Lecture 5: Short and Medium term predictions and risks in politics and economics

David Aldous

February 3, 2016

How good are people at predictions for short- and medium-term politics and economics?

Here we are not thinking of "routine" issues – predicting election results from opinion polls, or predicting macroeconomic indicators 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, few people imagined that Russia would annex Crimea, or that Scotland would almost become independent, or the emergence of an entity like ISIL. Where is the line between predictable and unpredictable, and what do these words actually mean?

Of course there is no magic crystal ball that will tell you actual predictions. This lecture's focus is on how to assess how good other people's past predictions have been, and we look at

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

Some conceptual issues.

It is often said that "nobody predicted the peaceful ending of Soviet control of Eastern Europe (1989) and subsequent breakup of the Soviet Union (1991)". But what exactly does that mean?

Nice illustration of the difficulties of searching for pre-internet material. A quick search finds a Wikipedia page *Predictions of the dissolution of the Soviet Union* asserting than many such predictions were made. But these are of the style "it's a bad system that can't last forever" rather than any testable prediction.

A scholarly analysis of literature in the International Relations discipline was given in 1993 by Gaddis (*International relations theory and the end of the Cold War*). What's relevant to this lecture is their 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).*

This "unlikely events don't happen" attitude strikes me as very strange. To me it's self-evident that, in such cases, instead of saying "this will or will not happen" one should think of alternative outcomes and assign probabilities.

I happen to have a book (Dunnigan – Bay *A Quick and Dirty Guide to War, 1st edition*) published in 1985 that 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 assessment for 1985-1995 was

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

and their phrase "the empire crumbles" for the latter was rather accurate.

I believe that anyone else who, seriously considered possibilities in 1985 would also assign some small probability to "the empire crumbles". (**small project:** is this correct?)

Reading the actual history of the Soviet Union over 1985-91, my view (unprovable, of course) is that the outcome actually was unlikely.

Unlikely events do sometimes happen!

Course project: look at some similar source of past forecasts and judge how accurate they were. For instance, the 2008 edition of Dunnigan – Bay *A Quick and Dirty Guide to War, 4th edition*.

The Good Judgment Project

[show page; demo making prediction]

Course project: track some questions like these.

How to score a prediction tournament

Consider for a moment a scenario where two people, A and B, are asked to *predict* (as Yes/No) 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 prob.	B's forecast prob.	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?

One's first reaction might be

> *it's impossible to score a prediction tournament because we don't know the true probabilities.*

This assertion might be made by analogy with assessing the quality of a movie – we can't say which movie reviewer is best at assessing the "true quality" of movies because we have no standard of "true quality". Then a second reaction might be

> *over a long series of predictions by an individual, look at those where the predicted probability was around (say) 60%; and see whether around 60% of those events actually happened.*

If this happens, the individual is called *calibrated*. Being calibrated is clearly desirable, but it's not sufficient because one can "cheat" to attain calibration.

[board]

Digression: analogy with golf.

Recall that in golf, each hole has a "par", the score (number of shots) that an expert is reckoned to need. Imagine you are a non-expert golf player, participating in a tournament with other non-experts, on a new golf course with eccentric design (some holes are very easy, some are very hard) on which no-one has played before, so there is no known par. Suppose your score, over the 18 hole course, is 82. Is this good? In absolute terms, you have no way of knowing – there is no "par" for the course with which to compare your score. But a relative comparison is easy – you are better, or maybe just luckier, than someone who scores 86.

Our scoring system for prediction tournaments will have the same property – we can tell who is relatively better, but not how good they are in absolute terms. Also, like golf, we are trying a get a **low** score.

The Good Judgment project is an instance of a *prediction tournament*, where contestants make forecasts for a series of future events. To analyze the resulting 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 **squared error:** is

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

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

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

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

$$(p-q)^2 + q(1-q).$$

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, (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 $\mathrm{MSE}(A)$ and $\mathrm{MSE}(B)$, 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 (2) implies that

$$s_A - s_B \text{ is a sample estimate of } \mathrm{MSE}(A) - \mathrm{MSE}(B) \tag{3}$$

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 $\mathrm{MSE}(A) - \mathrm{MSE}(B)$ 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.

This scoring rule is used for mathematical convenience, but the fact that we are measuring the "cost" of an incorrect probability forecast as the squared error $(p - q)^2$ is consistent with a calculation [later in lecture] that, in a very simple "decision under uncertainty" model, the cost scales as order $(p - q)^2$.

