

POLS 7012 Final Exam 2022

Joe Ornstein

Due December 14, 2022 @ 5pm

[Pereira & Fernandez-Vazquez \(2022\)](#) investigate whether electing women to public office reduces corruption on Spanish municipal councils, taking advantage of a population-based cutoff that Spain used to enforce gender quotas. You can find a repository of the study's data and code [here](#). For the final exam, we will replicate and extend some of their findings. Please write and submit a PDF report rendered from an R script (optionally: a Quarto document) that conducts the following analyses:¹

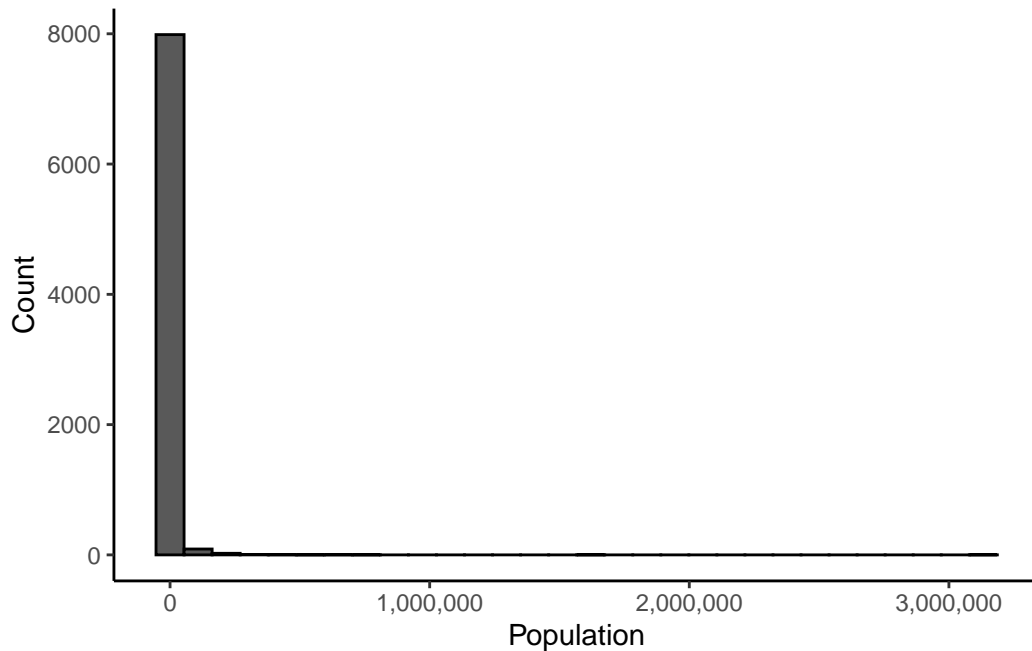
1. Plot a histogram of the Spanish city population distribution in 2007. What is the mean, median, and standard deviation of the distribution? Why are the mean and median so different?

```
library(tidyverse)

m07 <- read_csv('m07.csv')

ggplot(data = m07,
       mapping = aes(x=pop)) +
  geom_histogram(color = 'black') +
  theme_classic() +
  labs(x = 'Population',
       y = 'Count') +
  scale_x_continuous(labels = scales::comma_format())
```

¹Unlike the problem sets, this exam must be completed individually, without help from others.



```
m07 |>
  summarize(mean(pop),
            median(pop),
            sd(pop))
```

