Reflections on devising a "probability in the real world" course

David Aldous

22 May 2012

Textbook probability examples

Most textbook examples and questions are either "just maths" – X's and Y's – or unrealistic little stories, for example

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?

I claim to be the world expert in . . . . . .

As a hobby I have read and reviewed every (around 100) non-technical book on probability I've found. Many focus on specific topics (sports; stock market). Those that tackle "probability in general" fall into two opposite styles.

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.

Popular science without math, treating historical development and fashionable modern topics.

Representative examples of each genre are
Jeffrey Rosenthal: *Struck by Lightning: the curious world of probabilities*

Leonard Mlodinow: *The Drunkard's Walk: How Randomness Rules Our Lives*.

But . . . neither style works as a college course.

My Berkeley course consists of 20 lectures, on topics chosen to be maximally 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 itself.
- There is some concrete bottom line conclusion, which can be said in words ...
- ... but where mathematics has been used to derive conclusions ...
- ...... and where mathematics leads to some theoretical quantitative prediction that I or **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 the actual lectures fail to attain the ideal!

Here are the topics from 2011, and student feedback "like minus dislike".

(22) Psychology of probability: predictable irrationality
(18) Global economic risks (*)
(17) Everyday perception of chance (*)
(16) Luck
(16) Science fiction meets science
(14) Risk to individuals: perception and reality
(13) Probability and algorithms.
(13) Game theory.
(13) Coincidences and paradoxes. (*)
(11) So what do I do in my own research? (spatial networks)
(10) Stock Market investment, as gambling on a favorable game (*)
(10) Mixing and sorting
(9) Tipping points and phase transitions
(9) Size-biasing, regression effect and dust-to-dust phenomena
(6) Prediction markets, fair games and martingales (*)
(6) Branching processes, advantageous mutations and epidemics
(5) Toy models of social networks (*)
(4) The local uniformity principle
(2) Coding and entropy (*)
(-5) From neutral alleles to diversity statistics.

So that's my manifesto. The rest of this talk will be 8 illustrations of what I actually do in the lectures.

My students are mostly Stat majors – no Math majors. They have taken a first course in math probability and in math statistics. About half of them have taken/are taking/will take our courses in Stochastic Processes and Game Theory (and other Stat courses).

Re math, I try to
(i) quote textbook results
(ii) do back-of envelope style calculations on the board

rather than develop any theory. And the main objective – like a New Year's Resolution that's hard to keep – is to talk only about topics for which we can get actual data.

Illustration 1. I mostly avoid "standard topics" but the **birthday problem** is such a paradigm example .... I emphasize two points.

(a) We can readily get data, e.g. 30 realizations of the form [next slide]

(b) The conclusion is surprisingly robust to assumptions; "23" arises as $0.5 + 1.18/\sqrt{\sum_i p_i^2}$ for $p_i \equiv \frac{1}{365}$. What happens if (hypothetically) $p_i = \frac{4}{3} \times \frac{1}{365}$ for half the days and $= \frac{2}{3} \times \frac{1}{365}$ for the other half?

Answer: "23" becomes "22".

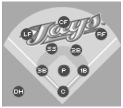
(c) Can we find data to test the non-uniform prediction?
Student project. Canadian professional ice hockey players!

Blue Jays Roster
Active Roster
40-Man Roster
Depth Chart
Search

Coaches
Transactions
Injuries
Draft Results
Front Office
Broadcasters
Minor League Affiliates

Player Search
Enter Last Name Go

Select a Team



Go to a player's page by moving the cursor over a position and clicking on a player's name.

