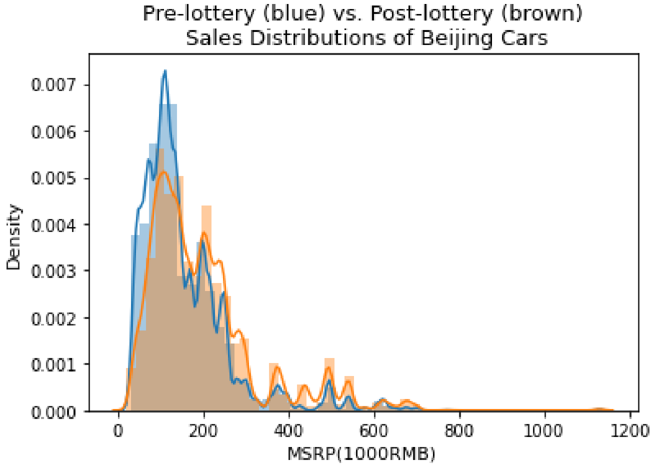
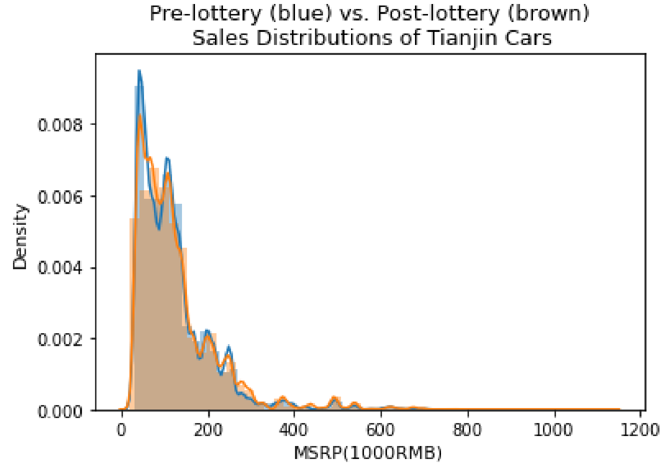
Jayoung Kang

Minilab 2

Exercise 3.2

a. *Overlay the histograms that describe the 2010 and 2011 distribution of Tianjin car sales. Be sure to normalize the histograms so the area of the bars in each histogram sum to 1.*





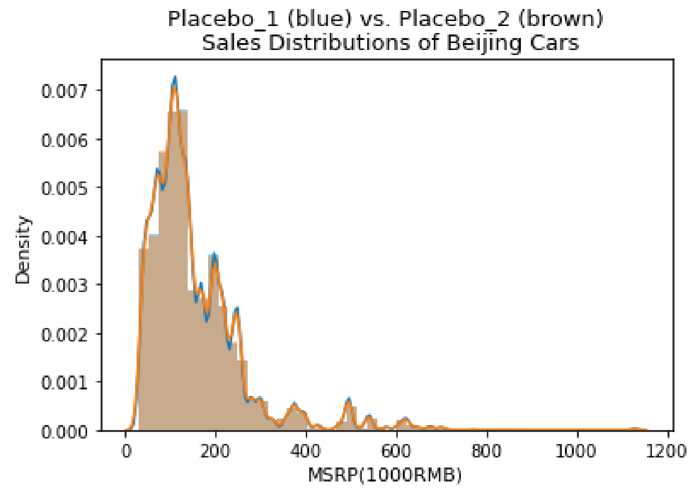
b. *Compare and contrast the shift between the Beijing distributions with the shift between the Tianjin distributions. Based on the shift in Tianjin car sales, should we be surprised to see the shift in Beijing car sales?*
Compared to the shift in the Beijing distributions pre and post-lottery, the Tianjin distributions did not shift as much during the same time period. The Beijing car sales for post-lottery distribution have shifted to the right slightly with fewer proportion of the sales distributed at the lower MSRP of around 100,000 RM when compared to the distribution of pre-lottery. A higher proportion of sales are observed around the MSRP of 200,000 RMB in Beijing post-lottery. However, for the Tianjin car sales distribution, although there is a slight shift in the post-lottery period where the proportion of sales of very cheap cars decreased compared to the pre period, this shift is very small compared to that of Beijing. Also, we don't see that high of an increase in the proportion of sales around 200,000 RMB as we did for Beijing.
We should not be surprised to see the upward shift in Beijing car sales compared to the relatively smaller shift in Tianjin car sales because it points to the possibility of the black market for license plates, which has led to a larger share of wealthy individuals to obtain license plates and purchase more expensive cars, compared to Tianjin, the control group.

Exercise 3.3

a. *Run the preceding code block so you have access to placebo_1.*

b. *Use rmultinom to sample observations from Beijing_pre. Store the resulting data frame in placebo_2. Be careful to draw the correct number of observations.*

c. *Compare placebo_1 and placebo_2. Do they appear to be drawn from the same distribution?*
The distribution of the two are almost identical, making it appear that they are drawn from the same distribution.