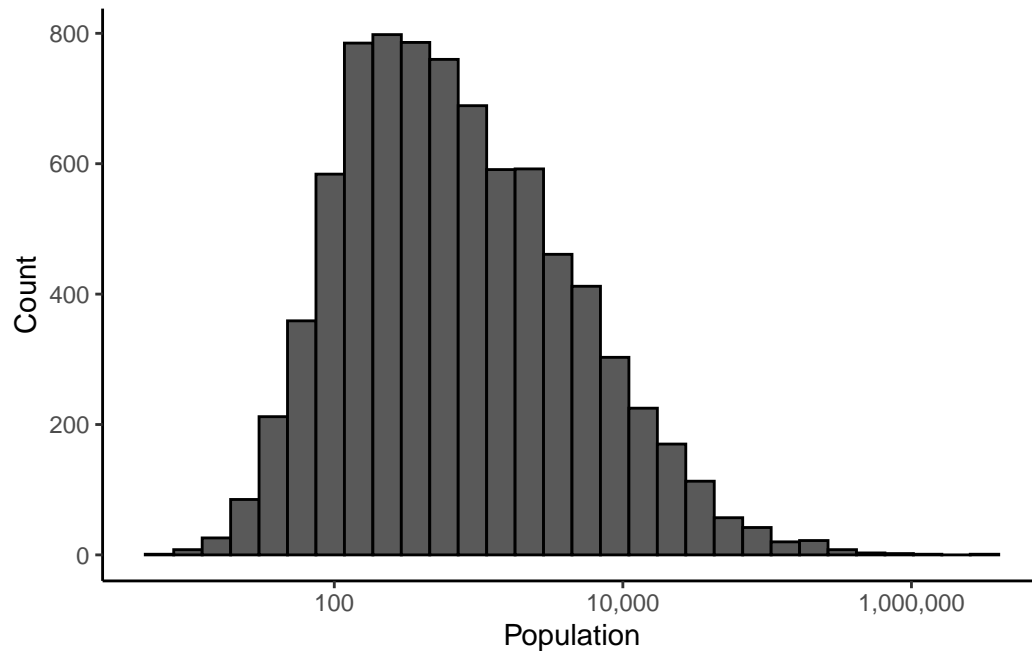
```
# A tibble: 1 x 3
  `mean(pop)` `median(pop)` `sd(pop)`
    <dbl>         <dbl>         <dbl>
1    5578.         582.         45916.
```

The city size distribution in Spain (and every other country, for that matter) is *long-tailed*. The largest city, Madrid, has over 3 million people, but the vast bulk of cities and towns have fewer than 1,000 people. Whenever you have a distribution that is so heavily skewed, the mean will be larger than the median.

2. What about the *logarithm* of population? What does that distribution look like?

```
ggplot(data = m07,
       mapping = aes(x=pop)) +
  geom_histogram(color = 'black') +
  theme_classic() +
```

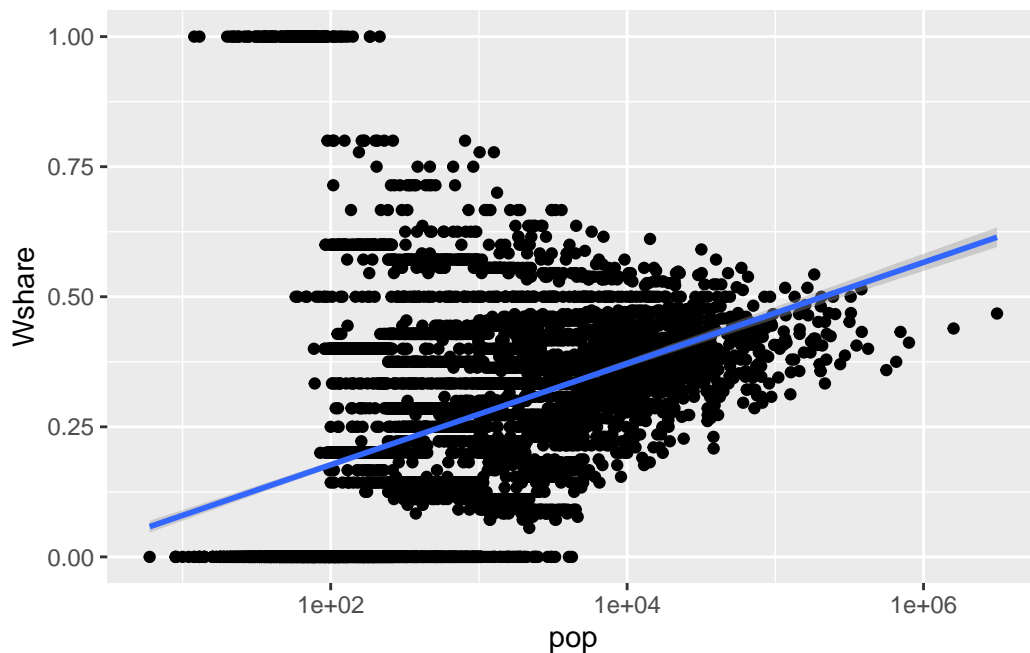
```
labs(x = 'Population',
     y = 'Count') +
scale_x_log10(labels = scales::comma_format())
```



City size distributions are roughly *log normal*, i.e. their logarithm is normally distributed.

