University of Waterloo W. H. Cherry

PARADOXES: Silver cabs/grey cabs

Experience shows that probability is a topic many people find difficult, in that mistakes are easier to make than in (some) other subject areas. This is one of six Statistical Highlights (#46 to #51) which discuss probabilistic subtleties and mistakes (which sometimes lead to so-called 'paradoxes'), with a view to helping the reader recognize and deal correctly with such matters. [These and related statistical issues are also discussed in Figure 7.12 of the STAT 220 Course Materials.]

The silver cabs/grey cabs context of this Highlight #50 involves the same probabilistic structure as Statistical Highlight #49 but the focus is different – trying to understand how people reason when dealing with (what are modelled as) conditional probabilities. A more usual name of this context is *Green cabs/blue cabs* – the change here is solely for alphabetic convenience; the choice of silver and grey tries to retain the possibility of a witness confusing the two colours under conditions of poor visibility.

1. The Context

A city has two taxi companies, the larger Silver cabs and smaller Grey cabs – 85% of the cabs are Silver, 15% Grey. A cab was involved in a hit-and-run accident at night; a witness reported that the cab was grey. In subsequent testing under visibility conditions similar to those at the accident, the witness identified cab colour *correctly* in 80% of tests for *both* cab colours. The Question of interest in this Highlight #50 is: *What is the probability the cab involved in the accident really was grey?*

2. Conditional probability calculations

A useful approach to this questions involves using conditional probability; a *dis*advantage is that the reasoning may be less accessible to those without the relevant specialized knowledge.

Define: event G – the cab in the accident was grey;
event S – the cab in the accident was silver;
event W_G – the witness reports the cab as grey;
event W_S – the witness reports the cab as silver.

$$\Pr(G) = 0.15 \qquad \text{-----(1)}$$

$$\Pr(S) = 0.85 \qquad \text{-----(2)}$$

$$\Pr(W_G|G) = \Pr(W_S|S) = 0.8 \qquad \text{-----(3)}$$

Assuming that *any* of the cabs in the city is equally likely to have been involved in the accident, the first two events have probabilities determined by the proportions of the two colours of cab in the city, as in equations (1) and (2) at the right above. Equation (3) expresses the two equal success rates for the witness colour identification. The discussion below in this Section 2 focuses on G because it is the cab color reported by the witness.

Also, the events S and G are complements – that is: $S \equiv {}^cG$ and $G \equiv {}^cS$; their probabilities add to 1 and so equation (4) follows.

$$\Pr(W_G|S) = \Pr(W_S|G) = 0.2 \qquad \text{-----(4)}$$

As shown at the right, we first use Bayes' rule as in equation (5) and then expand the denominator as in equation (6) – the vertical lines (|) denote *conditional on* (or *given gthat*).

$$\Pr(G|W_G) = \frac{\Pr(W_G|G)\times\Pr(G)}{\Pr(W_G)} \qquad \text{-----(5)}$$

$$= \frac{\Pr(W_G|G)\times\Pr(G)}{\Pr(W_G|G)\times\Pr(G) + \Pr(W_G|S)\times\Pr(S)} \qquad \text{-----(6)}$$

Using the probabilities provided by equations (1), (2), (3) and (4) in equation (6) yields equation (7).

$$\Pr(G|W_G) = \frac{0.8\times0.15}{0.8\times0.15 + 0.2\times0.85} = {}^{12}\!/_{29} \simeq \mathbf{0.4138}. \qquad \text{-----(7)}$$

Reasoning as for equation (7) leads to equation (8),

$$\Pr(S|W_S) = \frac{\Pr(W_S|S)\times\Pr(S)}{\Pr(W_S|S)\times\Pr(S) + \Pr(W_S|G)\times\Pr(G)} = \frac{0.8\times0.85}{0.8\times0.85 + 0.2\times0.15} = {}^{68}\!/_{71} \simeq \mathbf{0.9577}. \qquad \text{-----(8)}$$

A key issue in this Highlight #50 is to contrast the probabilities of 0.8 for both $\Pr(W_G|G)$ and $\Pr(W_S|S)$ in equation (4) with the differing values of $\Pr(G|W_G)$ and $\Pr(S|W_S)$ in equations (7) and (8) – this comparison is between:

- the accuracy of cab colour identification by the witness [equation (4)], **AND:**
- the probability the colour of a cab *is* as stated by the witness [equations (7) and (8)],

in the context of the two taxi companies with fleets of substantially *differing* sizes (with, respectively, 85% and 15% of the city's cabs). Examining the components of equations (7) and (8) identifies this size difference as the source of the probability differences – the two values in equations (7) and (8) being (respectively) appreciably smaller (about 41%) and larger (about 96%) than the common *accuracy* of 80%. [The key fleet size difference here may elsewhere be (opaquely) described as a *base-rate* difference.]

This (surprising?) comparison raises two matters:

- the *subtle* distinction between the *descriptions* of the two pairs of conditional probabilities and their *clear* differences in *value*;
- the practical implications of confusing two conditional probabilities with respect to *which* is the conditioning event.

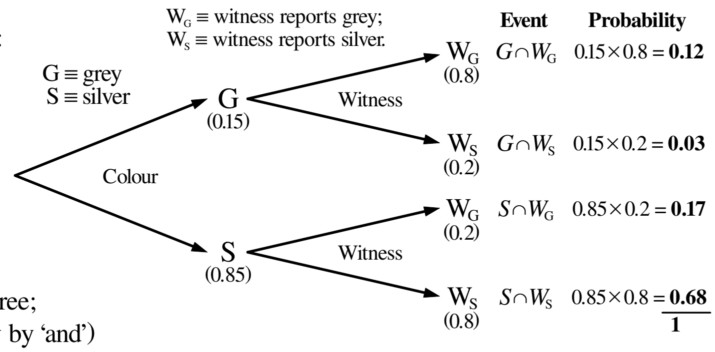
The second of these matters could arise in a law court if jurors were to accord witness testimony too high an accuracy resulting in an untoward effect on their verdict. [A possible role of Simpson's Paradox in interpreting data on the effect of *jury challenges* in the U.K is discussed at the top of page HL51.8 in Statistical Highlight #51.]