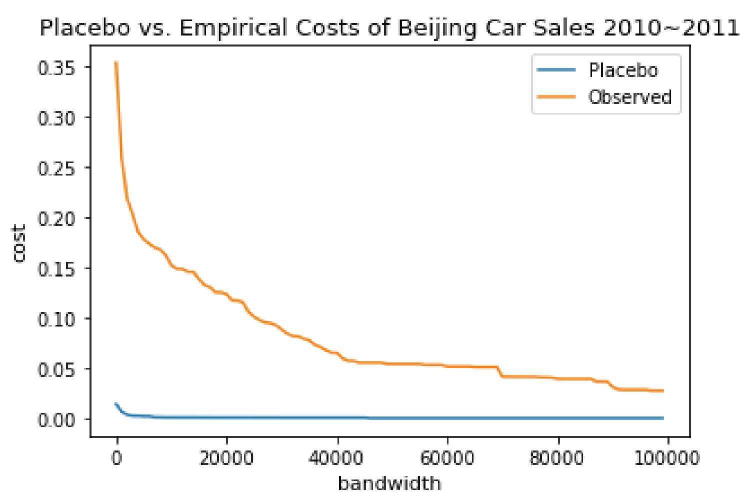


Exercise 3.4

a. *Compute the transport cost between the two placebo distributions for different values of d from 0 to 100,000.*

b. *For the same values of d, compute the transport cost between the observed distributions for 2010 and 2011 Beijing car sales.*

c. *Plot the placebo costs and the empirical costs obtained in the previous two steps with the bandwidth as the x-axis.*



d. *For which values of d is the placebo cost less than 0.05%?*
The values of d equal or above 25000 have placebo cost less than 0.05%

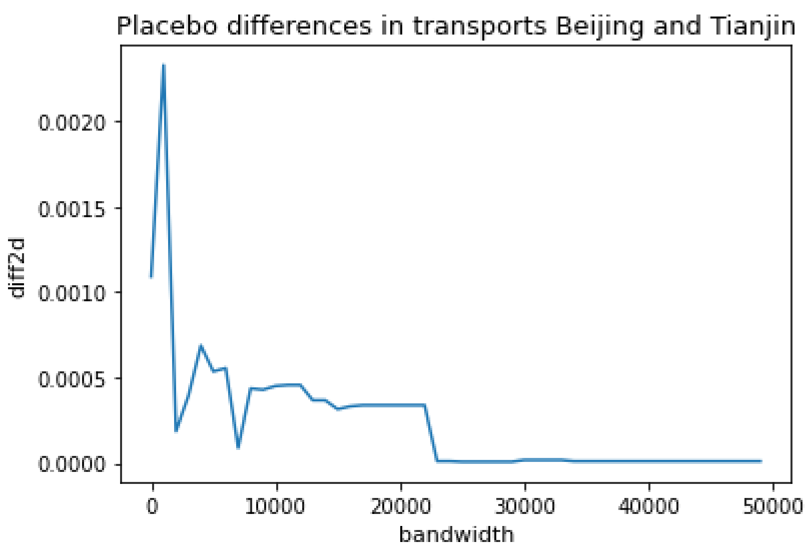
```
In [53]: lessthan05 = placebo_at_100k[placebo_at_100k['main'] < 0.0005]
    ...: lessthan05.head()
Out[53]:
    bandwidth      main
26      25000  0.000438
27      26000  0.000438
28      27000  0.000438
29      28000  0.000438
30      29000  0.000438
```

e. *For the smallest value of d found in the previous step, what is the empirical transport cost? This estimate for the lower bound on the volume of black market transactions is what we call the before-and-after estimate*
The empirical transport cost at bandwidth 25000 is 0.000438

```
In [52]: print(observed_at_100k[observed_at_100k['bandwidth'] == 25000])
    bandwidth      main
26      25000  0.100769
```

Exercise 3.5

g. *Create a plot of the absolute value of the placebo differences-in-transports estimator on the y-axis and the bandwidth on the x-axis.*



h. *For which values of d does the absolute value of the placebo differences-in-transports estimator stay below 0.05%? Note that the absolute difference is not a monotonically decreasing object, so this difference may even increase as we increase the bandwidth. Temporary increases above the 0.05% threshold can be ignored.*
The values of d in which the absolute value of the placebo differences in transports stay below 0 are 2000, 3000 and 7000 and above. Like the question noted, the absolute difference is not monotonically decreasing so, that is why the bandwidths between 3000 and 7000 do not have the absolute value below 0.05%.

```
In [51]: lessthan05_2.head()
Out[51]:
    bandwidth      main    main2d   control      diff    diff2d
3        2000  0.003298  0.002683  0.002496  0.000802  0.000186
4        3000  0.002869  0.001235  0.001627  0.001242 -0.000392
8        7000  0.000753  0.000578  0.000489  0.000264  0.000089
9        8000  0.000731  0.000564  0.000129  0.000601  0.000435
10       9000  0.000650  0.000556  0.000128  0.000521  0.000427
```

i. *Among all the values of d that you found in the previous step, which one yielded the largest value of (3) from part a? This is the difference-in-transports estimator.*
The one with d=7000 yielded the largest value of the difference in transports estimator.

```
In [50]: print(dit_at_50k[dit_at_50k.diff2d == dit_at_50k.diff2d.max()])
   bandwidth      main    main2d   control      diff    diff2d
8       7000  0.169615  0.145225  0.025194  0.144421  0.120031
```