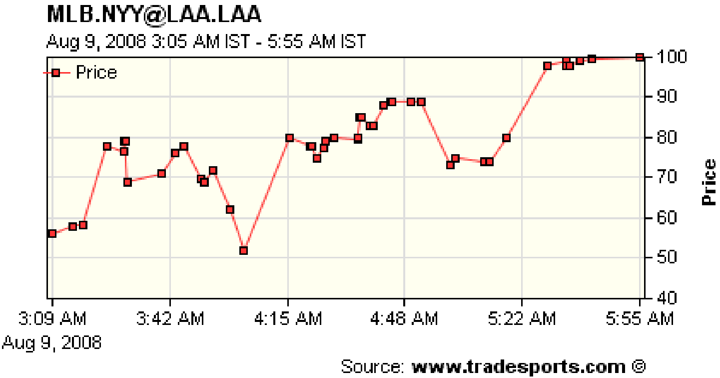
Active Roster

Pitchers	B/T	Ht	Wt	DOB
49 Jeremy Accardo	R/R	6-2	190	12/08/81
34 A.J. Burnett	R/R	6-4	230	01/03/77
37 Scott Downs	L/L	6-2	190	03/17/76
54 Jason Frasor	R/R	5-10	170	08/09/77
32 Roy Halladay	R/R	6-6	225	05/14/77
22 Brandon League	R/R	6-3	190	03/16/83
51 Jesse Litsch	R/R	6-1	195	03/09/85
28 Shaun Marcum	R/R	6-0	185	12/14/81
29 Dustin McGowan	R/R	6-3	220	03/24/82
56 Brian Tallet	L/L	6-7	220	09/21/77
59 Randy Wells	R/R	6-3	230	08/28/82
40 Brian Wolfe	R/R	6-3	220	11/29/80
Catchers	**B/T**	**Ht**	**Wt**	**DOB**
20 Rod Barajas	R/R	6-2	230	09/05/75
9 Gregg Zaun	S/R	5-10	190	04/14/71
Infielders	**B/T**	**Ht**	**Wt**	**DOB**
11 David Eckstein	R/R	5-7	175	01/20/75
2 Aaron Hill	R/R	5-11	195	03/21/82
6 John McDonald	R/R	5-11	185	09/24/74
17 Lyle Overbay	L/L	6-2	235	01/28/77
19 Marco Scutaro	R/R	5-10	185	10/30/75
Outfielders	**B/T**	**Ht**	**Wt**	**DOB**
21 Buck Coats	L/R	6-3	195	06/09/82
15 Alex Rios	R/R	6-5	195	02/18/81
24 Matt Stairs	L/R	5-9	220	02/27/68
23 Shannon Stewart	R/R	5-11	210	02/25/74
10 Vernon Wells	R/R	6-1	225	12/08/78
Designated Hitters	**B/T**	**Ht**	**Wt**	**DOB**
35 Frank Thomas	R/R	6-5	275	05/27/68

Official Info Terms of Use Privacy Statement Help/Contact Us Job Opportunities Corporate Partners

Illustration 2: Martingales and prediction markets.

In textbooks, martingales are a "non-elementary" topic, treated as math without any data.
I want to start with data.



In 17 baseball games from 2008 for which we have the prediction market prices as in Figure 1, and for which the initial price was around 50%, the prices (as percentages) halfway through the match were as follows.

$$07, 08, 26, 26, 32, 33, 43, 43, 48, 54, 66, 73, 76, 80, 85, 86, 96$$

The average of these 17 numbers is fortuitously very close to 50, which (by the martingale property) is the theoretical expectation of halftime price.

These figures look roughly uniform on $[0, 100]$ – is there an explanation?

Consider the point differences Z_1 and Z_2 in the first half and the second half. A fairly realistic model is to assume
(i) Z_1 and Z_2 are independent, with the same distribution function F;
(ii) their distribution is symmetric about zero.
For simplicity we add an unrealistic assumption
(iii) the distribution is continuous.
Under these assumptions, the chance the specified team wins, given the first half point difference is z, is

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

and therefore the conditional probability of winning at halftime is

$$P(Z_1 + Z_2 > 0 | Z_1) = F(Z_1).$$

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

Romney for Republican presidential nominee.

Palin for Republican presidential nominee.

Widely believed that the number of candidates whose prospects have risen and fallen is unusually large. Is it really?

Theory: If each candidate's initial price (% chance) is $< x$ then the mean number of candidates whose price ever exceeds x equals $100/x$.

Data: Maximum price (so far) for each candidate.
Romney 98
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
Conclusion: Only slightly more than expected.

actual bet

Illustration 3: Paradoxes, helpful and unhelpful.
Many of the standard paradoxes, e.g. the **two envelopes problem**, strike me as too unrealistic to be taken seriously. Conversely, there are many "paradoxical" phenomena that are both interesting and have real-world significance. For instance

A bet that is advantageous to both parties. In February 2012, I (resident in the U.S.) and my friend Sir Jonathan (resident in the U.K.) agreed that there is a 50-50 chance that, one year ahead in February 2013, the pound/dollar exchange rate will be on either side of 1.60 dollars per pound. (The figure didn't matter, just that we agreed on the figure). We then made a bet:

If 1 pound is at that time worth more than 1.60 dollars then he will pay me 1 pound; if not then I will pay him 1.60 dollars.

So what will happen?

