Slow Thinking & Deep Learning; Tversky & Kahneman's Cabs.

Post-publication note by author:

I regret being unable to send you a copy of the final version of my article as published on-line on 11/06/15; this is for copyright restrictions which a number of supportive members of the statistics community are currently challenging.

However a positive outcome of the dispute is that I can presume myself at liberty to send to anyone I choose my original version below. This was submitted to Teaching Statistics *on 18/11/14, and was considerably longer, mainly because the Appendix was censored before publishing version. I speculate, but was not told explicitly told, that this because it is logically incompatible with Taleb's exposure of the 'ludic fallacy' that I refer to on my p5 below.*

More positive still, it provides me with the opportunity to share with others an appeal for help: 'what can we teach about Bayes?'. The fundamental difficulty is finding examples where the input probabilities are independent; this axiomatically is not the case where one of them is subjective, as in the Cab-problem. At the back of my mind are examples of two apparently promising areas:-

1) *Where the probabilities are provided by objective measures. For example, locating the best position for a new underwater oil rig might be obtained from data on (a) distance from the nearest successful extant rig and (b) composition of a sample from the seabed.*
2) *Where information is flowing in too fast for the human mind to process it, e.g. in fighter or drone control.*

My search on Google has revealed nothing I see useful for teaching purposes. But I'd be delighted to hear from any colleague who knows better, and might be interested in writing a joint article on this vitally important topic.

Mike Bedwell
Ukrainain Education Center, Kyiv

Michael_Bedwell@hotmail.com

Key words: Deep Learning, Slow thinking, Cab Problem, Bayesianism

Abstract

This note describes my experience in encouraging the 'deep learning' that has long been advocated in the pedagogic literature. My students have typically studied statistics only as part of another discipline, such as economics, business or law. So while they have mostly been aged 17 or over, my approach has necessarily presumed only elementary mathematics and thus should be adaptable to younger people.

My main example is a problem story constructed by Amos Tversky in the 1970s to evaluate human beings' intuitions about statistical inference, and which in 2012 was revisited in a best-seller by his colleague, the Nobel prize-winner Daniel Kahneman. In his book he describes this problem as 'standard' and .unequivocally answers with a simple fixed-point number. I describe how I have encouraged my students to challenge the certainty of this assertion by identifying ambiguities unexplained in the story; in the process I strive to stimulate individuals' *Thinking, Fast and Slow,* to use the title of Kahneman's book, arguing that his 'slow thinking' is a prerequisite of deep learning.

While Kahneman more fully describes the problem as one of ' Bayesian inference', his story can be de-constructed without reference to the work of Thomas Bayes. However, the bitterest

conflicts in the statistical academic community continue to arise from the Bayes-frequentist controversy; this we cannot expect our students to resolve, but we owe it them to explain its causes. So my article includes as an appendix a 'Bayes Icebreaker' where I show an analogy between the cab story and an exercise previously described in *Teaching Statistics*.

Introduction.

Much of my recent teaching has been to a 17+ age group of Ukrainian students, eager to understand western education. They often want to to take direct advantage of it, their first step being to prepare for such computer-implemented tests as the TOEFL, GRE and GMAT. The principle of 'deep learning' is often novel to these young people; even though born well after the break-up of the USSR, they have usually been taught in schools and universities where the culture has remained one of rote-learning, corruption and nepotism.

The internet has progressively become my major tool in helping these students; the others have included two best-selling paperbacks that are cheap to buy but which have proved priceless in enlightenment, to me as much as my students. For these two books have provoked me to ruminate on my own *Thinking, Fast and Slow*, the title of one of them (Kahneman, 2012) . One clear conclusion I have come to is that deep learning can and should be inculcated in much younger people, and that as teachers of statistics we are well placed to take the lead.. I will argue that the level of mathematics needed is within the grasp of even primary school children anywhere in the world.

Discussion of the Pedagogy

As might be crudely explained, the 'slow thinking' of Kahneman's title corresponds roughly to the popular adages 'count to ten before replying' or 'engage brain before opening mouth', and I have interpreted it as a prerequisite for the 'deep learning' that the pedagogic literature has long advocated. Below is a problem from Kahneman's book which I have regularly posed to my students to stimulate such thinking, and which he and his co-worker Amos Tversky had developed in their research some years earlier:

'A cab was involved in a hit-and-run accident at night.
Two Cab companies, the Green and the Blue, operate in the city.
You are given the following data:

- 85% of the cabs in the city are Green and 15% are Blue.
- A witness identified the cab as Blue. The court tested the reliability of the witness under the circumstances that existed on the night of the accident and concluded that the witness had correctly identified each one of the two colors 80% of the time and failed 20% of the time.