3. Are larger cities likely to have a larger share of women on the council? Plot the relationship. Estimate a linear model and interpret the slope of the relationship.

```
ggplot(data = m07,
       mapping = aes(x=pop,
                     y=Wshare)) +
geom_point() +
scale_x_log10() +
geom_smooth(method = 'lm')
```



```
lm(Wshare ~ pop, data = m07)
```

Call:

```
lm(formula = Wshare ~ pop, data = m07)
```

Coefficients:

(Intercept)	pop
2.586e-01	3.591e-07

Larger cities tend to have more women on their municipal councils. On average, when the population increases by 100,000, the share of women on the council increases by about 3.6 percent.

4. What percent of municipal councils had a corruption scandal during the 2007-2011 period? Compare the average share of women on councils facing corruption scandals versus those not facing corruption scandals.

```
m07 |>
  count(corr0711) |>
  mutate(pct = n / sum(n) * 100)
```

```
# A tibble: 2 x 3
  corr0711     n  pct
    <dbl> <int> <dbl>
1         0  7987 98.4
2         1   129  1.59
```

Only 1.6% of the councils faced a corruption scandal during this period.

```
m07 |>
  group_by(corr0711) |>
  summarize(avg_women_share = mean(Wshare))
```

```
# A tibble: 2 x 2
  corr0711 avg_women_share
    <dbl>         <dbl>
1         0           0.259
2         1           0.355
```

On average, there is a *higher* share of women on councils with corruption scandals than the baseline.

5. Estimate the *difference-in-means* – the difference in women’s share between the cities with corruption scandals and those without. Compute and interpret a 95% confidence interval and p-value associated with that difference-in-means.

```
lm(Wshare ~ corr0711,
  data = m07)
```

Call:

```
lm(formula = Wshare ~ corr0711, data = m07)
```

Coefficients:

```
(Intercept)      corr0711
    0.25905         0.09551
```

```
t.test(Wshare ~ corr0711,
  data = m07)
```

Welch Two Sample t-test

```
data: Wshare by corr0711
t = -8.6803, df = 139.26, p-value = 9.403e-15
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -0.11726431 -0.07375511
sample estimates:
mean in group 0 mean in group 1
      0.2590513      0.3545610
```

Councils with corruption scandals have, on average, 9.5% more women on their councils. The p-value associated with this relationship is roughly 9×10^{-15} , meaning there is basically zero chance that we would have observed a statistic this large due purely to sampling error. The 95% confidence interval ranges from 7.4% to 11.7% – the true difference-in-means could be anywhere in that range and it would not be *surprising* for sampling error to have generated the statistic we observed.

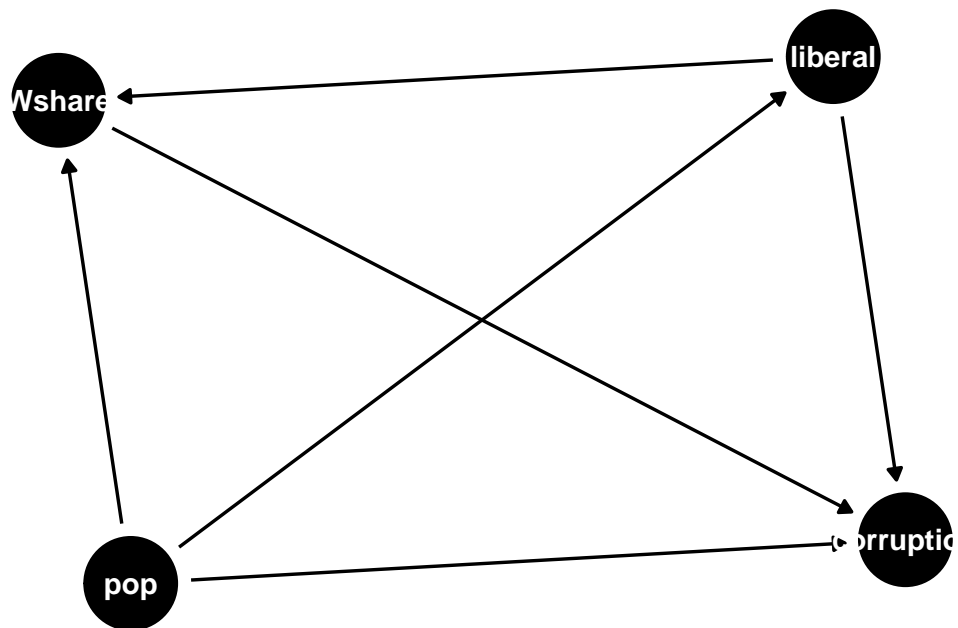
6. Can we interpret that relationship you just estimated as causal? Why or why not? Draw a DAG illustrating some potential alternative explanations.