- From his viewpoint he either loses 1 pound or wins an amount that by definition is worth more than 1 pound;
- from my viewpoint I either lose 1.60 dollars or win an amount that by definition is worth more than 1.60 dollars;
- and since we agreed it was a 50-50 chance, to each of us this is a favorable bet.

Illustration 4: The Kelly criterion marks the borderline between aggressive and insane investing.

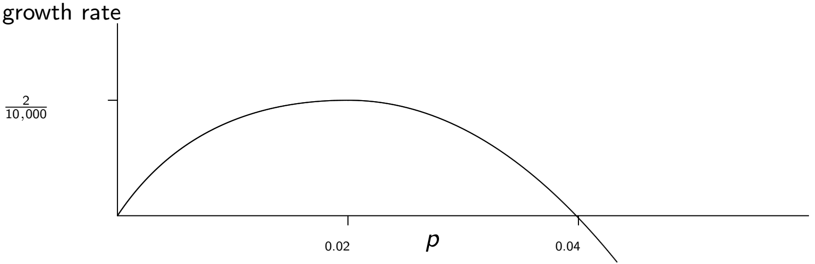
I focus on **long-term** investment. Imagine you inherit a sum of money at age 25 and you resolve to invest it and not start spending it until age 65. We envisage the following setting.
(i) You always have a choice between a safe investment (pays interest, no risk) and a variety of risky investments. You know the probabilities of the outcomes of these investments. [of course in reality you don't know probabilities – unlike casino games – so have to use your best guess instead].
(ii) Fixed time period – imagine a year, could be month or a day – at end you take your gains/losses and start again with whatever you've got at that time ("rebalancing").

The Kelly criterion gives you an explicit rule for how to divide your investments to maximize long-term growth rate.

To illustrate, imagine day-trading scheme with stocks based on some statistical non-randomness; within one day

51% chance to double money;   49% chance to lose all money.

Looks good – expected gain 2% per day – but don't want to risk all your money on one day. Instead use strategy: bet fixed proportion p of money each day. Theory says: long-term growth rate, depends on p, but in an unexpected way.



Optimal strategy: bet $p = 2\%$ of your capital each day; this provides growth rate $\frac{2}{10{,}000}$ per day, which (250 trading days per year) becomes 5% per year.

The numbers above depended on hypothetical assumptions. But the conceptual point is completely general. We are not assuming you can predict the future, just that you can assess future **probabilities** correctly. Provided there is **some** risky investment whose expected payoff is greater than the risk-free payoff, the **Kelly criterion** is a formula that tells you how to divide your portfolio between different investments.

There's one remarkable consequence of using this strategy. To get the maximum possible long-term growth rate, using "100% Kelly strategy", you must accept a short-term risk, of the specific form

50% chance that at some time your wealth will drop to only 50% of your initial wealth.

And 10% – 10% too! Of course, if not comfortable with this level of risk, you can use "partial Kelly strategy" combining with risk-free assets.

This story is told in the popular book **Fortune's Formula** by William Poundstone. Maybe nothing in this story seems intellectually remarkable, but in fact something is. Consider an analogy: the light speed barrier.

[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 low-term reward. Common sense says this risk-reward trade-off spectrum continues forever. But in fact 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".

Some data . . . . . .

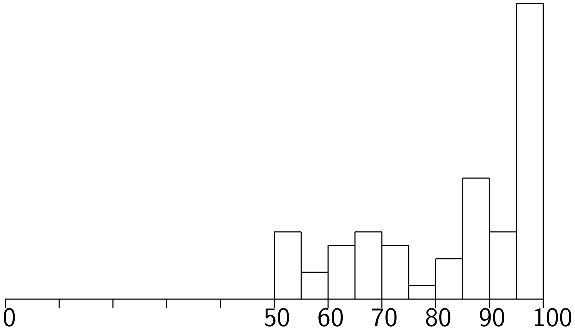


Figure 2. Historical distribution of the minimum future value of a 100 investment. Based on hypothetical purchase of S&P500 index on first day of each year 1950-2009 and on subsequent monthly closing data.

This is hardly uniform on $[0, 100]$.

IFA plot. Note 50 years.

Illustration 5: How do people think about chance in everyday life?

There are many ways one might study that question, for example by searching blogs to examine casual usage of specific words or phrases. I will show results of examining a sample of queries submitted to the search engine Bing containing the phrase "chance of" or "probability of". We manually examined about 1,000 queries, retained those where the user was seeking to discover the chance of something, and sorted these 675 retained queries into 66 groups (each containing about 10 queries) of queries on some similar topic. I then chose one representative query from each of these groups. All 66 are on my web page – below I show 30 of them, to indicate the range and frequency of topics that occur in such searches.

