

ORIE 5355/INFO 5370 HW 1: Survey Weighting

- Name:
- Net-id:
- Date:
- Late days used for this assignment:
- Total late days used (counting this assignment):
- People with whom you discussed this assignment:

After you finish the homework, please complete the following (short, anonymous) post-homework survey:

<https://forms.gle/AM1x5qEnLCvxsgJ7>

We have marked questions in **green**. Please put answers in the default color. You'll want to write text answers in "markdown" mode instead of code. In Jupyter notebook, you can go to Cell > Cell Type > Markdown, from the menu. Please carefully read the late days policy and grading procedure [here](#). In that link, we also give some tips on exporting your notebook to PDF, which is required for GradeScope submission.

A few notes about this homework:

1. This homework is purposefully heavy in using the Pandas package. Being able to explore data is an essential data science skill that you'll use throughout this class and your career -- even if the polling/politics application is not interesting to you. I encourage you to practice Pandas and learn how to use it well. Your code will NOT be graded on efficiency.
2. Some of the questions can be interpreted in multiple ways. That is always true in data science. You'll need to make judgment calls for what analysis to do. For the homework, you'll still receive full points for any "reasonable" choice. Also feel free to ask questions on EdStem.

Note: We sometimes provide sample images of what your output should look like. These are for clarification of what we are looking for. The numbers in the images are not necessarily correct, and your output does not need to look exactly like the images.

Conceptual component

1) Reading

Please read Sections 3 and 4 (pages 6-13) [here](#):

https://www.nber.org/system/files/working_papers/w20830/w20830.pdf, and answer the following questions.

Please summarize the sections in no more than two sentences.

In []:

Do you think it's a problem that most ratings are positive? If so, why? Answer in no more than four sentences. Please incorporate concepts discussed in class in your answer.

In []:

2) Personal reflection

Think back to a time that you trained a model on data from people or gathered opinions via a survey (an informal one is fine). If you have not done that before, you may answer these questions about an article in the news that reported

on public opinions or a model that you think might be in deployment at a company or organization with which you interact (for example, Amazon, google maps, etc)

Briefly summarize the scenario in no more than two sentences.

In []:

What was the construct that you cared about/wanted to measure? What was the measurement (numerical data)? In what ways did the measurement not match the construct you cared about? Answer in no more than 4 sentences.

In []:

What selection biases/differential non-response issues occurred and how did it affect your measurement? (If your answer is "None," explain exactly why you believe the assumptions discussed in class were met). Answer in no more than 3 sentences.

In []:

Given what we have learned in class so far, what would you do differently if faced with the same scenario again? Answer in no more than 3 sentences.

In []:

Programming component

In this part of the homework, we provide you with data from a poll in Florida before the 2016 Presidential election in the United States. We also provide you with (one pollster's) estimates of who will vote in the 2016 election, made before the election. You will use this data and apply the weighting techniques covered in class.

Preliminaries to load packages and data

```
In [ ]: import pandas as pd
import numpy as np
```

```
In [ ]: dfpoll = pd.read_csv('polling_data_hw1.csv') # raw polling data
dfpoll.head()
```

```
Out [ ]:
```

	candidate	age	gender	party	race	education
0	Someone else	30-44	Male	Independent	White	College
1	Hillary Clinton	45-64	Male	Republican	Hispanic	College
2	Hillary Clinton	30-44	Male	Independent	Hispanic	College
3	Hillary Clinton	65+	Female	Democrat	White	College
4	Donald Trump	65+	Female	Republican	White	High School

```
In [ ]: dfdemographic = pd.read_csv('florida_proportions_hw1.csv') # proportions of population
dfdemographic.head()
```

```
Out [ ]:
```

	Electoral_Proportion	Demographic_Type_1	Demographic_Type_2	Demographic_1	Demographic_2
0	0.388	party	NaN	Democrat	NaN
1	0.399	party	NaN	Republican	NaN
2	0.213	party	NaN	Independent	NaN
3	0.446	gender	NaN	Male	NaN
4	0.554	gender	NaN	Female	NaN

```
In [ ]: dfdemographic.tail()
```

```
Out [ ]:
```