There are a number of potential backdoor paths. For one, we already showed that population is strongly correlated with women's share. If larger places are also more likely to have corruption scandals, this could help explain the result. Maybe more liberal places are more likely to elect women to their councils and face corruption scandals.

```
library(dagitty)
library(ggdag)

dag <- dagify(corruption ~ Wshare + pop + liberal,
              Wshare ~ pop + liberal,
              liberal ~ pop)

ggdag(dag) +
  theme_dag()
```



7. Estimate a *conditional* difference-in-means for cities that don't have any banks. Interpret the results.

```
lm(Wshare ~ corr0711,
  data = m07 |> filter(bancos == 0))
```

Call:

```
lm(formula = Wshare ~ corr0711, data = filter(m07, bancos ==
  0))
```

Coefficients:

(Intercept)	corr0711
0.25258	0.03046

```
t.test(Wshare ~ corr0711,
  data = m07 |> filter(bancos == 0))
```

Welch Two Sample t-test

```

data: Wshare by corr0711
t = -0.73409, df = 13.567, p-value = 0.4754
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
95 percent confidence interval:
 -0.11972687  0.05880443
sample estimates:
mean in group 0 mean in group 1
    0.2525843      0.2830455

```

The conditional difference-in-means is smaller than the unconditional difference-in-means (3% compared to 9.5%) and statistically insignificant, suggesting that smaller cities without banks are less likely to elect women and also less likely to have problems with corruption.

8. Using the bandwidth reported in Table 2, just keep the cities slightly above and below the quota threshold in the 2007 election. Why is this subset of cities more appropriate for estimating the effect of women councilmembers on corruption than using the entire dataset?

```

threshold <- 5000
bandwidth <- 1042

d <- m07 |>
  filter(pop >= threshold - bandwidth,
         pop <= threshold + bandwidth)

```

Cities with a population of 5,000-6,000 are required to have more women on their councils by national law, while those with population 4,000-4,999 are not. Previously we worried that there might be a number of confounding variables that cause there to be more women on a council and also cause more corruption. But by comparing the cities just above and just below the population threshold, we ensure that the increase in the number of women on council is being caused by the *quota*, not by any of the other potential confounders. There should be a big increase in women's representation at that point, but not a big change in any of the confounders.

9. What is the average women's share of elected municipal councilmembers for the cities just above the cutoff? Just below? Plot the difference between the groups. Is the difference statistically significant?

```

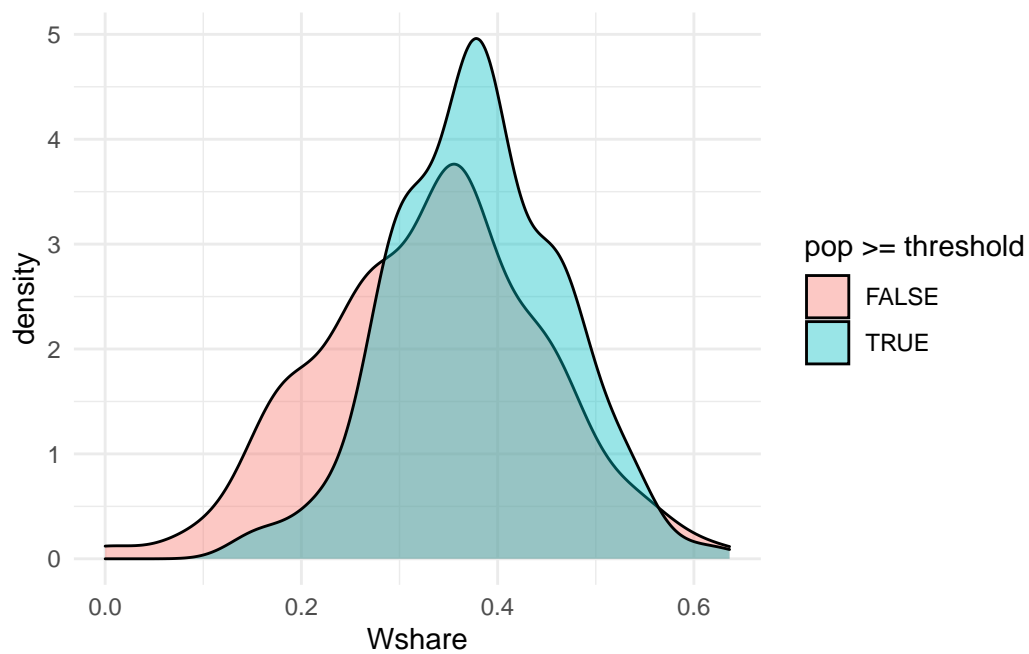
d |>
  group_by(pop >= threshold) |>
  summarize(avg_pct_women = mean(Wshare),
            num = n())