Can you guess which topics appear most often?

Do you think they will have much connection with typical textbook topics?

Query: chance of pregnancy on pill
Query: how to improve chance of getting pregnant
Query: chance of getting pregnant at age 41
Query: chance of getting pregnant while breastfeeding

Query: can you increase your chance of having a girl
Query: if twins run in my family what's my chance of having them?

Query: does a father having diabetes mean his children have a 50% chance of getting diabetes
Query: chance of siblings both having autism

Query: chance of miscarriage after seeing good fetal movement heartbeat at 10 weeks
Query: chance of bleeding with placenta previa
Query: any chance of vaginal delivery if first birth was ceaserian

Query: probability of having an adverse reaction to amoxicillin
Query: does hypothyroid in women increase chance of liver cancer?
Query: does progesterone increase chance of breast cancer
Query: which treatment has the least chance of prostate cancer recurring?
Query: what is the chance of relapse in a low risk acute lymphoblastic lukemia patient
Query: chance of getting a brain tumor

Query: probability of flopping a set with pocket pair in poker
Query: does a ring of wealth affect the chance of the dragon pickaxe drop in runescape?

Query: chance of surviving severe head injury

Query: chance of snow in Austin Texas
Query: is there chance of flood in Saint Charles, Illinois today?

Query: calculate my chance of getting to university of washington
Query: what are the chance of becoming a golf professional

Query: chance of closing airports in Mexico because of swine flu

Query: any chance of incentive packages for government employees to retire

Query: chance of children of divorce being divorced
Query: chance of food spoiling if left out over night

Query: what does it mean 50/50 chance of living 5 years
Query: probability of life and evolution

A quick rant

90 years of bad teaching of freshman Statistics, and shorthand like "the data is statistically significant", has led to the widespread blunder of thinking that statistical significance is an attribute of observed data; of course it's really an attribute of the hypothetical probability model that may or may not be generating the data.

I suspect this all started with the bad choice of using the same words (mean, s.d., etc) for data and for models.

Illustration 6; from the lecture on entropy and data compression.

For any probability distribution $\mathbf{p} = (p_s) = (p(s))$ on any finite set S, its entropy is the number

$$\mathrm{ent}(\mathbf{p}) = -\sum_s p_s \log_2 p_s .$$

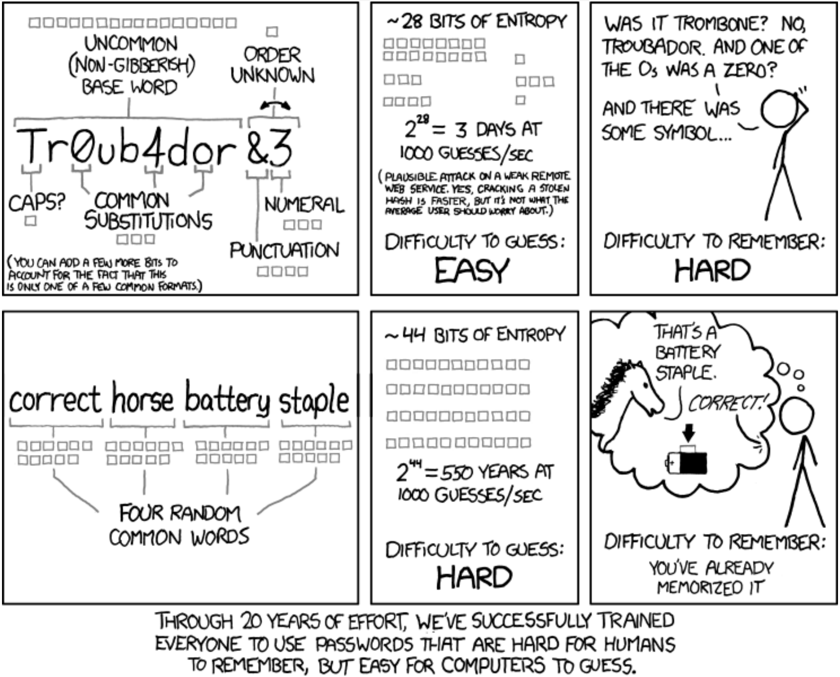
Suppose, from a library of about one million books, you pick one book uniformly at random and write out the entire text of that book. What is the entropy of what you write?

Answer [according to our definition]: 20 bits.