What is the probability that the cab involved in the accident was Blue rather than Green?
'

(Tversky and Kahneman, 1980, in Kahneman, 2012, pp166, 167)

Whenever possible, it has been my practice to have this problem printed out, double-spaced and with wide margins to allow ample room for annotations. For younger students I would imagine it necessary to have some discussion of the meaning of' 'percent' and 'probability', although in the UK, most children will have acquired some familiarity with these words; only too early in their lives they are exposed both to TV lottery shows in the home and to state-imposed, statistically-processed testing at school.

With my maturer students, I immediately get them into groups and then walk round amongst them, eliciting and answering questions. I encourage those engrossed in any electronic device to find the problem on Google, especially when I suspect them to be indulging in some sort of private communication. Elsewhere (Bedwell, 2009) I have argued that technology in the classroom serves only to disrupt the all-too-precious time we can spend in face-to-face discussion with our students. I expect mine to behave accordingly.

Discussion of the Problem

My approach is to put to each group the following questions:-

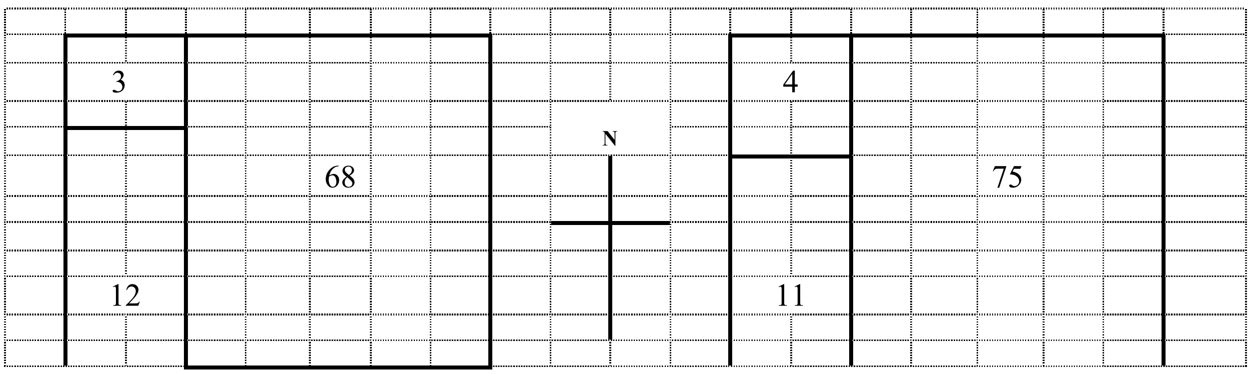
What is your answer to Tversky's problem?

Lack of motivation is common among the students I have known in Ukraine, where I have mostly taught in private universities; these have their admission policy determined less by students' academic interest than by the income of their parents. As a statistical aside, I have found this problem diminishes with class size; the greater the number of students, the higher the probability of there being among them enthusiasts with whom the others learn to compete. Nonetheless, I have to be prepared to deal with such answers as 'Dunno', or the Slavic equivalent; after indeed mentally biting my tongue in the effort to 'think slowly', I try as a prompt asking ' Do you mean you don't have enough information to decide – or that “it all depends” ? '.

Less discouragingly, other students will ask ' what formula should we use?'. I parry this by writing on the whiteboard such numerical answers as I have gleaned from the rest of the class, all of whom I direct to the next question.

How do you imagine the detail of the test undergone by the witness?

I am aware that I should allow time for the students to develop their own scenarios, but to date have lacked the patience to do other than steer them to the following of my own contrivance: 15 of the city’s Blue cabs and 85 of the Green were paraded in random order before the witness. She was envisaged as calling out in turn the colour she judged the cabs to be, so that each could be labeled with either a ‘B’ or a ‘G’ sticker, before being directed to one corner of a rectangular parking area. This has space for 10 cabs in each row and each column, 100 in all. To facilitate discussion of the diagram I’d sketch on the whiteboard, we imagined the park to be oriented on the cardinal directions so, for example, Blue cabs wrongly identified as Green ended up in the NW corner. My students have then had little difficulty in using the quantitative information in Tversky's story to calculate that number as 20% * 15% * 100 = 3, as shown Fig 1(a), before similarly confirming the numbers shown in the other three corners. The precision with which the reliability was specified in the problem, and the simplicity of the subsequent arithmetic once inspired a student to speculate that 'this was probably the scenario the experimenters had in mind'. I return to this comment in discussing the next question.