```



```
# A tibble: 2 x 3
  `pop >= threshold` avg_pct_women  num
  <lgl>                <dbl> <int>
1 FALSE                0.332    206
2 TRUE                 0.379    189
```

```
ggplot(data = d,
       mapping = aes(x=Wshare,
                     fill = pop >= threshold)) +
  geom_density(alpha = 0.4) +
  theme_minimal()
```



```
d |>
  mutate(treatment = as.numeric(pop >= threshold)) |>
  t.test(Wshare ~ treatment, data = _)
```

Welch Two Sample t-test

```
data: Wshare by treatment
t = -4.71, df = 378.79, p-value = 3.483e-06
```

```

alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
95 percent confidence interval:
-0.06657140 -0.02735889
sample estimates:
mean in group 0 mean in group 1
0.3318063      0.3787715

```

There is a small, but statistically significant increase in women's representation at the quota threshold, roughly 4.5 percentage points. It hardly seems like enough to cause a big change in corruption!

10. It can often be useful to conduct a *placebo test*. Do the same thing you just did with the 5,000 population cutoff, but for a 4,000 person cutoff. There's no law implementing a quota at that point, so we shouldn't expect to find a significant difference. What do you find?

```

threshold <- 4000

d2 <- m07 |>
  filter(pop >= threshold - bandwidth,
         pop <= threshold + bandwidth)

d2 |>
  group_by(pop >= threshold) |>
  summarize(avg_pct_women = mean(Wshare),
            num = n())

```

```

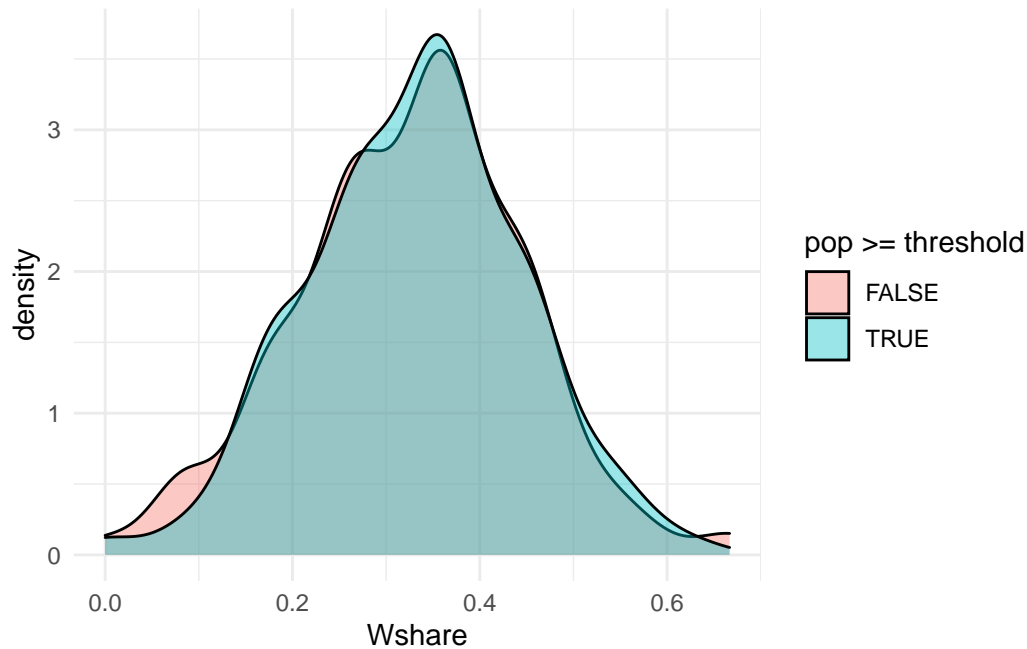
# A tibble: 2 x 3
  `pop >= threshold` avg_pct_women  num
  <lgl>              <dbl> <int>
1 FALSE              0.326   319
2 TRUE               0.331   201