In the actual Good Judgment project, there an extra issues concerning scoring – an individual can (and should) update predictions. Here are the stated scoring rules (actually from earlier GJP).

1. When a question closes, we compute your Brier (MSE) score for a given question by calculating your Brier score separately for each day and then averaging all scores.
2. Every day counts equally towards your score. After you make your first forecast, each subsequent day we count you as making the same forecast, and therefore you get the same Brier score for each subsequent day, until you update your forecast. This means that you should not wait until the last minute to make a forecast.
3. If you don't make a forecast on the first day a question is open, then you are assigned the average Brier score of the group.
4. Your overall Brier score across all of the questions on which you have made forecasts is the average of your overall Brier scores for each of those questions.

Digression: A mistake in the scoring rules in the earlier GJP.

If implemented precisely as stated above, there was a "mistake" in the rules in the original GJP, in the sense that one can sometimes "game the system" to get a better (mean) score by making a dishonest (not your true assessed probability) forecast. Fortunately the mistake is easily fixable. In brief, the trick is

> *Your forecast should overstate your assessed probability, in the early stages of question duration.*

The trick depends on the fact that, in a typical question, the questions"closes" if the event happens, but continues until the deadline if not; there is an asymmetry between "event happens" and "event does not happen" which is not handled correctly by the scoring rules.

We illustrate with an artificially simple example.

Suppose the question is open for duration $2T$, and your initial probability assessment is as follows.

> *With chance* $1/2$ *the event will happen at time* T*. Otherwise it will not happen.*

Suppose also you believe that, if the event is going to happen before the deadline, then it will be a "surprise", in the sense of no advance warning. So your honest forecast would be $p = 1/2$ before time T, and then (if the event did not happen at T) $p = 0$ until time $2T$. Now let's do the calculation supposing you forecast a dishonest value p before time T (but still forecast $p = 0$ subsequently).

• If the event happens at T, the question "closes" at T, and your average (over days) score is the score on each day: $(1 - p)^2$.

• If not, your score on days before T is p^2, and your score on days after T is 0, so your average score is $p^2/2$

Combining these possibilities, your mean score equals

$$\tfrac{1}{2}((1 - p)^2 + p^2/2)$$

and this is maximized by $p = 2/3$.

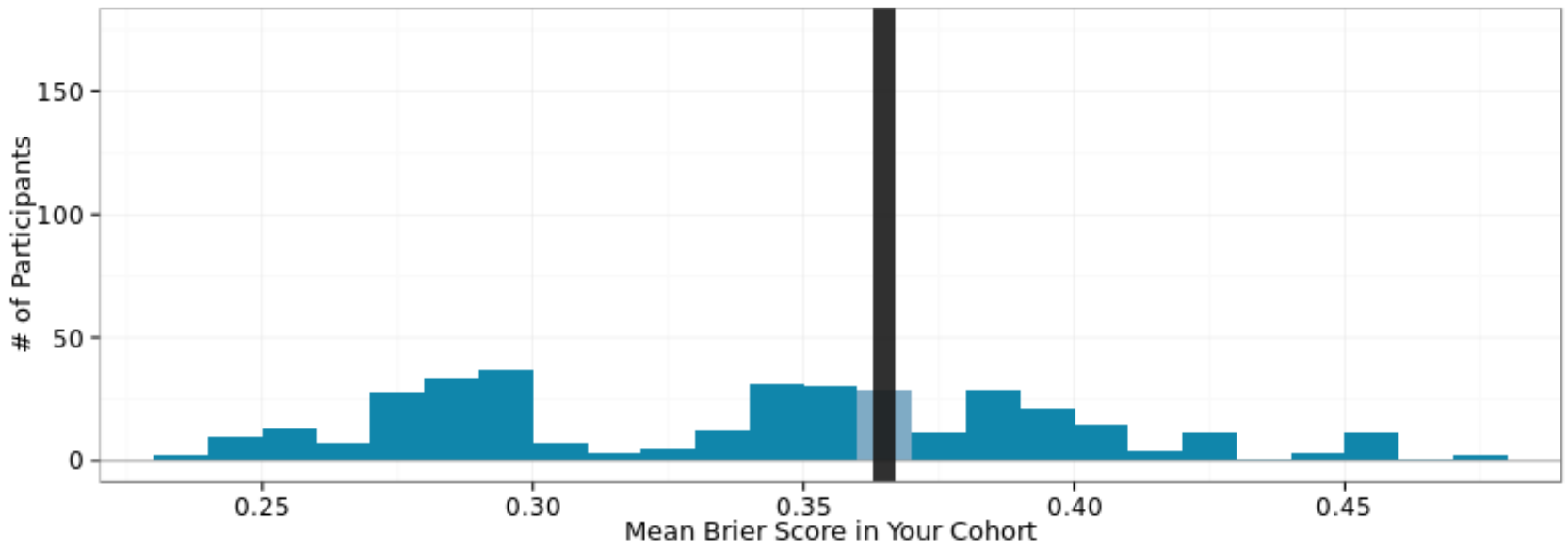
One can fix this mistake by not closing the question until the deadline (time $2T$ in our example) and imputing the forecast value $p = 1$ for the days after the event happens (if it happens before the deadline). In our example, with this modified rule, if the event happens at time T, your score would be $((1-p)^2 + 0)/2$, and then your overall mean score becomes