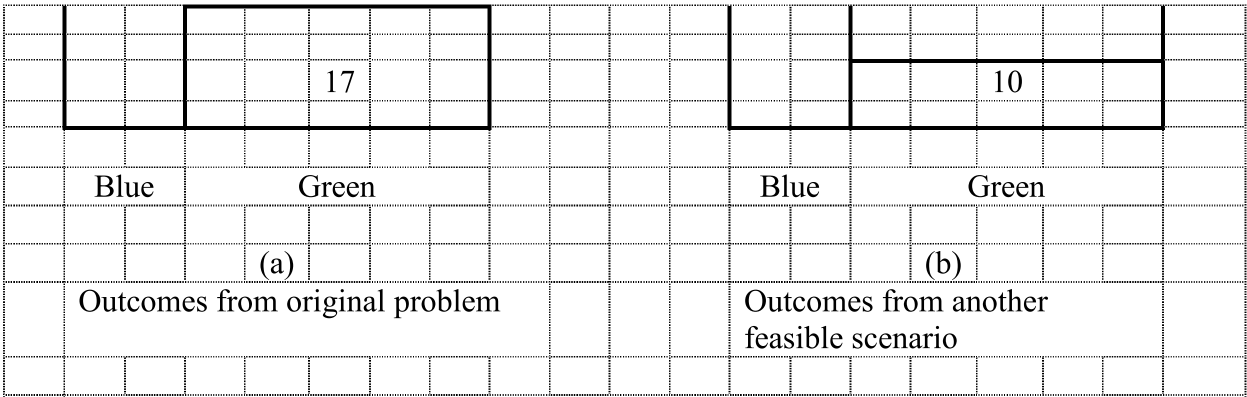


Figure 1. Final disposition of Cabs in the Car-Park

Why should the witness prove equally reliable with both colours?

Many years living in a student hall of residence have made me only too aware of the fire alarm dilemma: the more sensitive the setting of the detector, the smaller the risk of it failing to warn of a real emergency, but the greater the incidence of false alarms. By analogy, the witness will generally prove better at detecting one colour than the other, so a perfectly feasible result of the test could be the values in Fig. 1(b). Some students have objected that here the numerical values do not give a simple, whole-number value for the reliability, as was the case in the original problem. I address this by asking; 'are the reliability values in either case experimental data, or are they the calculated outcomes from such data?' They then appreciate that the reliability value of 80% in the original problem can only have been calculated from the results from the 'test', even if its construction in Tversky's imagination had been different from our own. I have found no need to introduce here any new terminology such as 'Type I 'and 'Type II' errors.

I am further exploring a boundary case, not shown diagrammatically, which has a third set of test results; my the aim is to explicate one of the common fallacies reported by Falk and Greebaum (1995, pp 81-82), namely that of applying to problems of inference arguments that are valid only in deductive logic. In this scenario we suppose the witness's reliability to increase to 100% when identifying Blue, but to remain less than 100% when identifying Green. We can express this as two conditional statements:
1) "If the witness says 'green', the cab is *certainly* Green". But if the test organizers know the cab to be Blue, then they deduce that the witness is *certainly* wrong.
2) "If the witness says 'blue', the cab is *probably* Blue". But if the test organizers know the cab to be Green, they should again deduce that the witness is *certainly* wrong,
However, subjects often conclude in (2) that the witness is only *probably* wrong, in the mistaken belief that any conclusion is subject to the same measure of doubt as the complement of the conditional statement. This is a fallacy which Falk and Greenbaum report as prevalent even among professional statisticians.

