UNIVERSITY OF SAN FRANCISCO
CHANGE THE WORLD FROM HERE

MSDS 601 Final Presentation

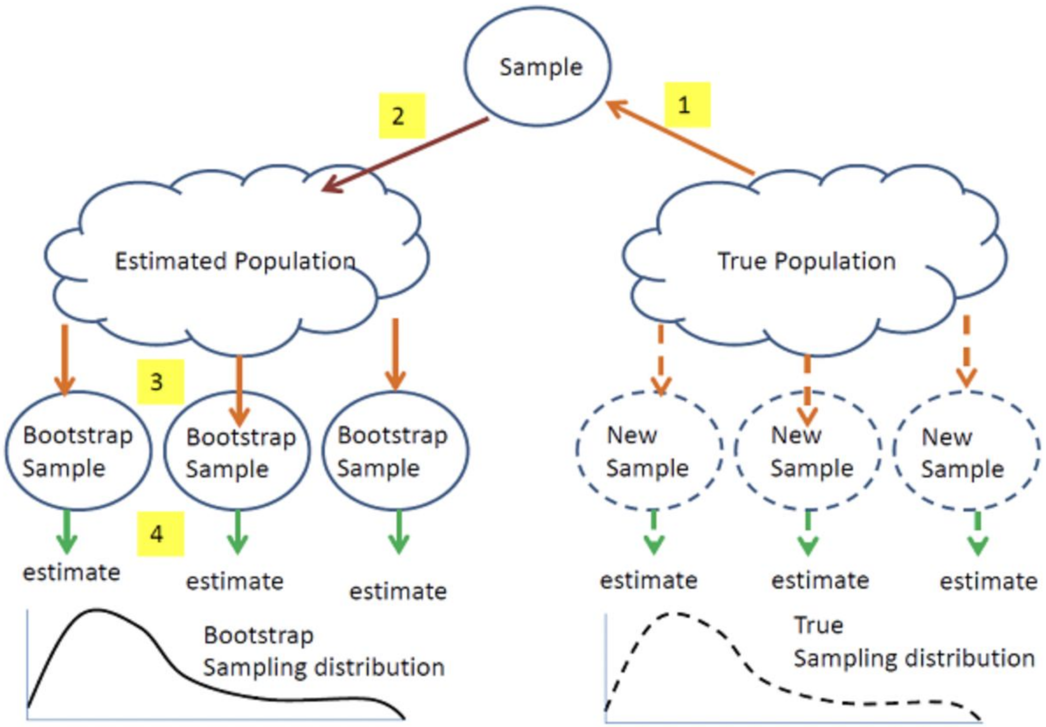
Bootstrapping Interview Guide

Jazz Sun, Harrison Yu

Interview Question-1

Could you briefly introduce bootstrapping concept in 1 minute?

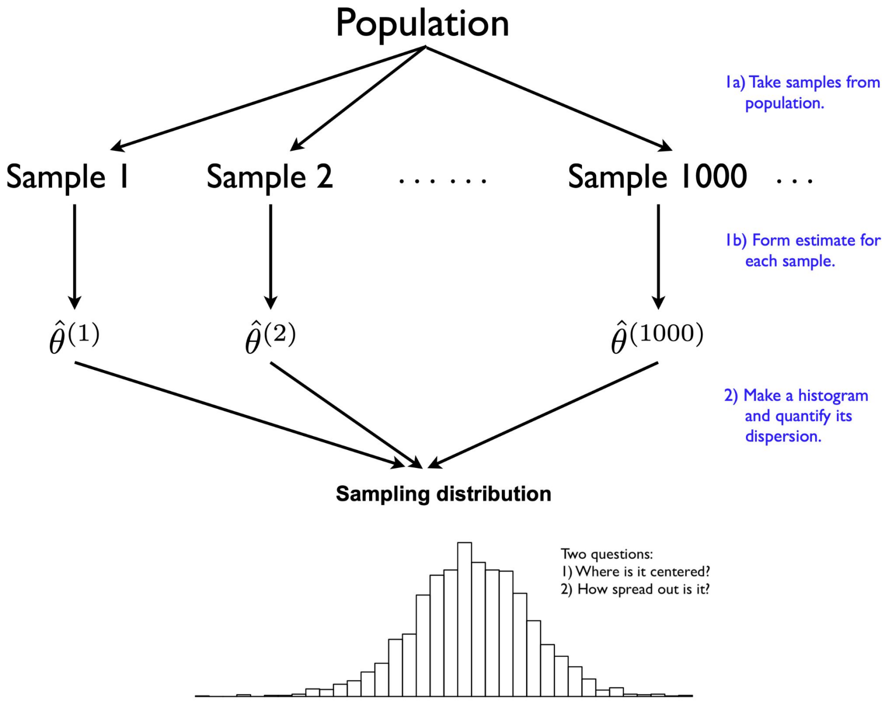
To sample with replacement to simulate population