On the other hand, widely claimed "entropy of English language is about 1.3 bits per letter", so the text of a book with 300,000 letters should have entropy about 400,000 bits.

Point (again): there is a distinction between actual data and hypothetical probability models by which the data might have been generated.

Here's a less artificial example. From the Hall of Fame for back-of-an-envelope calculations.


UNCOMMON (NON-GIBBERISH) BASE WORD
ORDER UNKNOWN
Tr0ub4dor &3
CAPS?
COMMON SUBSTITUTIONS
NUMERAL
PUNCTUATION
(YOU CAN ADD A FEW MORE BITS TO ACCOUNT FOR THE FACT THAT THIS IS ONLY ONE OF A FEW COMMON FORMATS.)
~28 BITS OF ENTROPY
2^{28} = 3 DAYS AT 1000 GUESSES/SEC
(PLAUSIBLE ATTACK ON A WEAK REMOTE WEB SERVICE. YES, CRACKING A STOLEN HASH IS FASTER, BUT IT'S NOT WHAT THE AVERAGE USER SHOULD WORRY ABOUT.)
DIFFICULTY TO GUESS: EASY
WAS IT TROMBONE? NO, TROUBADOR. AND ONE OF THE Os WAS A ZERO?
AND THERE WAS SOME SYMBOL...
DIFFICULTY TO REMEMBER: HARD
correct horse battery staple
FOUR RANDOM COMMON WORDS
~44 BITS OF ENTROPY
2^{44} = 550 YEARS AT 1000 GUESSES/SEC
DIFFICULTY TO GUESS: HARD
THAT'S A BATTERY STAPLE.
CORRECT!
DIFFICULTY TO REMEMBER: YOU'VE ALREADY MEMORIZED IT
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.

password checker

Illustration 7: Toy models and bad science.

II.—*A Mathematical Theory of Evolution, based on the Conclusions of Dr. J. C. Willis, F.R.S.*

By G. Udny Yule, *C.B.E., M.A., F.R.S., University Lecturer in Statistics, and Fellow of St. John's College, Cambridge.*

(Received June 11, 1923.—Read February 7, 1924.)

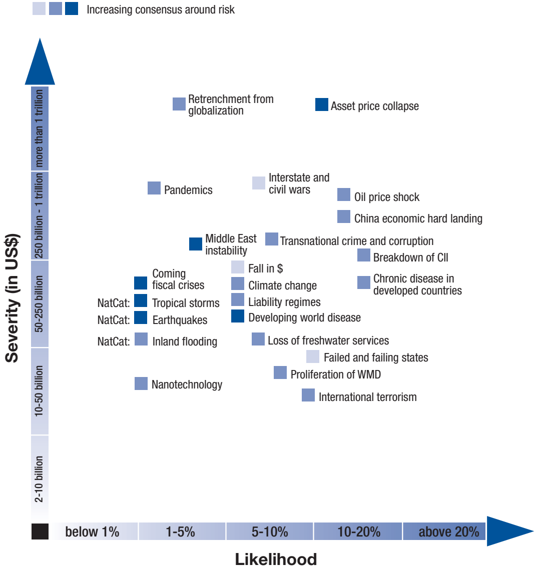
Contents.

I regard it as self-evident that interesting aspects of the future are uncertain and therefore that forecasts, predictions, risks etc should be expressed in probability terms. Of course I don't have any magic way of doing this, but we can study how accurate other people's predictions have been.

Illustration 8: Accuracy of assessments of global economic risks

A cynical view of retrospective analysis of the late-2000s worldwide financial crisis 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?

Each year since 2006 the OECD has produced a "global risks" report for the World Economic Forum annual meeting in Davos. The 2008 report, written in mid-2007 (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) gave, as in other years, a list of '"core global risks", summarized using the next graphic. The horizontal axis shows "likelihood" and the vertical axis shows economic effect.