I used to make a practice of pointing out that the E-W division of the park (depicted by the line running N-S in the diagram) could be marked out *before* the test, which is why the base- rate is often called the '*Prior*'. However, I shall henceforth soft-pedal on this, having only recently appreciated from the writings of Falk (1989, p 178) that a conditioning probability does not necessarily have to precede the acquisition of experiential evidence.
More importantly, I find no difficulty in provoking students to question the plausibility of what they are trying to imagine; as Falk (1989, p 175) points out, there is inevitable 'ambiguity about the "given" in probability story-problems'. Even if 'the conditions that existed on the night of the accident' can be preserved throughout the parade, there is an obvious problem of temporality: will the condition of the witness not change, during both the interval between the accident and the test, as well as in the course of the test itself? Won't her power of discrimination then be

either improving with experience, or deteriorating through boredom, thus invalidating the assumption that her judgments during the test were representative of her judgment at the time of the accident?

.

Why should the witness's reliability be independent of the base rate?

To discuss this question, I put two others to the students:

- 'Have you ever been to New York?
- What colour are most cabs in New York?

Even those who answer 'No' to the first question universally answer 'Yellow' to the second, and the class readily grasps that any answer to Tversky's question that is based on the numbers in Fig 1(a) must rest on the strong assumption that experimental evidence – here the witness's reliability -- is independent of the base rate. To drive the point home we replace "Green' in the original story with 'Yellow'; then, as one student put it, 'I can almost hear the witness asserting in court "yes, of course I know that most cabs are yellow, that's why I'm so sure that the cab I saw was different" '. Yet in his book, Kahneman states unequivocally that the ' correct answer ' to this 'standard problem' is 41%, which from Fig. 1(a) we can check as

$$\{12/ (12+17)\} \times 100\%$$

By contrast, most of my students have given values close to the witness's 'reliability' score of 80%. This is in common with most of the subjects in Tversky's original experiment, whom Kahneman accuses of 'base-rate neglect' (Kahneman 2012, p88), and hence concludes that we human beings are by intuition imperfect statisticians. But perhaps Tversky and Kahneman are no better than the rest of us?

Conclusion

I used to invoke Tversky and Kahneman's Cab problem in introducing Bayes's formula to my class, but hope to have shown the story's effectiveness in promoting deep learning without mention of Bayesianism. This is a highly controversial area that has been more thoroughly explicated by others, notably Falk (1989, pp180-182) ; writing well after the first publication of the cab problem, she and a colleague (Falk & Greenbaum, 1995, p91) concluded that 'no single procedure can be offered to replace that ritual; surely not any mechanical recipe'. Their 'ritual' and 'recipe' refer to hypothesis testing, which necessarily subsumes Thomas Bayes's theories. Surprisingly, however, the authors do award Bayes the accolade of referring to his *theorem* (p78), a term avoided by Kahneman(2012) who instead speaks of Bayes's *rule* (p166) and of Bayesian *statistics* (p154)*, thinking* (p169) and *reasoning* (p172). Teachers, at grade-school level at least, could be forgiven for deciding that Bayesianism is a minefield where they should fear to tread.

However, the choice is not open to that majority of Western teachers who are constrained by syllabuses which they may aspire to influence, but which are nonetheless state-imposed. Moreover, Googling 'Conflicts in the Classroom' will reveal the controversies already extant in the classroom throughout all stages of education, at least in the Anglosphere. Their causes are rooted not in what is taught, but in racism and sexism; while it is in this context that Varmi-Joshi (2007), for instance, has asserted that '…teaching controversial issues is all-important', there is no reason to suppose that this should apply to the substance of our teaching any less than its conduct.

So. feeling it imperative to grasp the nettle of Bayesianism with at least my abler students, I direct them to another best-selling paperback, *The Black Swan,* by Nicolas Taleb (2008).

Ironically, this is despite his making not a single reference to Bayes; rather it is because he draws extensively on Tversky and Kahneman’s work while writing dismissively about statistics and statisticians in general.

More particularly, among the sins Taleb lays at statisticians' door is what he dubs the ‘ ludic fallacy’ (Taleb, 2008, pp122-123); by this he means that there is nothing to be learnt about the untidy, inferential 'real' world from games involving dice or other artifacts of randomness. Given that it is precisely the comparison to such artifacts that forms the bedrock of statistical hypothesis testing, it is vital for our students -- indeed, for all of Taleb’s readers -- to confute this as one of the more outrageous of his claims. So once embarked on the teaching Bayesianism, I call on the dice-based ‘Bayes Icebreaker’ developed by Jessop (2010); in the Appendix I show how this can be related to Tversky's cab-story.

