ECON626_Lab1

Qingwen Wang

1.1 Law of large numbers

1. μ is 'loc=0', σ^2 is 'scale=1'
2. X_n corresponds to (i) $X_1, X_2, \ldots,$; partial_sums corresponds to (ii) Sn, and partial_means corresponds to (iii) $\frac{1}{n}S_n$.
3. According to the law of large numbers, as a sample size grows, its mean gets closer to the average of the whole population. In this case, the average is set to be zero, as the n increases, the average will be more closer to zero. Thus, the line us more stable at the end.

1.2 Selection bias

1. $E[Y^1 - Y^0] = 0$
2. $NATE = E[Y^1 | D = 1] - E[Y^0 | D = 0] = 0.17217832824951973$
3. $Selectionbias = E[Y^0 | D = 1] - E[Y^0 | D = 0] = 0.17217832824951973$
4. The probability of searching for Doctors Without Borders is a function of the charitability level. People who are more charitable are more likely to search for DWB. D depends on prob_search_dwb and charitability, which means higher charitability, higher probability to search web, and more likely to be in the treatment group.

```
prob_search_dwb = scipy.special.expit(charitability)
D = np.random.binomial(n=1, p=prob_search_dwb)
```

5. $P(D = 1) = 27.92\%$

```
percent = df['D'].value_counts(normalize=True)
percent
0    0.720779
1    0.279221
Name: D, dtype: float64
```

1.3 Data-generating process for a null experiment

1. data_generating process of a true randomized experiment

```
N = 10 ** 6
```

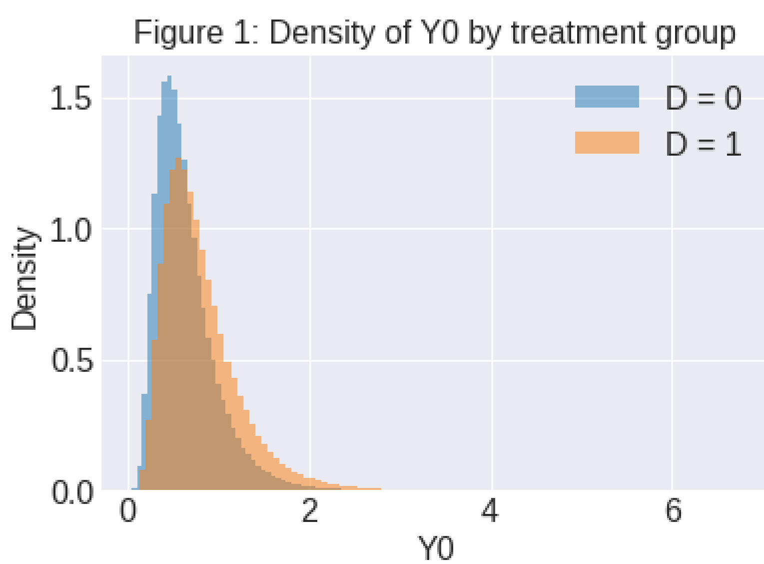
```
charitability = np.random.normal(loc=-1, scale=0.5, size=N)

Y0 = np.random.lognormal(mean=(charitability + 0.5), sigma=0.
1, size=N)

Y1 = Y0.copy()

D_exp = np.random.binomial(n=1, p=0.5, size=N)
```

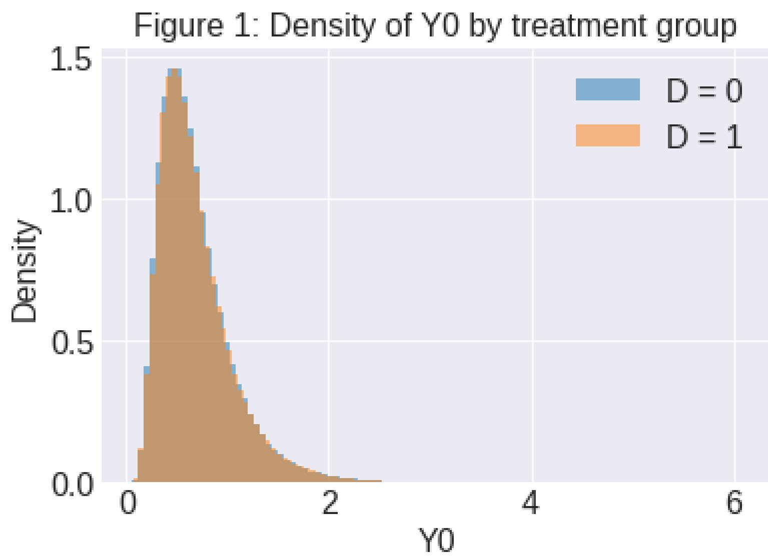
selection bias becomes smaller from 0.17 to -0.00037

$SelectionBias = E[Y^0 | D = 1] - E[Y^0 | D = 0] = -0.000372001236214281 05$

$NATE = E[Y^1 | D = 1] - E[Y^0 | D = 0] = -0.000372001236214281 05$

2. In the figure, the orange shade is more right skewed than the blue shade. People in the treatment group have a higher probability to donate than the people in the control group even they are not exposed to the ads. Thus, we can not say $(Y^1, Y^0) \perp D$. This is not a true randomized experiment.
3. The overlap shade in picture2 becomes larger than that in picture1.
   the original figure(p1): the density of Y^0 when D=1 is not same with the density of Y^0 when D=0; which means these two groups do not have the same probability to donate even there is no treatment.
   the new figure(p2): The shape of the density of Y^0 when D=1 and the density of Y^0 when D=0 are the same; they are almost perfectly overlapped; which means they have the same probability to donate if there is no treatment. This is a true randomized experiment.





