DS 100/200: Principles and Techniques of Data Science Date: January 24, 2020

Discussion #1

Name:

In Data 100 you will typically work with multiple variables and large data sets. But before we get carried away by complexity, let's make sure we have our feet on the ground when it comes to interpreting simple quantities like proportions.

Prosecutor's Fallacy

1. Investigators at the scene of a crime find a footprint that shows a distinctive pattern on the sole. They identify the type of shoe, and then they find a person owns that kind of shoe and could have committed the crime. They put this person on trial for the crime.

   After looking at sales patterns and so on, the investigators find that of the 10,000 other people who could have committed the crime, 1 in 1,000 own that kind of shoe.

   The prosecution says that given these findings, the chance that the defendant is not the guilty person is 1 in 1000.

   The prosecution has made an error called the "prosecutor's fallacy." Unfortunately it's rather common. Let's see what the error is and what conclusions we can draw from the evidence.

   (a) There are 10,001 people who could have committed the crime. Define a person to be "Consistent with Evidence" if the person owns the kind of shoe identified by the investigators. Fill in the table below with the counts of people in the four categories. The four counts should add up to 10,001, and you should assume, as the prosecution did, that only one person is guilty.

	Guilty	Not Guilty
Consistent with Evidence	1	10
Not Consistent with Evidence	0	9990

   (b) The prosecution has reported a proportion as a chance. Whether they know it or not, this implies they are assuming that the defendant is like a person drawn at random from the group who could have committed the crime. So let's assume that too. That is, we assume the defendant is drawn at random from 10,001 people of whom 1 is guilty.

   Use the table in Part **a** to fill in the blanks with choices from among "Guilty", "Not Guilty", "Consistent with Evidence", and "Not Consistent with Evidence". The vertical bar is the usual notation for "given".

Under this assumption, $\frac{1}{1000} = P(\underline{\text{Consistent}} \mid \underline{\text{not guilty}})$.

we know that of not guilty ppl, 1 in 1000 are consistent

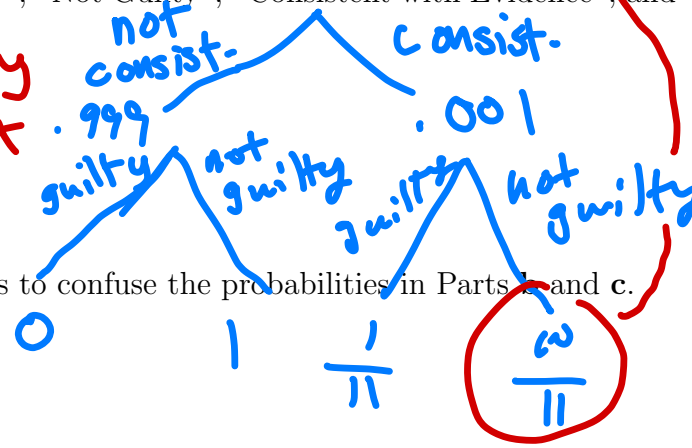
(c) What the investigators know is that the defendant has the fateful type of shoe. Fill in the blanks:

Given the findings of the investigators, the chance that the defendant is not guilty is $P(\underline{\text{①}} \mid \underline{\text{②}}) = \underline{\text{③}}$.

The last blank should be filled with a fraction, and the first two should be filled choices from among "Guilty", "Not Guilty", "Consistent with Evidence", and "Not Consistent with Evidence".

1) not guilty
2) Consistent
3) $\frac{10}{1+10} = \frac{10}{11}$

we know defendant is consistent



this is what we're looking for, which is ①

Note: The prosecution's error is to confuse the probabilities in Parts **b** and **c**.

Election Forecasts

2. People have a hard time understanding polls. In September 2016, the New York Times tried to explain aspects of polls that tend to get overlooked. To illustrate the issues, they gave all the data in one of their own polls to four well-known forecasters and asked them to make predictions.

   The data were from a poll of 867 Florida voters and the exercise was to predict Trump/-Clinton result in Florida. In the election, Trump beat Clinton in Florida, 49% to 47.8%.

   Here are the forecasts. The Times' own results, derived jointly with researchers at Siena College, are in the last line of the table.

Pollster	Clinton	Trump	Margin
Charles Franklin Marquette Law	42%	39%	Clinton +3%
Patrick Ruffini Echelon Insights	39%	38%	Clinton +1%
Omero, Green, Rosenblatt Penn Schoen Berland Research	42%	38%	Clinton +4%
Corbett-Davies, Gelman, Rothschild Stanford University/Columbia University/Microsoft Research	40%	41%	Trump +1%
NYT Upshot/Siena College	41%	40%	Clinton +1%

(a) Pick one of the options (i) and (ii); if you pick (ii), provide the reason.

The predictions were different from each other because

(i) samples come out differently due to randomness so the forecasters all had different data.

(ii) different analyses use same data but different models and weights, & ∴ predict different outcomes

(b) Point out one other interesting aspect of the data in the table. This question doesn't have just one right answer; just describe something you noticed.

e.g. all predictions sum to << 100%, so "other" was significant in the election

(c) If you were going to forecast an election result, which of the following groups would you most want to focus on, and why? Pick at most two groups.

(i) adults in the Census

(ii) eligible voters

(iii) registered voters

(iv) likely voters

(v) undecided voters

we want people who CAN vote (v) & ppl who likely will vote (iv) as these are the ppl who determine the outcome of an election BY VOTING

(d) Of the two main methods for identifying likely voters, described below, one does a better job at predicting whether the person will show up and vote. Which do you think it is, and why? Could it systematically exclude some likely voters?

- Self-reported voting intention
- Voting history (in which past elections did the person vote; data available in the voter registration database)

this is more accurate than self-reported data.

Yes, people who did not register to vote previously, e.g. young voters who turned 18 after the last election