	Electoral_Proportion	Demographic_Type_1	Demographic_Type_2	Demographic_1	Demographic_2
112	0.034	race	education	Hispanic	Some College
113	0.028	race	education	Hispanic	College
114	0.011	race	education	Other	High School
115	0.011	race	education	Other	Some College
116	0.015	race	education	Other	College

`dfdemographic` contains estimates of likely voters in Florida in 2016. When `Demographic_Type_2` is `NaN`, the row refers to just the marginal population percentage of the group in `Demographic_1` of type `Demographic_Type_1`. When it is not `NaN`, the row has the joint distribution of the corresponding demographic groups.

For example, row 0 means that 38.8% of the electorate is from the Democrat party. Row 113 means that 2.8% of the electorate is Hispanic AND graduated college.

Part A: Raw visualization

Here, we'll visualize whether the respondents in the poll match the likely voter estimates. Create a scatter-plot where each point represents one Demographic group (for example, people who identify as Independent: party-Independent), where the X axis is the `Electoral_Proportion` in `dfdemographic`, and the Y axis is the proportion in `dfpoll` (Hint: you will need to calculate this yourself).

```
In [ ]:
```

`dfdemographic` tells us the true proportion of each demographic group in the entire population, the calculated proportion from `dfpoll` tells us the proportion of each demographic group represented in the poll. For a poll to be representative of the population, ideally we would want the latter to be the same as the former. However, this is hard to achieve in practice.

In your view, which group is most over-represented? Most under-represented? Why? Answer in no more than 3 sentences. There are multiple reasonable definitions of "over" or "under" represented; any choice is fine as long as you justify your answer.

```
In [ ]:
```

Part B: Weighting

For this question, we'll ignore people who answered anything but "Hillary Clinton" or "Donald Trump."

You'll notice that some of the groups in the polling data ("refused") do not show up in the population percentages. For the questions that require weighting by demographics, ignore those respondents.

1) Raw average

Below, report the "raw polling average," the percentage of people "Hillary Clinton" divided by the number who answered either Hillary or Trump.

In []:

2) Single dimensional marginal weighting (on just 1 demographic type)

For each demographic type separately -- age, gender, party, race, and education -- weight the poll by just that demographic type, in accordance to the population proportions given. Report the resulting poll results, and briefly (at most 3 sentences) describe what you observe.

For example, when weighted by race, you'll report:

Weighted by race --- Clinton: 0.531, Trump: 0.469

(your results might be different due to rounding/precision, but we do not expect the difference to be large)

In []:

2-dimensional joint distribution weighting

Now, for each pair of demographic types in `dfdemographic`, do the same -- weight the poll by that pair of demographic types, in accordance to the given joint distributions, and briefly (at most 3 sentences) describe what you observe.

For example, when weighted by race and age, you'll find:

Weighted by age and race: Clinton: 0.523, Trump: 0.477

In []:

3) 2-dimensional marginal

We don't always have access to joint distributions across the population -- for example, it may be hard to estimate from past exit polls (surveys done as people are leaving the polling station) what the joint distribution of education and gender is, for example. However, access to marginal distributions are often available.

As discussed in class, one strategy when you don't have access to joint distributions -- only marginals -- is to *multiply* the marginal distributions. For example, if 50% of your population is Democratic and 50% is a woman, then pretend that 50% times 50% = 25% of your population is a Democratic women. Clearly this technique is not perfect, but it is sometimes a useful heuristic. (Hint: you can use the marginal distribution provided in `dfdemographic`)

For the following pairs of Demographic types, report the weighting results if you use the joint distributions in `dfdemographic` versus if you approximate the joint distribution using the marginals. Briefly (at most 3 sentences) describe what you observe.

(party, gender)

(race, gender)

As an example output, here's the results for two other pairs of demographics (your results might be different due to rounding, but we do not expect the difference to be large):

Demo1	Demo2	Joint	Marginal
age	race	0.523431	0.525669

Demo1	Demo2	Joint	Marginal
age	education	0.525068	0.523938

In []:

4) Bonus points (up to 2 points): Implement a "cheap" version of the MRP technique mentioned in class.

