charleyferrari_week9hw

Charley Ferrari

December 10, 2015

Question 2.1

Suppose we have a four-sided die from a board game. On a tetrahedral die, each face is an equilateral triangle. When you roll the die, it lands with one face down and the other three faces visible as a three-sided pyramid. The faces are numbered 1-4, with the value of the bottom face printed (as clustered dots) at the bottom edges of all three visible faces. Denote the value of the bottom face as x. Consider the three mathematical descriptions of the probabilities of x. Model A: $p(x) = \frac{1}{4}$. Model B: $p(x) = \frac{x}{10}$. Model C: $p(x) = \frac{12}{25x}$. For each model, determine the value of $p(x)$ for each value of x. Describe in words what kind of bias (or lack of bias) is expressed by each model.

A: $p(x) = \frac{1}{4}$ $p(x = 1) = \frac{1}{4}$ $p(x = 2) = \frac{1}{4}$ $p(x = 3) = \frac{1}{4}$ $p(x = 4) = \frac{1}{4}$

There is no bias in this model. Each face of the die has an equal probability of landing face down.

B: $p(x) = \frac{x}{10}$ $p(x = 1) = \frac{1}{10}$ $p(x = 2) = \frac{2}{10}$ $p(x = 3) = \frac{3}{10}$ $p(x = 4) = \frac{4}{10}$

This model is biased towards larger values. Given the fact that each face is the same, these larger values shouldn't be more likely to land face down.

C: $p(x) = \frac{12}{25x}$ $p(x = 1) = \frac{12}{25}$ $p(x = 2) = \frac{12}{50}$ $p(x = 3) = \frac{12}{75}$ $p(x = 4) = \frac{12}{100}$

This model is biased towards smaller values. Again, this shouldn't be, since smaller values shouldn't b e more likely to land face down either.

Question 5.1

This exercise extends the ideas of Table 5.4, so at this time, please review Table 5.4 and its discussion in the text. Suppose that the same randomly selected person as in Table 5.4 gets re-tested after the first test result was positive, and on the re-test, the result is negative. When taking into account the results of both tests, what is the probability that the person has the disease? Hint: For the prior probability of the re-test, use the posterior computed from the Table 5.4. etain as many decimal places as possible, as rounding can have a surprisingly big effect on the results. One way to avoid unnecessary rounding is to do the calculations in R.

First lets recalculate the probability of having the disease after being tested positive:

```r
psad <- 0.001

pplussad <- 0.99
pplushappy <- 0.05

psadplus <- (pplussad*psad)/(pplussad*psad + pplushappy*(1-psad))

psadplus
```

```
## [1] 0.01943463
```

So I get 1.9%, which is the same value mentioned in the book. Lets now use this as our new prior distribution. psad will change (to psadplus), but the conditional probabilities will remain the same.

```
psad <- psadplus

psadplusminus <- ((1-pplussad)*psad)/((1-pplussad)*psad + (1-pplushappy)*(1-psad))

psadplusminus
```

```
## [1] 0.0002085862
```

Looks like the probability is around 0.02%. For a sanity check, lets see what the probability of being sick given just a negative test is. It should be less than this:

```
psad <- 0.001

psadminus <- ((1-pplussad)*psad)/((1-pplussad)*psad + (1-pplushappy)*(1-psad))

psadminus
```

```
## [1] 1.053674e-05
```

As expected, it's lower.

Question 5.2

a. Suppose that the population consists of 100,000 people. Compute how many people would be expected to fall into each cell of Table 5.4. To compute the expected frequency of people in a cell, just multiply the cell probability by the size of the population. To get you started, a few of the cells of the frequency table are filled in here:

```
n <- 100000

freqplussad <- pplussad*psad*n

freqminussad <- (1-pplussad)*psad*n

freqplushappy <- pplushappy*(1-psad)*n

freqminushappy <- (1-pplushappy)*(1-psad)*n

freqsad <- psad*n

freqhappy <- (1-psad)*n

freqplus <- (pplussad*psad+pplushappy*(1-psad))*n

freqminus <- ((1-pplussad)*psad+(1-pplushappy)*(1-psad))*n

solutionmatrix <- matrix(c(freqplussad,freqminussad,freqsad,
                           freqplushappy, freqminushappy, freqhappy,
                           freqplus, freqminus, 100000),nrow=3)

rownames(solutionmatrix) <- c("Positive", "Negative", "Total")
```

```
colnames(solutionmatrix) <- c("Sad", "Happy", "Total")

solutionmatrix
```

```
##          Sad Happy  Total
## Positive  99  4995   5094
## Negative   1 94905  94906
## Total    100 99900 100000
```

b. Take a good look at the frequencies in the table you just computed for the previous part. These are the so-called "natural frequencies" of the events, as opposed to the somewhat unintuitive expression in terms of conditional probabilities (Gigerenzer & Hoffrage, 1995). From the cell frequencies alone, determine the proportion of people who have the disease, given that their test result is positive. Before computing the exact answer arithmetically, first give a rough intuitive answer merely by looking at the relative frequencies in the row D = +. Does your intuitive answer match the intuitive answer you provided when originally reading about Table 5.4? Probably not. Your intuitive answer here is probably much closer to the correct answer. Now compute the exact answer arithmetically. It should match the result from applying Bayes' rule in Table 5.4.

Intuitively, looking at the positive row, this should be a pretty low probability.

```
pdiseasepositive <-
  solutionmatrix["Positive","Sad"]/solutionmatrix["Positive","Total"]

pdiseasepositive
```

```
## [1] 0.01943463
```

c. Now we'll consider a related representation of the probabilities in terms of natural frequencies, which is especially useful when we accumulate more data. This type of representation is called a "Markov" representation by Krauss, Martignon, and Hoffrage (1999). Suppose now we start with a population of N = 10,000,000 people. We expect 99.9% of them (i.e., 9,990,000) not to have the disease, and just 0.1% (i.e., 10,000) to have the disease. Now consider how many people we expect to test positive. Of the 10,000 people who have the disease, 99%, (i.e., 9,900) will be expected to test positive. Of the 9,990,000 people who do not have the disease, 5% (i.e., 499,500) will be expected to test positive. Now consider re-testing everyone who has tested positive on the first test. How many of them are expected to show a negative result on the retest? Use this diagram to compute your answer.

```
n <- 10000000

freqsad <- psad*n
freqhappy <- (1-psad)*n

freqplussad <- pplussad*freqsad
freqplushappy <- pplushappy*freqhappy

freqplusminussad <- (1-pplussad)*freqplussad
freqplusminushappy <- (1-pplushappy)*freqplushappy

freqplusminussad
```

```
## [1] 99
```

```
freqplusminushappy
```

```
## [1] 474525
```

d. Use the diagram in the previous part to answer this: What proportion of people, who test positive at first and then negative on retest, actually have the disease? In other words, of the total number of people at the bottom of the diagram in the previous part (those are the people who tested positive then negative), what proportion of them are in the left branch of the tree? How does the result compare with your answer to Exercise 5.1?

```
pplusminussad <- freqplusminussad/(freqplusminussad+freqplusminushappy)

pplusminussad
```

```
## [1] 0.0002085862
```

```
psadplusminus
```

```
## [1] 0.0002085862
```

The two results match!