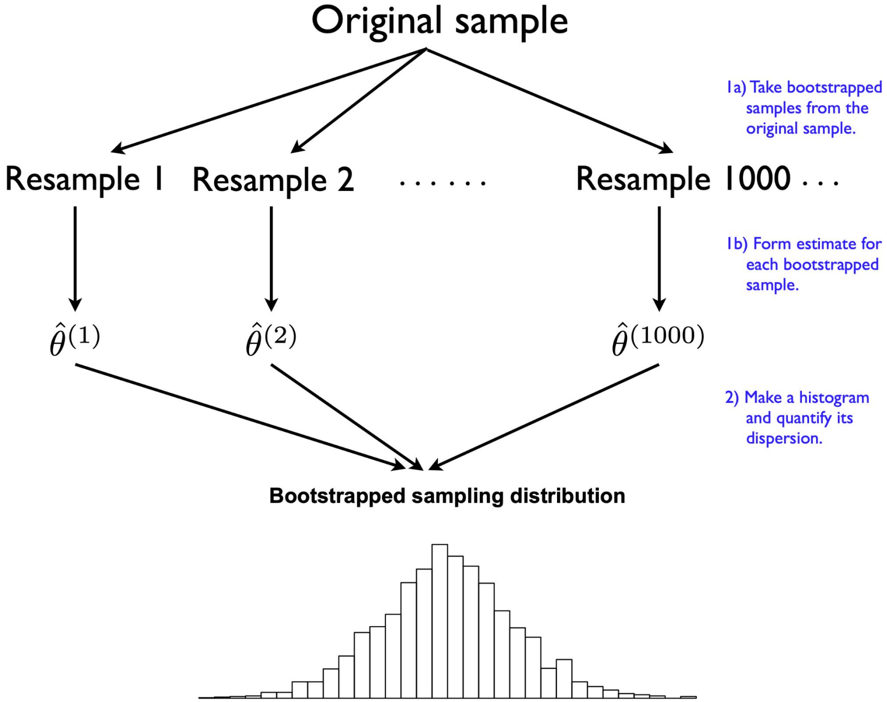


Interview Question-2

What are the difference between sampling distribution and bootstrapping distribution?

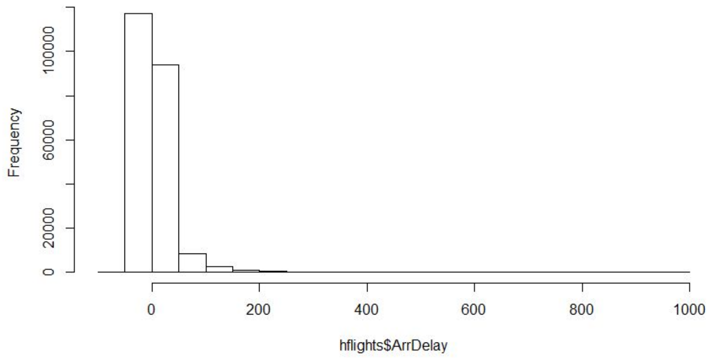
Bootstrapped distribution came from original sample, while sampling distribution came from population

