Answers to PS2 - Program Evaluation PPHA 34600

Diego Diaz

May 14, 2020

Other group members: Piyush Tank, Matthew Mauer.

1 - HARRIS are interested in answering the following question: What was the effect of FIONA on profits for the average farmer? To make sure everybody is on the same page, explain to them what the ideal experiment would be for answering this question. Describe the dataset that you'd like to have to carry out this ideal experiment, and use math, words, and the potential outcomes framework to explain what you would estimate and how you would do so. Make sure to be clear about the unit of analysis (ie, what is "i" here?)

Answer: We would need an experiment in which the assignment of the treatment is made randomly, so that the distribution of observables and unobservables is the same for both treated and untreated units. We are interesting in estimating the effect of FIONA on farmers' profits, therefore the treatment in our experiment is participation in the FIONA program, and the effect is the difference in profits that we would observe when a farmer participates in comparison for when he does not. According to the potential outcomes framework, naming the impact on profits τ_i, the effect for any farmer can be expressed as:

$$\tau_i = Y_i(1) - Y_i(0)$$

Where i is an index for each farmer, $Y_i(1)$ is the profit for farmer i after he participated in FIONA and $Y_i(0)$ is what we could have observed if he had not participated.

Although we can't observe τ_i because we can't observe a farmer in both states, HARRIS is interested in the effect on the average farmer, in other words, the Average Treatment Effect (ATE). This is:

$$\tau^{ATE} = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0]$$

We can estimate the (ATE) with our ideal experiment. By randomly assignment the treatment to obtain a dataset with the treatment assignment, that is, participation in the program, and an observed outcome, in this case, post-treatment profits. Given that both groups would be equal in every observable/unobservable before the program in this setting, differences between average profits afterwards would give us an unbiased estimate of the (ATE).

$$\hat{\tau}^{ATE} = \overline{Y(1)} - \overline{Y(0)}$$

Where $\overline{Y(1)}$ and $\overline{Y(0)}$ are simply the means between treatment and control groups.

2 - HARRIS like what you're suggesting, but think it's answering the wrong question. They aren't going to be able to get every single farmer to participate. They'd instead like to know: What was the effect of FIONA on profits among farmers who took up insurance? Describe in math and words, using the potential outcomes framework, what they'd like to estimate. Explain how this differs from what you described in (1), and describe what component of this estimand you will be fundamentally unable to observe.

Answer: HARRIS would like to estimate the effect of taking up the insurance on profits. This differs to what we described in (1) since before we were looking at the effect of participating in the program. Once in the program, farmers might have chosen not to take insurance (never-takers). Potentially we may also have farmers that take the program insurance even when not participating in the program (always-takers). With non compliers, we can define R_i as a dummy variable that takes the value of 1 if the insurance is offered and 0 if it's not, and D_i as whether the farmer takes the insurance or not. In this case the estimand from before does not give as the (ATE) since we would have a selection problem with the farmers that take insurance. We could obtain:

$$\tau^{ITT} = E[Y_i(1)|R_i = 1] - E[Y_i(0)|R_i = 0]$$

$$\hat{\tau}^{ITT} = \overline{Y(1)} - \overline{Y(0)}$$

Where τ^{ITT} is the intent to treat effect. We can still obtain an estimand for the effect of taking insurance on the compliers, which would be the τ^{ATT} (Average Treatment effect on the Treated) by calculating:

$$\hat{\tau}^{ATT} = \frac{\hat{\tau}^{ITT}}{P[D_i = 1|R_i = 1] - p[D_i = 1|R_i = 0]}$$

In other words, we can't estimate the effect on the average farmer but we can estimate the effect on the farmers that choose to take insurance when offered participation. This is because we can't observe the effect of insurance on farmers than choose not to take it after being offered to take it. However, in the specific case of constant treatment effects, we have that: $\tau^{ATT} = \tau^{ATE}$.

3 - HARRIS are on board with your explanation. Because FIONA already exists in the real world, they can't run an RCT to study it. However, they do know that not all farmers were offered insurance through FIONA. It turns out that FIONA only impacted certain districts. Non-FIONA districts were not offered any insurance products. Explain what you would recover if you simply compared FIONA farms to non-FIONA farms on average. Describe three concrete examples of why this might be problematic

Answer: If we compare FIONA farms to non-FIONA farms, we would be comparing observed outcomes and not potential outcomes. In this setting, we cannot estimate the (ATE) because the estimator compares potential outcomes. The problem is that participating farmers may be fundamentally different in one group in comparison to the other, and these differences might be unobservable, which would create omitted variable bias. If we have a problem in observables, for instance, if the farmers from the treated districts are more productive, healthy, educated, or with more access to capital; although we could potentially control for these variables, this would make it impossible for us to distinguish the treatment assignment to the fact that those farmers simply live in the treated district. Other effect may also play a role in confounding the results. For example, a local natural disaster in a district could ruin the crops for the farmers in that district, creating an upward bias in the estimand. Lastly, districts might have different distributions of crops, and that would also create bias in the results since the fiona farmers would have a different distributions of crops.