$$\tfrac{1}{4}((1-p)^2 + p^2)$$

which is indeed maximized by $p = \frac{1}{2}$.

A bottom line?

1. Some people really are better than others. Over 168 questions, the SE of each score in the figure is about 0.02.



2. It might be that the best people are able to assess the true probabilities. Or not. There's no way to distinguish.

The Good Judgment Project has roots in 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 expected . . . . . *the Spanish Inquisition*.
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.

Course project: (rather vague). Look at the technical statistics part of Tetlock's book, look at critiques of his work by others, look as subsequent academic literature. For instance Mandel - Barnes *Accuracy of forecasts in strategic intelligence* claim that Canadian experts are better than U.S. experts.

Note there is a 2015 follow-up book which is less technical: *Superforecasting*.

The annual World Economic Forum Global Risks Survey

[show WEF]

My link goes to the 100-page report; I will talk about the key graphic.

[show 2016 graphic]
[note axes are likelihood vs economic impact]

Predictions are for "next 10 years". We can look at past years and see (partially) how the predictions worked out. I will show the 2011 and 2007 reports.

We are all aware of what Wikipedia calls the *Financial crisis of 2007 - 08* and the subsequent *2008 - 2012 global economic recession*. It is often said (as with collapse of Soviet Union) that no-one predicted this. Here is the 2007 report, written in late 2006, before any widely-recognized signs of trouble.

[show 2007 graphic]

The entry "asset price collapse", defined via

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

appears as the 5th most likely of the 23 risks, but quantifying risk as likelihood times severity it is assessed as the greatest of these risks. So this assessment is actually as good as one could hope for.

As an aside, the "oil price shock" assessed as 4th most likely did almost occur in 2007-8 but was overtaken by the asset price collapse and did not have the severe impact predicted – see chart below.

[show chart]

In this Global Risks Survey the types of risk (as well as probability and size of economic effect) are very vague, so we can't really give a numerical "score", as in the GJP, for how accurate they were. However

Course project: look at the 2011 chart, and give a rough assessment of the subsequent economic effects of some of the identified risks.

Course project: compare the risks identified in the 2011 chart with the extent of media coverage (as end-2010) of future global risks.

[show 2011 chart]

Course project: look at the 2016 chart, and think of some risks they did not consider, and analyze in the style of the Global Risks Survey.

Other exercises in thinking about the future, on my "papers" page, are

- *Future Global Shocks*
- *Global Trends 2030: Alternative Worlds*
- *Shifting Gear: policy challenges for the next 50 years*.

As we look into the future, where does rationality end and science fiction begin?

Math digression: The cost of errors in assessing probabilities

What is the cost of an error in assessing a probability? This is a very vague question, and clearly the answer is very context-dependent. For instance in the context of betting against a human opponent at odds to be negotiated, having a less accurate estimate of the true probability than does your opponent is liable to be very costly. We consider instead 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 are in the classical setting of linear utility, not 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 $p_{\rm guess}$ and p are on the same side of $p_{\rm crit}$ 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

$$|p - p_{\rm crit}|z \text{ where } z = a - b - c + d > 0. \tag{4}$$

Let me outline an argument for 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 $p_{\rm crit}$ is independent of p and $p_{\rm guess}$. Then the proportion of times that $p_{\rm crit}$ happens to be in the interval between p and $p_{\rm guess}$ should be of order $|p - p_{\rm guess}|$, assuming the latter is small; and when this occurs the mean cost is also, by (4), of order $|p - p_{\rm guess}|$.

So in this particular "decision under uncertainty" context the cost of errors is order $(p - p_{\rm guess})^2$.

Finally, for fun I have a link to some future predictions from 1993. These were deliberately intended to be provocative – unlikely extreme changes – but a few have actually almost happened.