1.4 Data-generating process for an experiment with a treatment effect

1. Make a version of the biased data-generating process with an ATE=0.3, by changing the 4th line

```
N = 10 ** 6
charitability = np.random.normal(loc=-1, scale=0.5, size=N)
Y0 = np.random.lognormal(mean=(charitability + 0.5), sigma=0.
1, size=N)
# original Y1 = Y0.copy()
Y1 = Y0 + 0.3
prob_search_dwb = scipy.special.expit(charitability)
D_exp = np.random.binomial(n=1, p=prob_search_dwb)
```

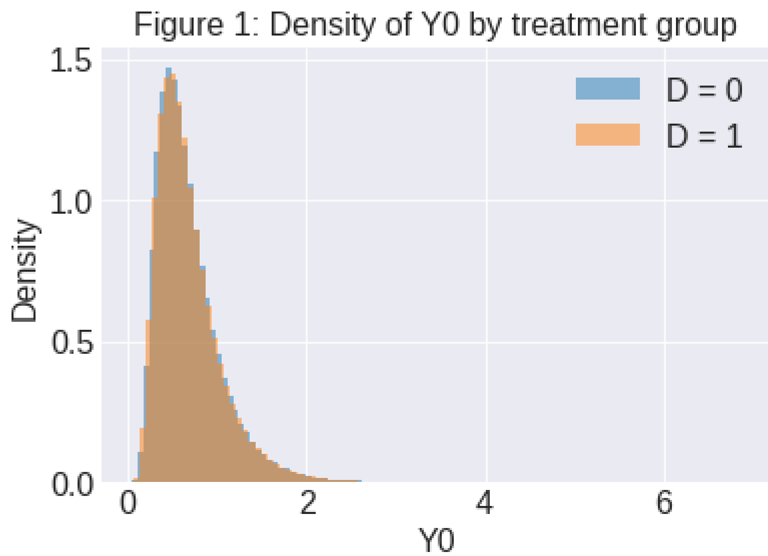
$NATE = E[Y^1 | D = 1] - E[Y^0 | D = 0] = 0.4717684935948674$

2.

```
N = 10 ** 6
charitability = np.random.normal(loc=-1, scale=0.5, size=N)
Y0 = np.random.lognormal(mean=(charitability + 0.5), sigma=0.
1, size=N)
Y1 = Y0 + 0.3
```

```
D_exp = np.random.binomial(n=1, p=0.5, size=N)
```

$NATE = E[Y^1 | D = 1] - E[Y^0 | D = 0] = 0.3011791308830063$



2 Causality warm-up

1. Selection bias is induced by preferential selection of units; selection bias exists when there is systematic difference in treatment group and control group. i.e. people who are more rich are more likely to buy luxury jewelries whether or not they are informed of promotion activities than people who has financial problems.
   $SelectionBias = E[Y0 | Treatmentgroup] - E[Y0 | Controlgroup]$, when $SelectionBias! = 0$, means there exists selection bias that the treatment group and control group will still behave differently even there is no treatment. Selection bias is a problem because when selection bias exists, it is difficult to observe the true treatment effect and the experiment results maybe totally wrong.
2. reasons selection bias may occur: 1. people choose it 2. treatment access is limited
3. P(X∈A,Y∈B)=P(X∈A)P(Y∈B) A random variable that doesn't have an effect on the other random variables in the experiment, which means receiving information about one random variable does not change our assessment of the probability distribution of the other random variables.
4. In a randomized experiments, participants are randomly assigned to the treatment or control groups so that there is no relationship between(Y^1, Y^0) and

D. Since independence assumption holds, $E[Y^0 | D = 0, 1] = E[Y^0] \rightarrow$ $E[Y^0 | D = 1] - E[Y^0 | D = 0] = 0 \rightarrow SelectionBias = 0 \rightarrow$ Experiments do remove selection bias.

