Notebook

November 1, 2018

Local date & time is : 11/01/2018 06:54:03 PDT

Deadline: This assignment is due Monday, October 29th at noon (12pm). Late work will not be accepted.

You will submit your solutions using both OKpy and Gradescope. You will find detailed submission instructions at the bottom of this notebook and on bCourses (here). **Please do not remove or add cells and please ignore the '#newpage' cells** (these are here to facilitate Gradescope submission).

You should start early so that you have time to get help if you're stuck. Post questions on Piazza. Check the syllabus for the office hours schedule. Remember that Connector Assistant office hours are for *coding questions only*.

1 newpage

1.1 Question 1: Surviving the Titanic

a. (4 points) Calculate and print separate survival rates for third class passengers and crew. Are crew substantially more likely to survive than third class passengers? [You do not need to conduct a formal hypothesis test.]

```
In [153]: titanic.group('pclass', collect = np.mean).drop(3, 4)

Out[153]: pclass | survived mean | sex mean
          crew   | 0.239548      |
          third  | 0.235489      |
```

Not so substantially. Their survival rate is about the same. The difference in the survival rate does not seem statistically significant.

b. (4 points) Calculate survival rates separately by both `sex` and `pclass`. How do third class passengers and crew compare here?

```
In [154]: #write code here
          t = titanic.group(['pclass', 'sex'], collect = np.mean).drop(3,4)
          t

Out[154]: pclass | sex    | survived mean
          crew   | female | 0.869565
          crew   | male   | 0.222738
          third  | female | 0.476471
          third  | male   | 0.140878
```

It shows that crew female and male passengers have significantly high survival rates than third class male and female passengers. But it's not unusual or suprising.

c. (5 points) How can your findings in **part (a)** and **part (b)** be reconciled?

There is certain possibility to have this kind of discrepancy, given that the the values we are using to find these rates are zeros and ones. When we aggregate these numbers by their class kind (third and crew) and their sex kind, their denominators and numerators change in both areas and this utimately ends up in different answers. It's just arithmetic. This now tells us that, per their ratios there are more female survivors in each respective class. Evidently, it matters what variable or covariate we choose investigate in order to determine association or causality

d. (7 points) Using a matching strategy to control for `sex`, compute and print the difference in survival rates for third class passengers and crew. Conditional on `sex`, were crew substantially more likely to survive than third class passengers?

[To answer this, first match each third class passenger to a randomly selected crew member of the same sex. Then compute the survival rates for third class passengers and the matched crew members. Finally, calculate and print the difference in survival rates between the two groups.]

```
In [156]: #define matching function here
          def match_sex_id(sex_value):
              match = crew.where('sex', sex_value)
              if (match.num_rows >= 1):
                  return np.random.choice(match.column('id')).item(0)
              else:
                  return None

In [157]: #write remaining code here; remember to print differences in survival rates
          matched_third_class =  third_class.apply(match_sex_id, 'sex')
```

```
matched_crew = crew.take(matched_third_class)

matched_crew.group(['pclass'], np.mean)\
.column('survived mean').item(0)-third_class.group(['pclass'], np.mean).column('survived mean
```

```
Out[157]: 0.175787728026534
```

Conditional on sex, crew passengers have higher survival rates in comparison with third class passengers. Based on this, I would think that crew passengers were more likely to survive than third class passengers. It is also worth noting that female crew and third class passengers had higher survival rates than their male counterparts.

2 newpage

2.1 Question 2: Revisiting the Oregon Health Study

a. (3 points) Why does the fact that the lottery winners are disproportionately drawn from larger households complicate the experiment analysis? Why can't we simply compare lottery winners and lottery losers to estimate the average causal effect of winning the lottery on, say, health?