```

```

ggplot(data = d2,
       mapping = aes(x=Wshare,
                     fill = pop >= threshold)) +
  geom_density(alpha = 0.4) +
  theme_minimal()

```



```
d2 |>
  mutate(treatment = as.numeric(pop >= threshold)) |>
  t.test(Wshare ~ treatment, data = _)
```

Welch Two Sample t-test

```
data: Wshare by treatment
t = -0.49617, df = 443.64, p-value = 0.62
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
95 percent confidence interval:
 -0.02566265  0.01531682
sample estimates:
mean in group 0 mean in group 1
  0.3259807      0.3311536
```

There is *no* significant difference in women's representation at the placebo threshold.

11. What about corruption? Is the rate of corruption scandals higher for cities just to the left or right of the threshold?

```

threshold <- 5000
bandwidth <- 1042

d <- m07 |>
  filter(pop >= threshold - bandwidth,
         pop <= threshold + bandwidth) |>
  mutate(treated = as.numeric(pop >= threshold))

d |>
  group_by(treated) |>
  summarize(pct_corruption = mean(corr0711))

```

```

# A tibble: 2 x 2
  treated pct_corruption
    <dbl>         <dbl>
1       0           0.0485
2       1           0.0423

```

```
t.test(corr0711 ~ treated, data = d)
```

Welch Two Sample t-test

```

data:  corr0711 by treated
t = 0.29601, df = 392.82, p-value = 0.7674
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
95 percent confidence interval:
 -0.03506717  0.04749847
sample estimates:
mean in group 0 mean in group 1
  0.04854369      0.04232804

```

The rate of corruption scandals in 2007-2011 does not seem to be significantly different between cities just barely above or below the threshold.

12. What about the *change* in corruption scandals between the 2003-2007 period and the 2007-2011 period? Is that higher for cities just barely above or below the threshold?

```

d |>
  group_by(treated) |>

```

```
summarize(change_in_corruption = mean(corr0711 - corr0307))
```

```
# A tibble: 2 x 2
  treated change_in_corruption
  <dbl>         <dbl>
1      0         -0.0146
2      1         -0.0688
```

On both sides of the threshold there was a slight decrease in the share of cities with a corruption scandal, but that decrease was slightly larger (about 5 percentage points) in the cities with a gender quota.

13. Re-estimate that difference-in-means using the local linear regression approach from the `rdrobust` package. What is the associated p-value and 95% confidence interval? Interpret the results.

```
library(rdrobust)

Y <- d$corr0711 - d$corr0307
X <- d$pop
threshold <- 5000

rd <- rdrobust(y=Y, x=X, c=threshold)
```

```
[1] "Mass points detected in the running variable."
```

```
summary(rd)
```

Sharp RD estimates using local polynomial regression.

Number of Obs.	395	
BW type	mserd	
Kernel	Triangular	
VCE method	NN	
Number of Obs.	206	189
Eff. Number of Obs.	71	66
Order est. (p)	1	1
Order bias (q)	2	2

BW est. (h)	361.540	361.540
BW bias (b)	518.496	518.496
rho (h/b)	0.697	0.697
Unique Obs.	187	169