Though often using statisticians as subjects, in his book Kahneman (2012) nowhere describes himself as a statistician. But a question he might ask in discussing the base-rate bias (pp 146-150) is ' Are statisticians likely to be in the majority of people who read either my book or Taleb's?' The answer is surely 'no'.   Identifying the intellectual conflicts latent in these two books in itself stimulates students' deep learning. Indeed, this is a message we teachers can proselytize more widely whenever these books are mentioned in social conversation. Axiomatically, as best-sellers, they often are.

Appendix

Bayes Ice-Breaker

The table in Fig 2 below is an adaption of the table in the 'Ice-breaker' developed by Jessop (2010). In the original. Jessop considers three six-sided dice one with the faces inscribed with

Constituents of sample Space			
Evidence	Vowel	Consonant	Sum
NATHAN	2	4	6
ANTHEA	6	6	12
	Likelihoods		
Evidence	Vowel	Consonant	Base Rates
NATHAN	2/6	4/6	1/3
ANTHEA	6/12	6/12	2/3

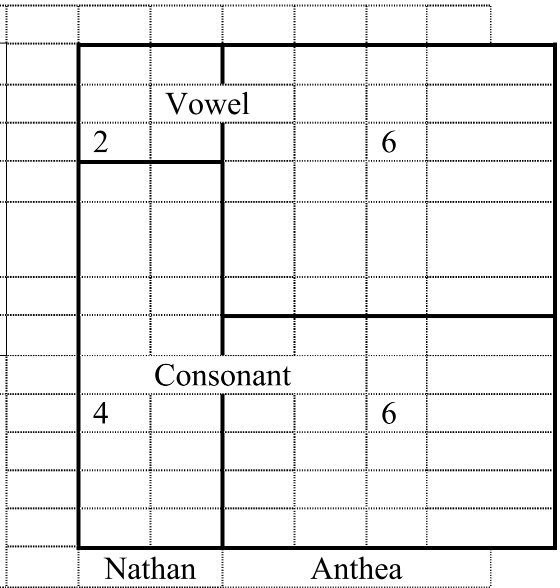


Figure 2. Jessop’s Icebreaker with two Antheas and one Nathan

tossed, the forename NATHAN, and two with the name ANTHEA. The problem is to calculate the relative diagnostic worth of each of the letters A-E-H-T-N when displayed after tossing just one of the three dice picked at random: if for example an N appears, the chances of either name are 50:50, while an E makes ANTHEA a 100% certainty. This nicely illustrates Bayes's formula. To permit a comparison with the cab problem, I reduce the evidence to the Vowel/Constant dichotomy. The table in Fig 2 thus corresponds to Jessop’s Tables 1 and 2, while the diagram on the right shows the 'cab-park' equivalent of his problem, From either the tables or the diagram, we readily deduce that the probability of the dice being Anthea is 6/ (6+2) = 75% if a vowel is tossed, but only 6/ (6+4) = 60% in the case of a consonant**.**

References

Bedwell, M (2009) Rescuing Statistics from the Mathematicians. *Proceedings of Conference on Modeling in Mathematics Education*, Dresden, September 2009, 52-57.

Falk, R. and Greenbaum, C.(1995). Significance Tests Die Hard. *Theory & Psychology*. **5**(1), 75-98.

Falk, R. (1989). Inference under uncertainty via conditional probabilities. In Morris, R (Ed) Studies in mathematics education, Vol 7: The teaching of statistics (175-184) Paris: Unesco, 1989 (reprinted with corrections 1991)

Jessop, A. (2010). Bayes Ice-Breaker. *Teaching Statistics*. **23**(1), 13-16.

Kahneman, D. (2012). *Thinking, Fast and Slow*. Penguin.

Taleb, N. (2008). *The Black Swan*. Penguin.122-123

Tversky, A and Kahneman, D.(1980), Causal Schemas in Judgement Under Uncertainty, in Kahneman, D. (2012),166, 167; Notes 402.

Varmi-Joshi, S (2007) in Foster, C. (2013), The essential guide to secondary mathematics, Routledge

Acknowledgements

I am indebted to

1) the students who over the years have stimulated my own deep learning
2) the anonymous referee for her/his helpful comments on my initial submission of this article, and
3) most of all, Professor Ruma Falk of the Hebrew University of Jerusalem for her severe but constructive criticism of much of my writing on statistics.