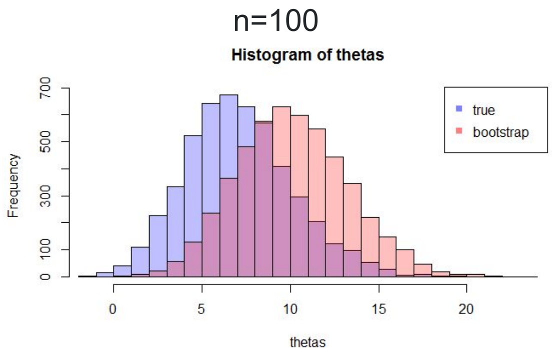


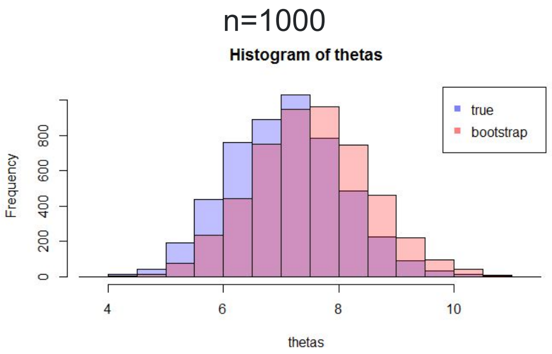


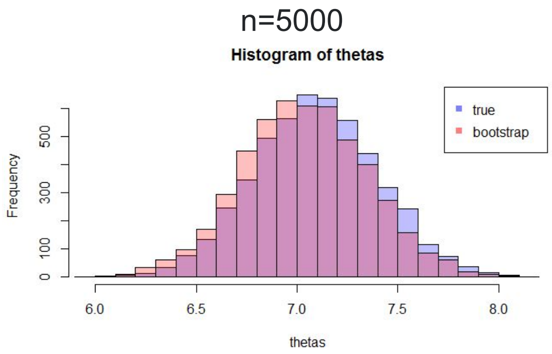
https://towardsdatascience.com/linear-regression-with-bootstrapping-4924c05d2a9

Bootstrapped distribution approximate sampling distribution as n gets larger, depended on bootstrap sample we get









https://towardsdatascience.com/linear-regression-with-bootstrapping-4924c05d2a9

Interview Question-3

What are pros and cons of bootstrapping?

Pros and Cons

Pros

1. Resolve resource limitation
2. Work with any population distribution

Cons

1. Excessive computing power
2. Rely on sample quality

Interview Question-4

Tell me about how bootstrapping can be applied in linear regression models

Tell me about how bootstrapping can be applied in linear regression models

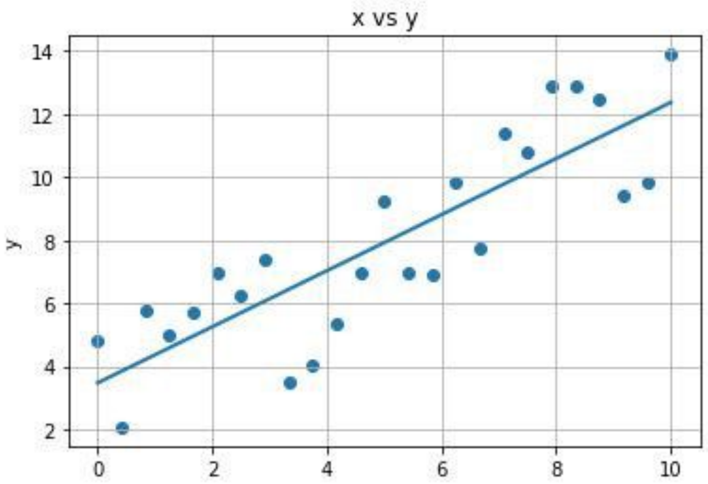
- Bootstrapping is a nonparametric approach to statistical inference that gives us standard errors and confidence intervals for our parameters

- Bootstrap can be applied to regression models to give insights into variability of our parameters (beta) with minimal assumption about parent distribution

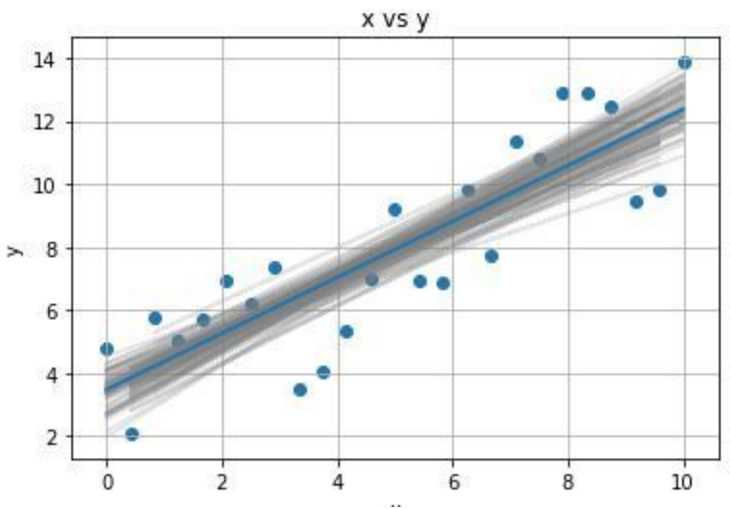
- Parametric bootstrapping — resampling from all of the points (X):
  1. Sample the data with replacement numerous times (100)
  2. Fit a linear regression to each sample
  3. Store the coefficients (intercept and slopes)
  4. Plot a histogram of the parameters
  5. Make inferences about true parameters

