Notebook

September 14, 2019

0.0.1 Question 3

Given our population of Pennsylvania voters, every possible SRS has a certain well defined probability of occurring. For example, if we collected a SRS of only 2 voters without replacement (instead of 1500), the chance of each SRS is given in the probability distribution table below:

N_T	N_C	N_O	p
0	0	2	0.00189373515
0	1	1	0.04131080160
1	0	1	0.04193604504
0	2	0	0.22529213625
1	1	0	0.45740422268
2	0	0	0.23216338697

As an exercise in probability, we will have you compute similar probabilities for a simple random sample of 4 people without replacement.

Part 1 Find the following probability. Give your answer as an exact number (i.e. as a product of fractions).

$$P(N_T = 4, N_C = 0, N_O = 0)$$

Hint: It is a product of four fractions, each of which has a distinct numerator and denominator.
$=\frac{(2{,}970{,}733)}{(6{,}165{,}478)} \cdot \frac{((2{,}970{,}732)}{(6{,}165{,}477)}) \cdot \frac{(2{,}970{,}731)}{(6{,}165{,}476)} \cdot \frac{(2{,}970{,}730)}{(6{,}165{,}475)}$

Part 2 The answer from part 1 was a bit unwieldy.

We can simplify the problem by assuming that the draws are *with replacement*. That is, once a draw is picked as a sample, it can be picked again in the future. In this case, what is the following probability:

$$P(N_T = 4, N_C = 0, N_O = 0)$$

Give your answer as an exact value.
$(\frac{(2{,}970{,}733)}{(6{,}165{,}478)})^4 = 0.0538998396$

Part 3 Under this same simplfying assumption (that we sample with replacement), find the following probabilities:

$$P(N_T = 2, N_C = 2, N_O = 0)$$

$$P(N_T = 2, N_C = 1, N_O = 1)$$

Hint: See the "Fun Problem Related to HW" from the September 5th lecture if you're not sure how to proceed.

$P(N_T = 2, N_C = 2, N_O = 0) = \frac{4!}{2!2!} \cdot (\frac{2{,}970{,}733}{6{,}165{,}478})^2 \cdot (\frac{2{,}926{,}441}{6{,}165{,}478})^2$

$P(N_T = 2, N_C = 1, N_O = 1) = \frac{4!}{2!} \cdot (\frac{2{,}970{,}733}{6{,}165{,}478})^2 \cdot (\frac{2{,}926{,}441}{6{,}165{,}478}) \cdot \frac{268{,}304}{6{,}165{,}478}$

0.0.2 Question 4

Can you generalize the above probability calculation and express the probability in terms of a probability mass function?

To set up the problem, let random variables N_1, N_2, N_3 be the number of Trump, Clinton, and Other voters selected, respectively.

Let p_1, p_2, p_3 be the chance of a Trump, Clinton, Other voter being chosen, respectively, and let n be the size of the sample drawn (with replacement).

In general, what is

$$P(N_1 = k_1, N_2 = k_2, N_3 = k_3) = ?$$

Hint: The answer involves$n!, k_1!, k_2!, k_3!$andp_1, p_2, p_3 raised to various powers. Also note this may be a particularly tough problem depending on your math background. Take your time, and please discuss with fellow students and instructors. You don't need to get this problem right to complete the questions later in this homework.

$P(N_1 = k_1, N_2 = k_2, N_3 = k_3) = \frac{n!}{k_1!k_2!k_3!} \cdot (p_1)^{k_1} \cdot (p_2)^{k_2} \cdot (p_3)^{k_3}$

Part 1 If we're trying to predict the results of the Clinton vs. Trump presidential race, what is the population of interest?

The population of the United States

Part 2 What is the sampling frame?

The population of Pennsylvania

0.0.3 Question 6

Why can't we assess the impact of the other two biases (voters changing preference and voters hiding their preference)?

Note: You might find it easier to complete this question after you've completed the rest of the homework including the simulation study.

If the voters kept their preferences hidden, it would be impossible to know so unless we had a look at their vote. And if the voters changed their preference, it is impossible to know if they were lying at a poll, and we would have to look at their vote to understand how the poll would have changed. No one is going to reveal their vote, and if they hid their preference or changed it last minute, they wouldn't want to share that with anyone. Hence it becomes really hard.

Part 4 Make a histogram of the sampling distribution of Trump's percentage advantage in Pennsylvania. Make sure to give your plot a title and add labels where appropriate. Hint: You should use the `plt.hist` function in your code.

Make sure to include a title as well as axis labels. You can do this using `plt.title`, `plt.xlabel`, and `plt.ylabel`.

```
In [22]: plt.hist?
```

Part 3 Compare the histogram you created in Q8.2 to that in Q7.4.

The positive proportion of Trump winning increases in the biased histogram. It shows how a biased poll can give fake results towards the election. The proportion advantage for Trump in the biased poll is more positive in the biased poll.

Write your answer in the cell below.

As you increase the sample size, the sampling error for the unbiased sample decreases as they have an inverse relationship when no bias is involved. But when we look at the biased case, the sampling error increases as the sampling size increases as the bias is more reflected in the larger sample size we are considering. Hence we can see Trump's winning percentage increase in the unbiased sample and decrease in the biased sample.

0.0.4 Question 10

According to FiveThirtyEight: ”... Polls of the November 2016 presidential election were about as accurate as polls of presidential elections have been on average since 1972.”

When the margin of victory may be relatively small as it was in 2016, why don’t polling agencies simply gather significantly larger samples to bring this error close to zero?

This is because their polls contain bias against specific candidates, and as we saw above, if the bias favors another candidate, then the other candidates get a lower estimated winning proportion, increasing the sampling error overall. Hence because of the bias, we get a larger sampling error as the population size increases

0.0.5 Question 11

Help us get to know you by filling out this survey! You will need to be logged into your UC Berkeley account to access it.

Once you've finished filling it out, you should see a secret string on the confirmation page. Assign `secret` to that secret string to get credit for taking the survey.

We will also go check that you actually filled out the survey before giving you points for this question, so make sure to submit it!

```
In [36]: secret = "4RQ9VBN"
```