability models with phase transitions, that is queues and branching processes. But I don't have any substantial interesting data to anchor the lecture.

10.6 Luck

I would like to give a lecture centered on **The Big Question**, which has been asked since time immemorial:

What are the relative contributions of skill and chance to success in different aspects of human life?"

Here are two recent discussions. The main message of Malcolm Gladwell's 2008 bestseller *Outliers* is that the time, place and socio-economic status of

one's birth, the surrounding culture, and luck, rather than pure individual merit, play more of a role in success than we might suppose. And in the 2009 book *Dance with Chance*, Spyros Makridakis et al. write

Hard work, determination, education and experience should count for a great deal [as regards professional success]. But, again the data available suggests that luck is almost entirely responsible for *which* hard working, determined, educated and experienced people make it in life.

But finding actual data, rather than anecdotes, to support these assertions seems too difficult. So instead my lecture meanders around the following topics.

(1) While few readers would admit to “believing in luck” in the superstitious sense, if I ask

do you ever take decisions on the basis (in part) of feeling lucky or unlucky?

then people often admit to doing so, and one can devise psychology experiments to test this.

(2) In a 1997 psychology paper by Peter R. Darke and Jonathan L. Freedman [The Belief in Good Luck Scale](#) the authors are interested in the spectrum between

- the view that luck is a somewhat stable characteristic that consistently favors some people but not others and is especially likely to favor oneself
- the rational view of luck as random and unreliable.

They devise a set of questions to place an individual on that spectrum; so I put these questions to my students.

(3) The 2003 book *The Luck Factor* by psychologist Richard Wiseman is based upon interviews with several hundred people who self-describe as being extremely lucky or unlucky. His conclusion:

- Lucky people create, notice and act upon the chance opportunities in their lives.
- Lucky people make successful decisions by using their intuition and gut feelings.

- Lucky people’s expectations about the future help them fulfill their dreams and ambitions.
- Lucky people are able to transform their bad luck into good fortune

To me, interpreting these consequences in terms of “luck” seems rather arbitrary; one could just say they are consequences of “adopting a positive attitude towards life”.

(4) **Real-world samples and categorizations of luck.** Section 1.3 described some, mostly hypothetical, examples of luck given by the philosopher Nicholas Rescher. A more serious project is to find a source of real examples and devise a useful categorization. It is interesting to contrast the psychologist’s advice above with the “philosophical” advice by Rescher:

- Be realistic in judgements (evaluate the probabilities and utilities as objectively as you can)
- Be realistic in expectations (there is only so much one can do)
- Be prudently adventuresome (don’t be so risk-averse as to lose out on opportunities)
- Be cautiously optimistic.

10.7 Toy models in Population Genetics: some mathematical aspects of evolution

I spend some time during the course giving examples of the use and abuse of toy models, in the following sense.

Some probability models of real-world phenomena are “quantitative” in the sense that we believe the numerical values output by the model will be approximately correct. At the other extreme, a **toy model** is a consciously over-simplified model of some real-world phenomenon that typically attempts to study the effect of only one or two of the factors involved while ignoring many complicating real-world factors. It is thus “qualitative” in the sense that we do not believe that numerical outputs will be accurate.

The topic of this lecture is a classical source of toy models – but what I say is all textbook stuff with no new data.

First I make the conceptual point

10.7. TOY MODELS IN POPULATION GENETICS: SOME MATHEMATICAL ASPECTS OF EVOLUTI

The theory of genetics shows that heredity is not like paint mixing.

Then I introduce the usual simple population genetics models and outline the math arguments for:

For a single mutation giving an allele with small selective advantage α , the chance that the allele becomes fixed is about $\frac{2\alpha}{\sigma^2}$

The duration of a selective sweep $\approx \frac{\log(2N)}{\alpha}$ generations .

Then I talk about the neutral theory and the Ewens sampling formula

the effective number of coexisting neutral alleles at equilibrium is $1 + 4Np$.

This requires discussion of diversity statistics, used also in other lectures (baby names etc). Finally

time back to MRCA in Wright-Fisher

and for fun

“How many of your 10th generation ancestors are you (genetically) related to?”