3 Critical Reading

1. The study I am interested in is “An fMRI study of error monitoring in Montessori and traditionally-schooled children”. It discusses which kind of students learn better over time, the students who are focused on monitoring their own learning process, or the students who are focused on getting right answers.
   https://rossier.usc.edu/files/2020/07/fmri-study-error-monitoring.pdf
2. The unit in this study is 8-12 year-old students from Montessori schools and students educated in traditional schools.
3. Students in the study were asked to solve math problems while an fMRI tracked brain activity. The researchers evaluates potential learning outcomes by both brain activity and students’ quiz answer. Only Montessori students showed coherent changes in brain activity following errors, suggesting that they were engaging with the errors strategically to learn. Traditionally-schooled students, by contrast, showed coherent activity only after correct answers, and the activity pattern suggested that they were trying to memorize that event. Though both groups got the same number of problems right, the Montessori students skipped far fewer and got more wrong, making them learn the task more efficiently by the end.
   Y^1: the potential learning outcomes for students from a Montessori school
   Y^0: the potential learning outcomes for students from a traditional school
   D: learning knowledge by being focused on monitoring their own learning process, instead of being focused on getting right answers
4. The treatment effect was characterized by both brain activity and students’ quiz answer results. Compared results can be found in fig.2, fig.3, fig.4 under the the results part of the article.
5. the treatments in this study mostly assigned by the parents, the children’s parents decides what kind of school, Montessori or traditional, their children enter.
6. Family’s economic status can be a factor causing selection bias. Since Montessori school usually has a higher tuition fee than the tradition public school so that family with a worse economic status are less likely to send their children to a Montessori school. So this selection bias will cause the potential learning outcomes to be better, ATE to be higher.

7. To control the selection bias of the study, I will probably do some survey about the participates' family incomes , parents' education level, and participates' study environment to make sure the selected participates have similar background.
8. This study is informative and give good points to the questions I am curious about. In my home country, most primary school students have a lot of homework to do everyday. They are evaluated only by exam scores. I think there may exist a better educational pattern.

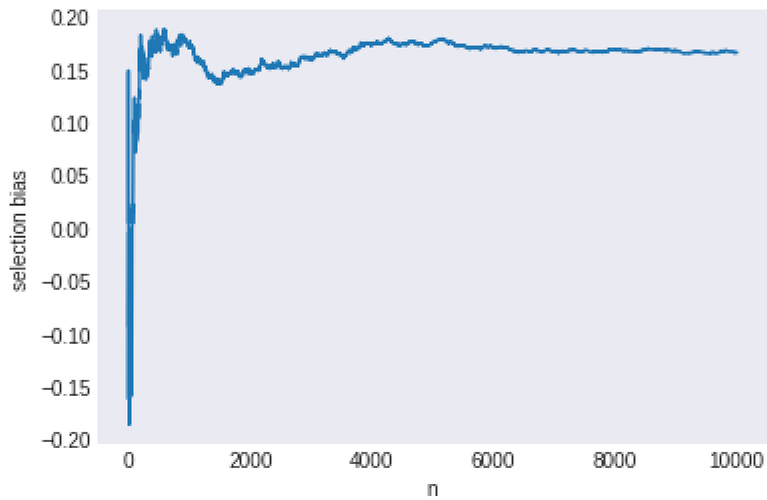
4 Selection bias vs. big data

1. $NATE = ATE + SelectionBias + Differentialeffectbias$
   $NATE = E[Y^1 | D = 1] - E[Y^0 | D = 0] = 0.17217832824951973$
   $ATE = E[Y^1 - Y^0] = 0$
   $SelectionBias = 0.17217832824951973$
   $Differentialeffectbias = 0$

2.

```
def selection_bias(n):
  df_pt = df[:n]
  bias = df_pt.query("D == 1")['Y0'].mean() - df_pt.query("D =
= 0")['Y0'].mean()
  return bias
```

```
ls = []
for i in range(0, 10000):
  ls.append(selection_bias(i))
df = pd.Series(ls)
ax = df.plot(legend=False)
ax.set_xlabel("n")
ax.set_ylabel("selection bias")
```



The line is converging to 0.17. The conclusion is that the law of large numbers applies to selection bias, however, big data will not remove selection bias.

Bonus 1: Why is it called identification

1. $NATE = E[Y^1|D=1] - E[Y^0|D=0]$, from the formulation, we can find that the data we need to calculate the NATE can be observed from the dataset.
2. No, we can not directly estimate these items.
3. ATE, Selection Bias, Differential effect bias are unknowns. The equation will have 3 solutions: (1) NATE = ATE + Selection Bias ; (2) NATE = ATE + Differential effect bias; (3) NATE = ATE + Selection Bias + Differential effect bias
4. if the independence assumption holds, we will get $NATE = ATE$, because independence assumption eliminates both selection bias and differential effect bias.
5. Including the independence assumption, we have 4 equations. ATE = NATE - (Selection Bias) - (Differential effect bias) can have 3 possible solutions. The relationship between ATE and NATE may not be one-to-one relationship, it depends on whether the independence assumption holds or not.

Bonus 2

Prove that NATE = ATT + selection bias:

$NATE = E[Y^1|D=1] - E[Y^0|D=0]$

add and subtract the counterfactual term to the right hand side

$NATE = E[Y^1|D=1] - E[Y^0|D=0] + E[Y^0|D=1] - E[Y^0|D=1]$

$NATE = (E[Y^1|D=1] - E[Y^0|D=1]) + (E[Y^0|D=1] - E[Y^0|D=0])$

Thus,

$$NATE = ATT + SelectionBias$$