Tell me about how bootstrapping can be applied in linear regression models

Best fit line with OLS



100 Best fit lines with Bootstrapping sample data

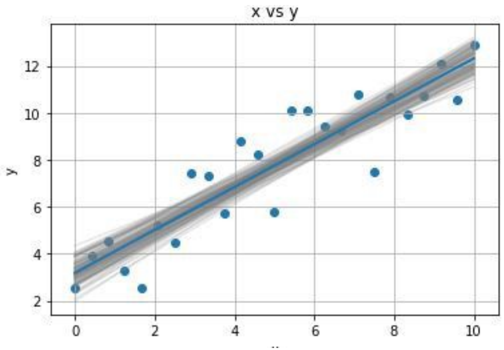


Problem: bootstrap on X treats X as random rather than fixed. Also might not work on sparse data

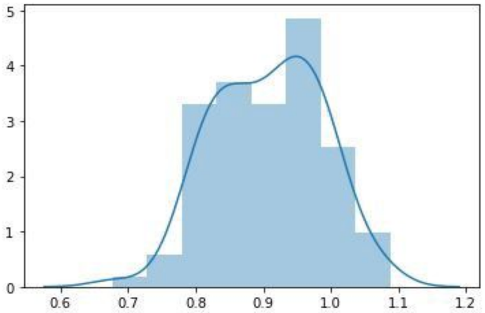
Can also do non-parametric bootstrap (on residuals) for sparse data to avoid outliers being sampled multiple times

Implicit Assumption: errors are IID

1. Find the optimal linear regression on all the original data
2. Extract the residuals from the fit
3. Create new y-values using the residual samples
4. Fit the linear regression with the new y-values
5. Store the slope and intercepts
6. Plot a histogram of the parameters
7. Make inferences about true parameters