The above techniques use the mean answer among people who share a demographic as the estimate for that demographic. But that wastes information across demographics. For example, maybe people who only have "Some College" are similar enough to people who have "High School" as to provide some useful information.

First, do the following: use a logistic regression (or your favorite prediction tool) to predict candidate choice, using the demographics. You might want to convert some demographics (like education) to ordered numeric (e.g., 1, 2, 3) as opposed to using discrete categories.

Here, you will earn partial bonus points by just reporting the predictions and comparing them to the means of each covariate group in the raw polling data. Give a scatter-plot, where each point is one combination of full demographics (age, gender, party, race/ethnicity, education), the X axis is the raw polling average for that combination, and the Y axis is your regression prediction for that combination.

Then, once you have predictions for each set of covariates, "post-stratify" to get a single population estimate by plugging them into the above weighting techniques, where you use the predictions instead of the raw averages in that cell. Report the resulting estimates if you do the 2-dimensional joint weighting (on every pair).

In []:

5) Bonus points (up to 2 points): Implement full "raking" using all the demographic covariates, i.e., match all the marginals without assuming independence, as opposed to just one or two marginal distributions.

Hint: at the heart of raking is calculating a weight for each survey respondent, so that the weights, when summed up, matches the population on desired marginals as much as possible.

For example, suppose we have a survey with two respondents, A likes chocolate and B hates chocolate. We know that in the population, 80% of people like chocolate (the number is made up). So in a representative sample with 2 respondents, we would expect to see 1.6 respondents who like chocolate. Thus, one possible thing to do is we assign A with weight 1.6 and B with weight 0.4, and then use these weights to weigh their answers to the question we care about.

Of course, when the number of respondents and the number of demographic dimensions that we want to match get larger, finding the weights itself becomes harder. There are mainly two ways to do this. The first is through an iterative approach known as Iterative Proportional Fitting (IPF). IPF iterates through each demographic variable at a time, and adjusts the weights of all respondents through post stratification. IPF is easier to implement by hand by just following the instructions. The second approach is known as generalized raking, which is done through solving an optimization problem, and proposed by [Deville, J. C., Särndal, C. E., & Sautory, O. \(1993\)](#). This approach attempts to solve one undesirable outcome of IPF that the resulting weights may be unbalanced, which leads to higher variance in the final estimate. [This blog post](#) gives a nice introduction to generalized raking, while providing a code snippet for implementing it in Python (you will have to make suitable changes, of course.)

In []:

Part C: Uncertainty analysis, choices, and discussion

1) Education weighting analysis and "refused" answers

i. In Part B, you should notice a discrepancy from what we said in class and the data -- weighting by education does *not* seem to help much in reducing the polling average from being pro-Clinton.

Here, we'll try to dig into the data to see why the methods we tried above might not be perfect, and what data you would want (such as demographic joint distribution) to do better.

First, aggregate (using the groupby function) the poll results by education. Second, aggregate by education and some of the other covariates (for example, education and race, or education and party). Discuss in 4 sentences or less.

In []:

ii. You'll notice that there are some responses with "refused," and that those people in particular are Trump-leaning. Furthermore, there are likely many people who refused to answer the poll at all, who do not show up in the data. The weighting techniques we used above would ignore these people. How would you adjust your procedures/estimates above to take them into account? Answer in at most 3 sentences.

In []:

None of the above techniques deal with selection biases/non-response on *un-measured* covariates. Do you think that may be an important concern in this dataset? Why or why not? Respond in 3 or fewer sentences.

In []:

2) Final estimates

Throughout this homework, you made many estimates of the same quantity -- the fraction of people who will vote for Clinton in Florida. Below, plot a histogram of all your estimates.

In []:

Given all your above analysis, if you were a pollster what would you report as your single estimate?

In []:

Justify your choice, in at most 3 sentences

In []:

Though we did not discuss how to calculate margin of error or standard errors with weighting in this course, what would you say if someone asked you how confident you are in your estimate? You may either qualitatively answer, or try to come up with a margin of error.

In []: