W20 Causal Analysis

May 1st notes

Don't panic id it is hard, grade will be curved base on class performance

Go through each question

Tues Thur Friday office hour

Tomorrow TA office hour/optional review (5-6:30)

Causal Inference Analysis :

> How change in one variable affect the other (how much change occur in the other)

- ***correlation does not mean causal (ex;*** direct impact in policy recommendation)
- **A/B Testing**
  - version a and b of V1 for A and B group
    - control group (old version of V1)
    - test group (new version of V1)
  - make sure assignment to each group is random → get rid of confounding factors
    - factor that affect both no longer in effect (seasonal change affect ice cream eating but in this case we assign the ice cream amount for each group)

  ▼ **Example : Game**

    - Gates at certain level
    - assignment to Design 1 and 2 base on time of download

- Design 1: Set the first gate at Level 40 (current design)
- Design 2: Set the first gate at Level 30 (proposal design)

In the data, the following features are collected

Field	Explanation
userid	player identifier
version	whether the player will encounter gate at level 30 or 40
sum_gamerounds	the number of game rounds played by the player during the first 14 days after install.
retention_1	did the player come back and play 1 day after installing?
retention_7	did the player come back and play 7 days after installing?

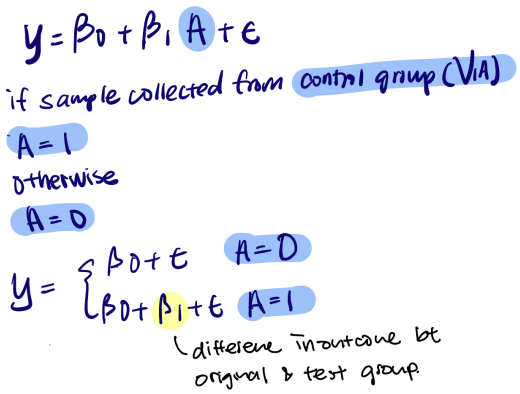
Is there any difference in the average of sum_gamerounds between the two versions?

Is there any difference in the 1-day retention rate between two versions?

Is there any difference in the 7-day retention rate between the two versions?

	userid	sum_gamerounds	retention_1	retention_
gate_30	4.987564e+06	52.456264	0.448188	0.190201
gate_40	5.009073e+06	51.298776	0.442283	0.182000

- **Regression** → **Analyze Result** (difference in outcome bt two groups)
  - **beta1** tells difference between group 0 and 1



- How much **variation/Uncertainty** do we expect round by round (repeat experiment)
  - **Robustness and consistency**
  - ***Regression*** BOOTSTRAPPING
    - `Pandas.get_dummies(data,columns=[" "], drop_first=False)`
    - original dataset N samples :(do this k times) 1. resample N sample (with replacement) 2. reestimate 3. find *α/2 and 1- α/2 percentile*
      - *α percent of estimated β1 fall into the interval*
      - 95% confidence interval, find 2.5 and 97.5 percentile
    - if 0 (no difference in y between 2 groups) does not fall into *α* →

  there is a difference in the *y* value due to _change in variable-.