The potential difficulty of identufying a correct conditional probability and the consequenes of a mistake in real-world situations is why a context like that of silver cabs/grey cabs has been of interest from as far back as 1972 – see the Source (starting on page 12) given overleaf at the bottom of page HL50.2 and the two references to authors Khaneman and Taversky quoted from it; five additional references involving one or both these authors are given by Oldford in the Source.

(*continued overleaf*)

3. Can conditional probability ideas be made more widely accessible?

In everyday living, we are routinely faced with making decisions under uncertainty; when (what are actually) conditional probabilities are involved, as they often are, the foregoing discussion in this Highlight #50 shows that mistakes are easy to make. It would therefore be useful if the (difficult) ideas of conditional probability could be made accessible more widely in society by trying to avoid the specialized terminology of introductory probability courses. A valuable insight in this regard, brought to the writer's attention by Prof. Chris Springer in the 1980s, is to use a *tree diagram* as an adjunct to a more formal approach like that in equations (1) to (8) overleaf on page HL50.1. Discussion elsewhere (for instance, in the Source below and its references) suggests that others have had this same insight.



An illustration is the tree diagram at the right, where:

- the left-hand two branches of the tree show the two cab colours and their probabilities in brackets ( ) [from equations (1) and (2) overleaf on page HL50.1];
- the right-hand four branches show, for each cab colour, the witness colour reports and their *conditional* probabilities in brackets ( ) [from equations (3) and (4)] – the conditioning reflects the two preceding left-hand branches of the tree;
- the *combination* of the two events (indicated verbally by 'and') that occur along a (complete) branch of the tree is their *intersection* – it is shown in the 'Event' column at the right of the tree with the component events in the order they occur in the tree, but the intersection would be *un*affected were they to be shown in the reverse order [see equation (9) below];
- invoking a modelling assumption of *probabilistic independence*, the probability of each intersection is the *product* of the probabilities of its component events, as shown in the rightmost column – these four probabilities (of course) sum to 1.

NOTE: The tree diagram above, with *three* branches each with *two* arrows (which yield *four* outcomes), is the simplest tree; more complicated situations may involve branches with more than two arrows (*e.g.*, three arrows for the three doors in the Monty Hall situation discussed in Statistical Highlight #49) and/or more than two sets of branches across the treee.

Unfortunately, tree diagrams do not circumvent all the difficulties of a more formal approach.

* Tree branches must be *ordered* correctly – we start on the left with those for events with *un*conditional probabilities, then add in to their right the branches for events that are conditioned by events of preceding branch(es).
* The conditional probabilities of interest in this Highlight #50 are not generated *directly* by the tree but they are easily calculated from the probabilties in the rightmost column.
  + The value of $^{12}/_{29}$ in equation (7) is the value from the first branch divided by the sum of the first and third branch values: $0.12 \div (0.12 + 0.17)$ – these are (unsurprisingly) the values in the central term of equation (7) overleaf.
  + The value of $^{68}/_{71}$ in equation (8) is the value from the fourth branch divided by the sum of the second and fourth branch values: $0.68 \div (0.03 + 0.68)$ – these are (unsurprisingly) the values in the third term of equation (8) overleaf.

We now see that the tree diagram imposes a 'natural' order on the two events in each intersection that corresponds, as does equation (7), to the *first* term of equation (9) at the right, whereas the ordering in the *last* term is that in equations (3) and (4). It is curious that the *same* intersection of two events can be the source of conditional probabilities, with their component events reordered, that are the central concern of the discussion overleaf on page HL50.1.

$$\Pr(G \mid W_G) \times \Pr(W_G) = \Pr(G \cap W_G) \equiv \Pr(W_G \cap G) = \Pr(W_G \mid G) \times \Pr(G) \qquad \text{-----(9)}$$

Equation (9), a conditional probability result that leads to Bayes' rule [see equation (5) overleaf], shows that conditional probabilities are the (unusual) *quotient* of two probabilities, which may be (part of) the reason they are difficult to illustrate pictorially.

Equation (9) also shows that a conditional probability involves a *relationship* of events, their intersection; in the real world, relationships come in immense variety and are usually difficult to mathematize, which may account for why conditional probabilities as models involve difficult ideas (and why the real-world implications of probabilistic *in*dependence are usually troublesome).

SOURCE: Oldford, R.W.: *Probability, problems, and paradoxes pictured by eikosograms*, April 21, 2003 (35 pages, 17 references); url http://math.uwaterloo,ca > examples > paper PDF

Prof. Oldford states on page 12 that; *This problem* (that is, the two cabs context) *was first introduced by Kahneman and Taversky (1972) and has been presented to many subjects in slight variations by different investigators. The above version is that given in Taversky and Kahneman (1982, pp. 156-157).*

The two citations are given on page 35 as:

Kahneman, D. and A. Taversky (1972): *On prediction and judgement*, ORI Research Monograph, 12(4).

Taversky, A. and D. Kahneman (1982): Ch. 10, *Evidential impact of base rates* in Kahneman, D., Slovic, P. and A. Taversky (editors, 1982): *Judgement under uncertainty: Heuristics and biases.* Cambridge University Press, Cambridge, UK.

The first citation is exactly as on page 535 of the book in the second; it is likely that ORI is *Oregon Research Institute.*