Histogram of slopes



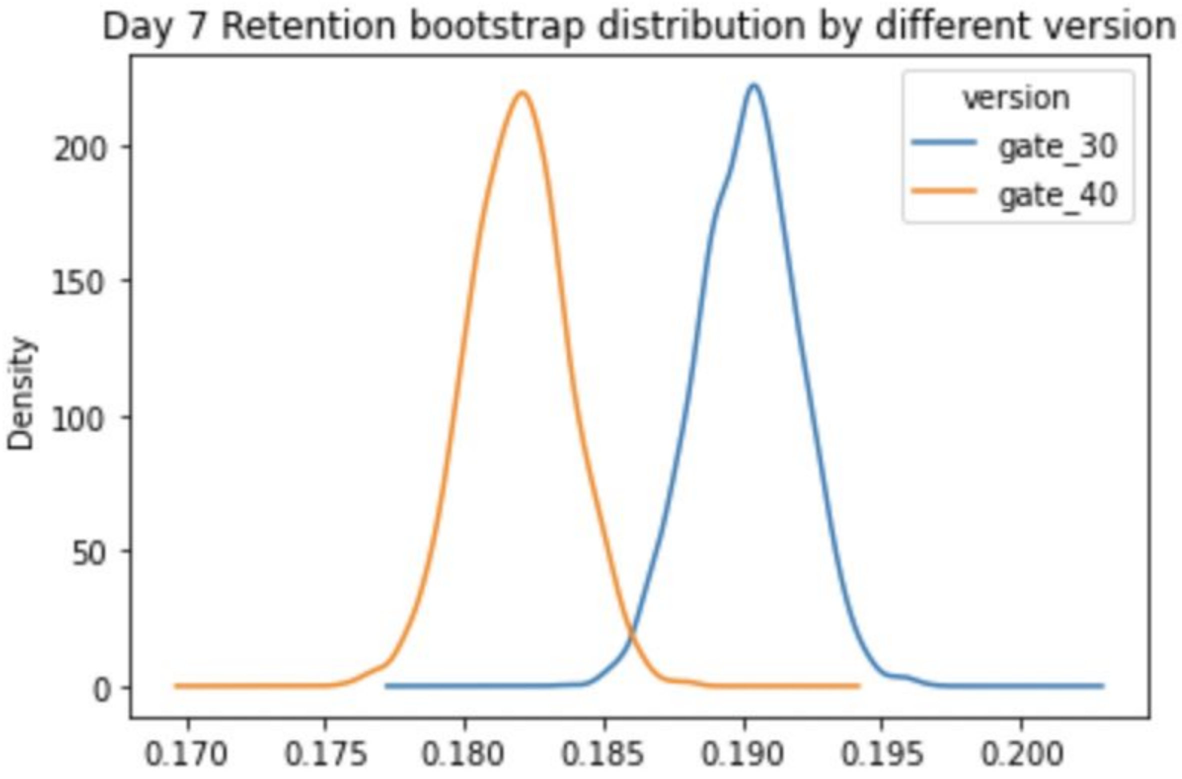
Interview Question-5

Tell me why and how to apply bootstrapping on AB Testing?

Hypothesis:

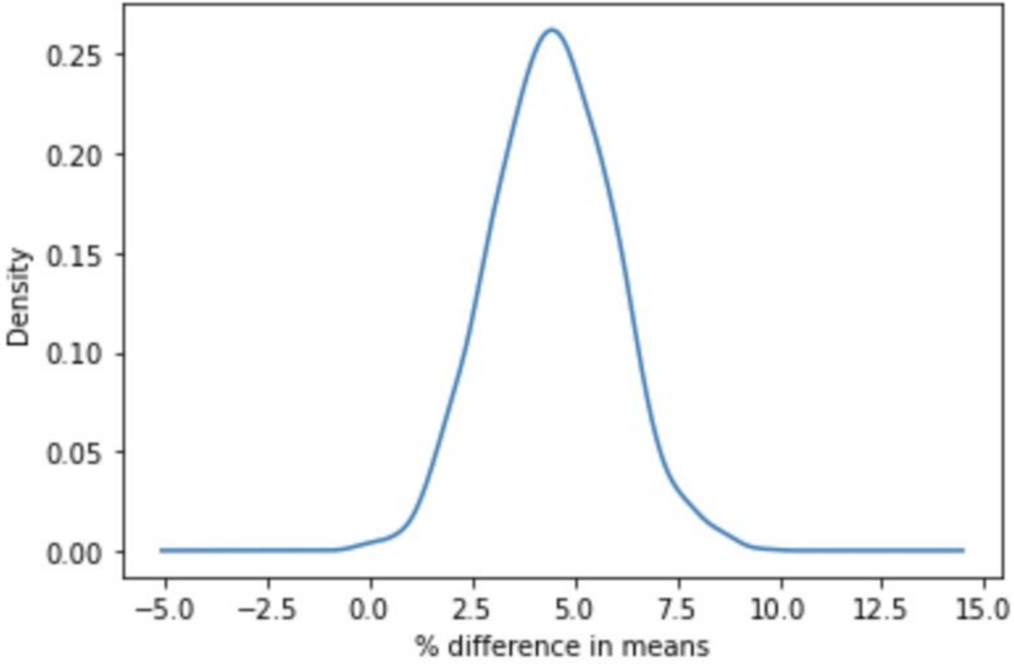
Where we should put the first gate? Level 30 vs. Level 40

Day 7 retention rate bootstrap distribution by Level 30 vs. Level 40



The probability that 7-day retention is greater when the gate is at level 30-99.9%

'99.9%'



Bootstrapping can make your stat life easier :)

UNIVERSITY OF SAN FRANCISCO
CHANGE THE WORLD FROM HERE

APPENDIX

Why we should apply bootstrapping on AB Testing?

The good thing about using bootstrapping is that I don't need to assume the distribution of the data, or consider if the sample size is large enough.