▼ Difference in co-effecient between 2 group of version

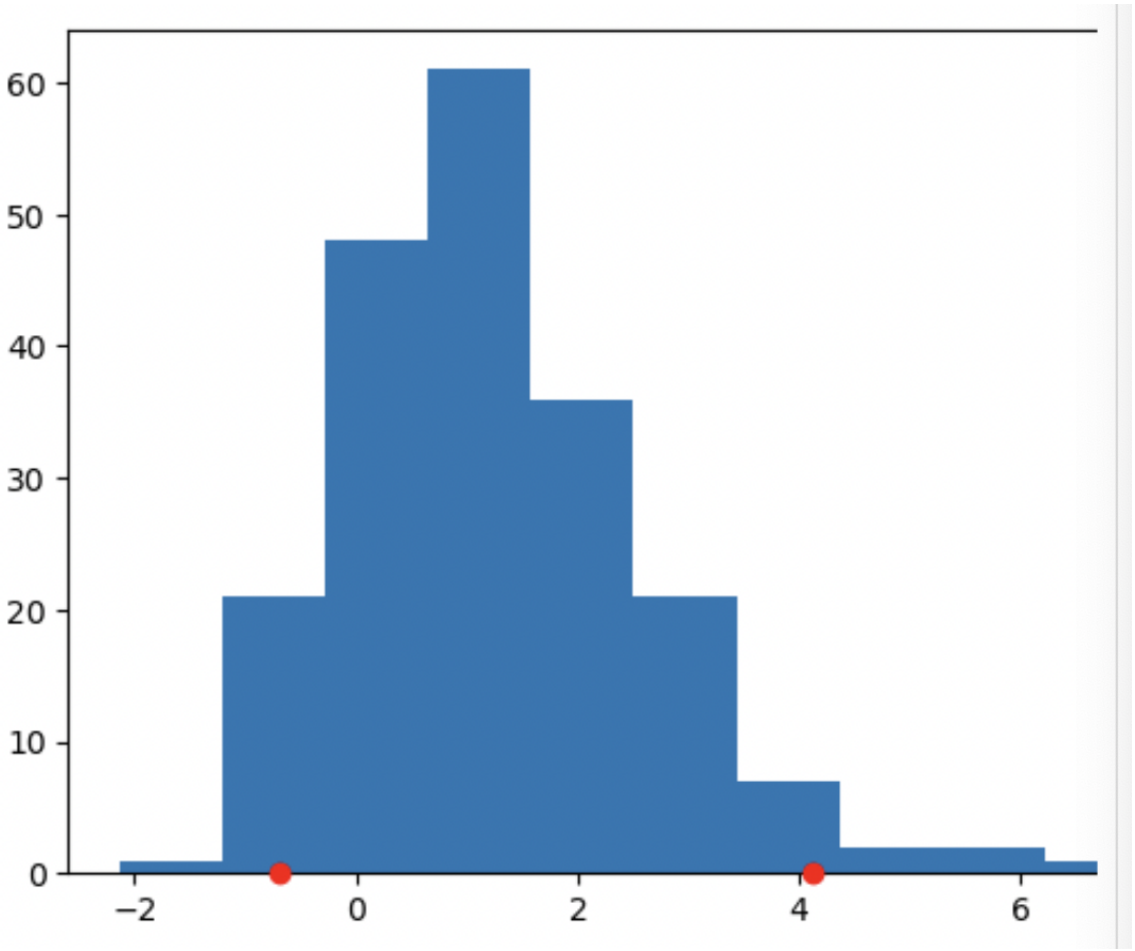
- Run function 200 rounds and creat a list to store each round

```
#Version 1
def bootrap():
    ccat1 = ccat.sample(n = ccat.shape[0], replace=True)
    # replace  = True allow us to construct variation in information
       # the item picked is equally likely to be picked
    model.fit(ccat1[["version_gate_30"]], ccat1["sum_gamerounds"])
    return model.coef_[0]
```

```
# Version 2
measurement="sum_gamerounds"
Model=LinearRegression()
def simulation(measurement):
    X=ccat.sample(replace=True, n=ccat.shape[0])
    return  Model.fit(X[["version_gate_40"]], X[measurement]).coef_[0]
simulation(measurement)

#Run function 200 rounds and creat a list to store each round
results = [bootrap() for i in range(200)]
plt.hist(results)
plt.show()

# distribution of Beta value collected


#95% confidence interval, chop top and tail
## show the dots marking 95% interval


plt.scatter(np.percentile(results,[2.5,97.5]),[0,0],zorder=3,clip_on=False,color="red")
```

95% time the beta is between the dots

not conclusive whether the two groups have real difference because 0 is inside the interval


60
50
40
30
20
10
0
−2
0
2
4
6