Increasing consensus around risk
Severity (in US$)
more than 1 trillion
250 billion - 1 trillion
50-250 billion
10-50 billion
2-10 billion
Retrenchment from globalization
Asset price collapse
Pandemics
Interstate and civil wars
Oil price shock
China economic hard landing
Middle East instability
Transnational crime and corruption
Breakdown of CII
Coming fiscal crises
Fall in $
Climate change
Chronic disease in developed countries
NatCat: Tropical storms
Liability regimes
NatCat: Earthquakes
Developing world disease
NatCat: Inland flooding
Loss of freshwater services
Failed and failing states
Proliferation of WMD
Nanotechnology
International terrorism
below 1%
1-5%
5-10%
10-20%
above 20%
Likelihood

Defininíng risk as "likelihood" multiplied by "economic effect", the 5 most serious risks (as assessed in 2008) were

Asset price collapse
Oil price shock
China economic hard landing
Inter-state and civil wars
Breakdown of civil informational infrastructure

The entry "asset price collapse" was defined as

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

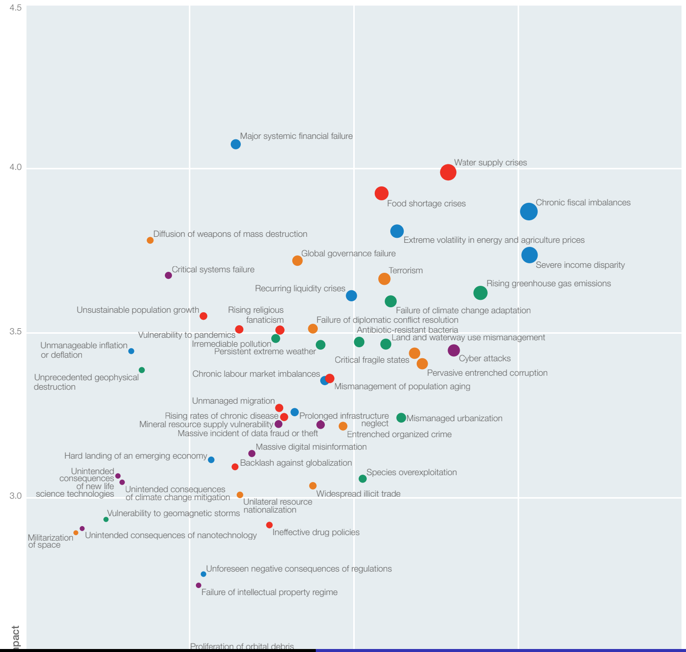
Given that these 5 risks were assessed to have 10-20% likelihood and that one of them occured with even more than predicted severity, this OECD assessment is actually as good as one could hope for. Note that the "oil price shock" assessed as 2nd most serious did almost occur in 2008 but was overtaken by the asset price collapse and did not have the severe impact predicted – next graphic.


Crude oil, Brent price chart
133.9
110.8
87.7
64.7
41.6
Sep-06
Mar-07
Sep-07
Mar-08
Sep-08
Mar-09
Sep-09
Mar-10
Sep-10
Mar-11
Sep-11
($/bbl)

What's my point? Interesting aspects of the future are uncertain, so whatever forecasting/prediction you do, say it in probability terms.

With this in mind let us look at the corresponding graphic from the 2012 report. Now the most serious risks (for the next 10 years) are assessed as

Chronic fiscal imbalances
Water supply crises
Severe income disparity
Food shortage crises
Extreme volatility in energy and agricultural prices
Rising greenhouse gas emissions



4.5
4.0
3.5
3.0
Major systemic financial failure
Water supply crises
Food shortage crises
Chronic fiscal imbalances
Diffusion of weapons of mass destruction
Extreme volatility in energy and agriculture prices
Global governance failure
Severe income disparity
Critical systems failure
Terrorism
Recurring liquidity crises
Rising greenhouse gas emissions
Unsustainable population growth
Rising religious fanaticism
Failure of climate change adaptation
Failure of diplomatic conflict resolution
Vulnerability to pandemics
Antibiotic-resistant bacteria
Unmanageable inflation or deflation
Irremediable pollution
Land and waterway use mismanagement
Persistent extreme weather
Critical fragile states
Cyber attacks
Unprecedented geophysical destruction
Pervasive entrenched corruption
Chronic labour market imbalances
Mismanagement of population aging
Unmanaged migration
Rising rates of chronic disease
Prolonged infrastructure neglect
Mismanaged urbanization
Mineral resource supply vulnerability
Massive incident of data fraud or theft
Entrenched organized crime
Massive digital misinformation
Hard landing of an emerging economy
Backlash against globalization
Unintended consequences of new life science technologies
Species overexploitation
Unintended consequences of climate change mitigation
Widespread illicit trade
Unilateral resource nationalization
Vulnerability to geomagnetic storms
Ineffective drug policies
Militarization of space
Unintended consequences of nanotechnology
Unforeseen negative consequences of regulations
Failure of intellectual property regime
Proliferation of orbital debris

misc lists ......

- List of student projects in 2011.
- list of 15 actually verifiable predictions of theory
- Wikipedia list of events that did not happen in 2011