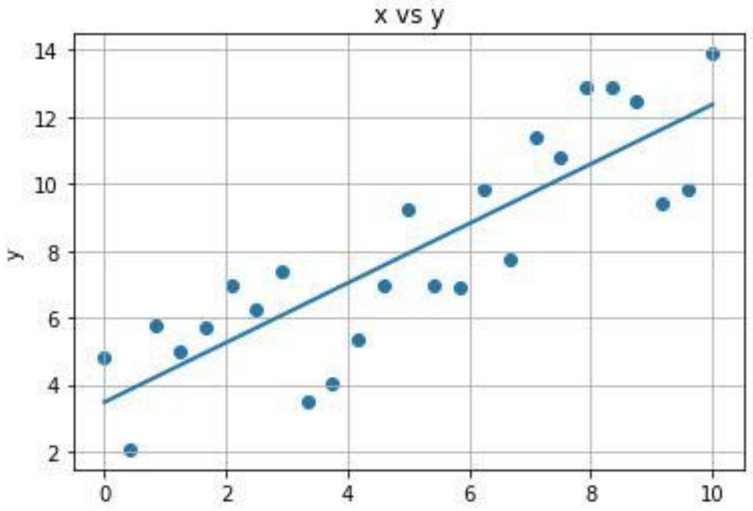
$$\widehat{\text{Prestige}} = \underset{(3.588)}{-7.289} + \underset{(0.1005)}{0.7104} \text{ Income} + \underset{(0.0825)}{0.4819} \text{ Education}$$

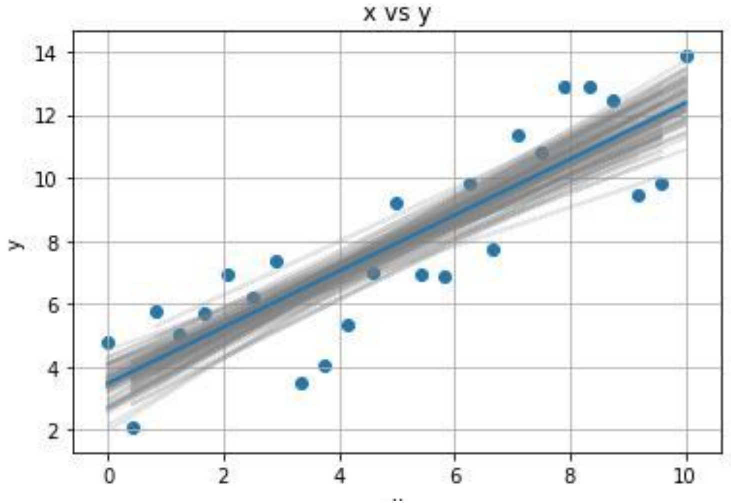
Table 21.5 Statistics for $r = 2,000$ Bootstrapped Huber Regressions Applied to Duncan's Occupational Prestige Data

	Coefficient		
	Constant	*Income*	*Education*
Average bootstrap estimate	−7.001	0.6903	0.4918
Bootstrap standard error	3.165	0.1798	0.1417
Asymptotic standard error	3.588	0.1005	0.0825
Normal-theory interval	(−13.423,−1.018)	(0.3603,1.0650)	(0.2013,0.7569)
Percentile interval	(−13.150,−0.577)	(0.3205,1.0331)	(0.2030,0.7852)
Adjusted percentile interval	(−12.935,−0.361)	(0.2421,0.9575)	(0.2511,0.8356)

NOTES: Three bootstrap confidence intervals are shown for each coefficient. Asymptotic standard errors are also shown for comparison.

Tell me about bootstrapping and regression models





Day 7 retention rate bootstrap distribution-code

```
#Day 7 retention
# Creating an list with bootstrapped means for each AB-group
boot_7d = []
for i in range(2000):
    boot_mean = df.sample(frac=1,replace=True).groupby('version')['retention_7'].mean()

    boot_7d.append(boot_mean)

# Transforming the list to a DataFrame
boot_7d = pd.DataFrame(boot_7d)

# A Kernel Density Estimate plot of the bootstrap distributions
boot_7d.plot(kind='kde',title="Day 7 Retention bootstrap distribution by different version")
```

The probability that 7-day retention is greater when the gate is at level 30-code

```
# Creating a list with bootstrapped means for each AB-group
boot_7d = []
for i in range(2000):
    boot_mean = df.sample(frac=1,replace=True).groupby('version')['retention_7'].mean()
    boot_7d.append(boot_mean)

# Transforming the list to a DataFrame
boot_7d = pd.DataFrame(boot_7d)

# Adding a column with the % difference between the two AB-groups
boot_7d['diff'] = (
    (boot_7d['gate_30']-boot_7d['gate_40'])/
             boot_7d['gate_40']*100)

# Ploting the bootstrap % difference
ax = boot_7d['diff'].plot(kind='kde')
ax.set_xlabel("% difference in means")

# Calculating the probability that 7-day retention is greater when the gate is at level 30
prob = (boot_7d['diff']>0).sum()/len(boot_7d['diff'])

# Pretty printing the probability
'{:.1%}'.format(prob)
```