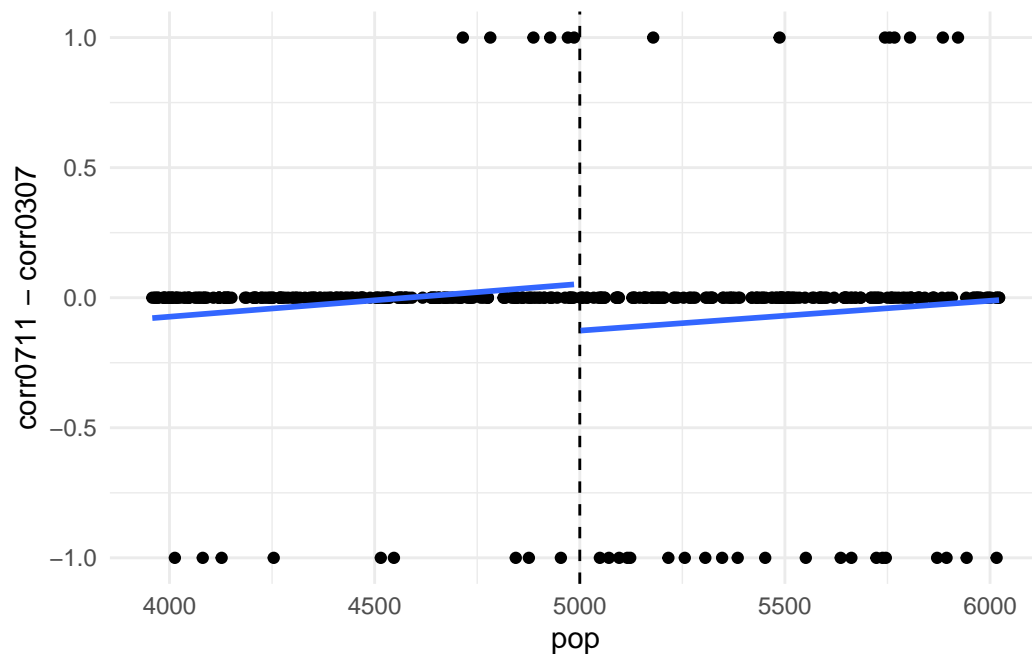
Method	Coef.	Std. Err.	z	P> z	[95% C.I.]
Conventional	-0.233	0.150	-1.551	0.121	[-0.528 , 0.061]
Robust	-	-	-1.342	0.180	[-0.602 , 0.113]

This estimate is much larger (about 23.3% difference between the groups), but the 95% confidence interval includes zero. This is a result that would not be surprising from sampling error alone if women had no effect on municipal corruption.

- Optional Bonus Fun:** Create a chart that visualizes what's going on in the local linear regression analysis. Keep data points only within the bandwidth, with population on the x-axis, outcome on the y-axis, a vertical line at the threshold, and two linear models on either side. What's going on here?

```
ggplot() +
  # points to the left of the threshold
  geom_point(data = d |>
    filter(pop < threshold),
    mapping = aes(x = pop,
      y = corr0711 - corr0307)) +
  # points on the right
  geom_point(data = d |>
    filter(pop >= threshold),
    mapping = aes(x = pop,
      y = corr0711 - corr0307)) +
  # vertical dashed line at threshold
  geom_vline(xintercept = threshold, linetype = 'dashed') +
  # linear model to the left of the threshold
  geom_smooth(data = d |>
    filter(pop < threshold),
    mapping = aes(x = pop,
      y = corr0711 - corr0307),
    method = 'lm',
    se = FALSE) +
  # linear model to the right of the threshold
```

```
geom_smooth(data = d |>
  filter(pop >= threshold),
  mapping = aes(x = pop,
    y = corr0711 - corr0307),
  method = 'lm',
  se = FALSE) +
theme_minimal()
```



There are five cities just to the right of the threshold that had corruption scandals during the 2003-2007 period, but did not during the 2007-2011. As best I can tell, these five cities are really driving the strong negative treatment effect estimate in the paper.

So it's worth asking, what happened in those cities? Did they get a lot of new women councilmembers, which caused them to reduce their corruption?

```
d |>
  filter(corr0711 - corr0307 < 0,
    pop > 5000,
    pop < 5200) |>
  select(Wshare03, Wshare, FemAlc03, FemAlc)
```

```
# A tibble: 5 x 4
```

	Wshare03	Wshare	FemAlc03	FemAlc
	<dbl>	<dbl>	<dbl>	<dbl>
1	0.25	0.143	1	0
2	0.588	0.333	1	1
3	0.308	0.4	0	0
4	0.364	0.312	0	0
5	0.231	0.294	0	1

Honestly, this does not appear to corroborate the theory. Three out of these five cities had *fewer* women on the council after the gender quota. One previously had a female mayor who lost in the 2007 election. It seems unlikely that the reason why these cities had less reported corruption post-quota is because they got a bunch of new female leadership.² Since the result seems to depend on the data from these five cities, I'm skeptical of drawing big conclusions about the effect of women on corruption.

²FYI I would have conducted this analysis using something called a *fuzzy* regression discontinuity, which takes into account the fact that cities to the right of the threshold are more likely to be treated, but not guaranteed to be treated.