We can't simply do that because, that would be factors that could potentially bias our results. The selection process already creates interference, family demographics, and health history etc are important covariates that can affect our estimate.

b. (5 points) Match each lottery winner to a randomly selected lottery loser with the same value of `household_size`. [Your code may take a while to run. If your notebook appears to be hung up, try restarting your kernel.]

```
In [163]: #write your code here
          def match_size_id(size_value):
              mach = losers.where('household_size', size_value)
              if (mach.num_rows >= 1):
                  return np.random.choice(mach.column('id')).item(0)
              else:
                  return None
```

```
In [164]: #use .apply get array of matched noncontacted voters
          match_indices = winners.apply(match_size_id, 'household_size')

          #use .take and `match_indices` to get a table of matched noncontacted voters
          losers_size_matches = losers.take(match_indices)
          losers_size_matches.show(5)
```

```
<IPython.core.display.HTML object>
```

c. (4 points) As above, check balance between lottery winners and matched lottery losers by comparing average values of `household_size`, `english`, `female`, and `age` for the two groups. Be sure to print the results.

```
In [165]: #Write code for lottery losers here
          losers_size_matches.select('win_lottery','household_size', 'english', 'female', 'age').group(
```

```
Out[165]: win_lottery | household_size mean | english mean | female mean | age mean
          0           | 1.34734             | 0.913989     | 0.584149    | 42.1931
```

```
In [166]: #write code for lottery winners here
          winners.select('win_lottery','household_size', 'english', 'female', 'age').group('win_lottery
```

```
Out[166]: win_lottery | household_size mean | english mean | female mean | age mean
          1           | 1.34734             | 0.911997     | 0.585188    | 42.1612
```

d. (5 points) Test whether the differences in English-speaking rates you measure are statistically significant. Has matching corrected the imbalance?

```
In [168]: #calculate observed difference in english as preferred language for combined
          eng_diff = combined.where('win_lottery', 1).column('english').mean() - combined.where('win_lo
          eng_diff

          #assign array of simulated test statistics under null
          eng_sim_diff = permuted_sample_average_difference(combined, 'english', 'win_lottery', 1000)
```

```
            #calculate and assign p-value
            com_eng_pvalue = np.count_nonzero(abs(eng_sim_diff) >= abs(eng_diff))/1000

            #print results
            print('estimated difference =', eng_diff)
            print('p_value =', com_eng_pvalue)
```

```
estimated difference = -0.0019922044174967013
p_value = 0.607
```

The matching has corrected the imbalance. The p_value supports the null hypothesis, hence the difference is not statistically significant.

e. (5 points) Estimate the average causal effect of winning the lottery on `cost_any_owe` (i.e., the reduced form or Intent to Treat) and conduct a hypothesis test for whether your estimate is statistically significant. [The null hypothesis is that the treatment effect is zero for each participant.] Be sure to print the results.

```
In [169]: #calculate observed difference in cost_owe_any as preferred for combined
            cost_diff = combined.where('win_lottery', 1).column('cost_any_owe').mean() - combined.where('
            cost_diff

            #assign array of simulated test statistics under null
            cost_sim_diff = permuted_sample_average_difference(combined, 'cost_any_owe', 'win_lottery', 1

            #calculate and assign p-value
            cost_pvalue = np.count_nonzero(abs(cost_sim_diff) >= abs(cost_diff))/1000

            #print results
            print('estimated difference =', cost_diff)
            print('p_value =', cost_pvalue)
```

```
estimated difference = -0.05326981377219575
p_value = 0.0
```

f. (3 points) Describe your findings in a complete sentence. Be sure to reference the *meaning* of the variables you're examining rather than just the column names (e.g. don't say '`cost_any_owe` decreases by ...'; instead say what that means in plain English).

The estimate is statistically significant. Which means it does not support the null hypothesis that winning the lottery for participants has zero effect on the medical debt accumulated 12 months after the lottery. Between lottery winners and losers, less proportion of winners had accumulated debt 12 months after.

3 newpage

3.1 Submission

```
In [ ]: from client.api.notebook import Notebook
        ok = Notebook('pset6.ok')
        _ = ok.auth(inline=True)
        _ = ok.submit()
```

```
=====================================================================
Assignment: Pset 6
OK, version v1.12.5
... Omitting 0 lines ...

Successfully logged in as jwonder001@berkeley.edu
```

```
In [ ]: gsExport.generateSubmission('pset6.ipynb')
```

```
In [ ]:
```