4 - HARRIS hears your concerns, but still wants an estimate of the impacts of FIONA. Given that you're unable to implement your ideal experiment, and you are worried about simple comparisons of FIONA-aided farmers and those without insurance, you'll need to do something a little more sophisticated. Luckily for you and for HARRIS, India makes data on farmers available to the public, in the form of ps2_data.csv. Read the data into R and, as always, make sure everything makes sense. Document and fix any errors. Use the variables contained in the dataset to describe, using math and words, two (related) potential approaches to estimating the effect of FIONA on profits. Make sure to be clear about your unit of analysis, and be explicit about how these designs apply to FIONA (ie, describe things in terms of "profits," not just "outcome"). Hint: HARRIS wants you to describe two selection-on-observables designs.

Answer: Some issues are observed in the data. The first thing we notice by making a summary in R is that we don't obtain statistics for *farmer_birth_year*. Taking a closer look we find that the data is not in numeric format, but character. This happens because some entries (about 500) are written as characters. Luckily, this only happened with two years, 1972 and 1973, so we can easily replace those values.

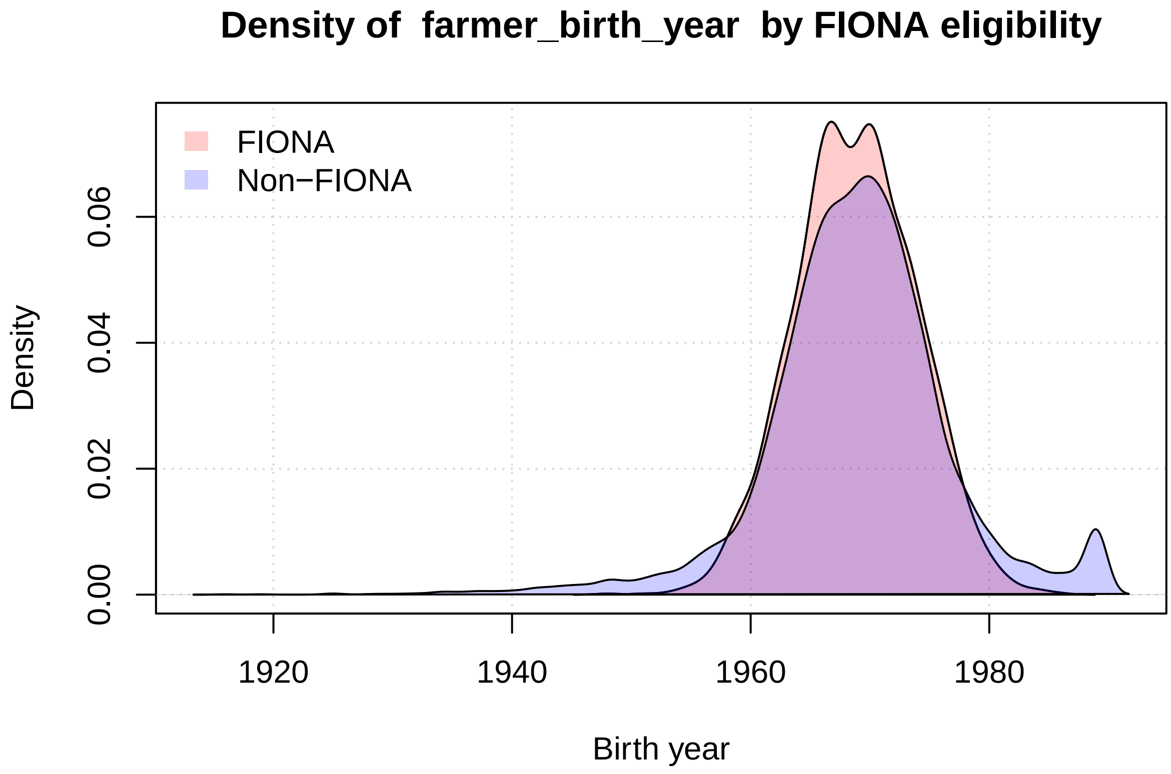
```
library(haven)
library(tidyverse)
library(dplyr)
library(knitr)
library(broom)

ps2_data <- read.csv("C:/Users/diego/Google Drive/Program Eval/Assignments/data/ps2_data.csv")

# summary(ps2_data) %>% kable()

# unique(ps2_data['farmer_birth_year']) %>% dplyr::filter(nchar(farmer_birth_year) > 5)

ps2_data[ps2_data['farmer_birth_year'] == 'nineteen seventy-three',]['farmer_birth_year'] <- 1973

ps2_data[ps2_data['farmer_birth_year'] == 'nineteen seventy-two',]['farmer_birth_year'] <- 1972

ps2_data['farmer_birth_year'] <- as.numeric(ps2_data$'farmer_birth_year')
# summary(ps2_data) %>% kable()

fiona = ps2_data[ps2_data$fiona_farmer == 1,]
non_fiona = ps2_data[ps2_data$fiona_farmer == 0,]
```

Looking at density plots and histograms for the data, we observe the following:

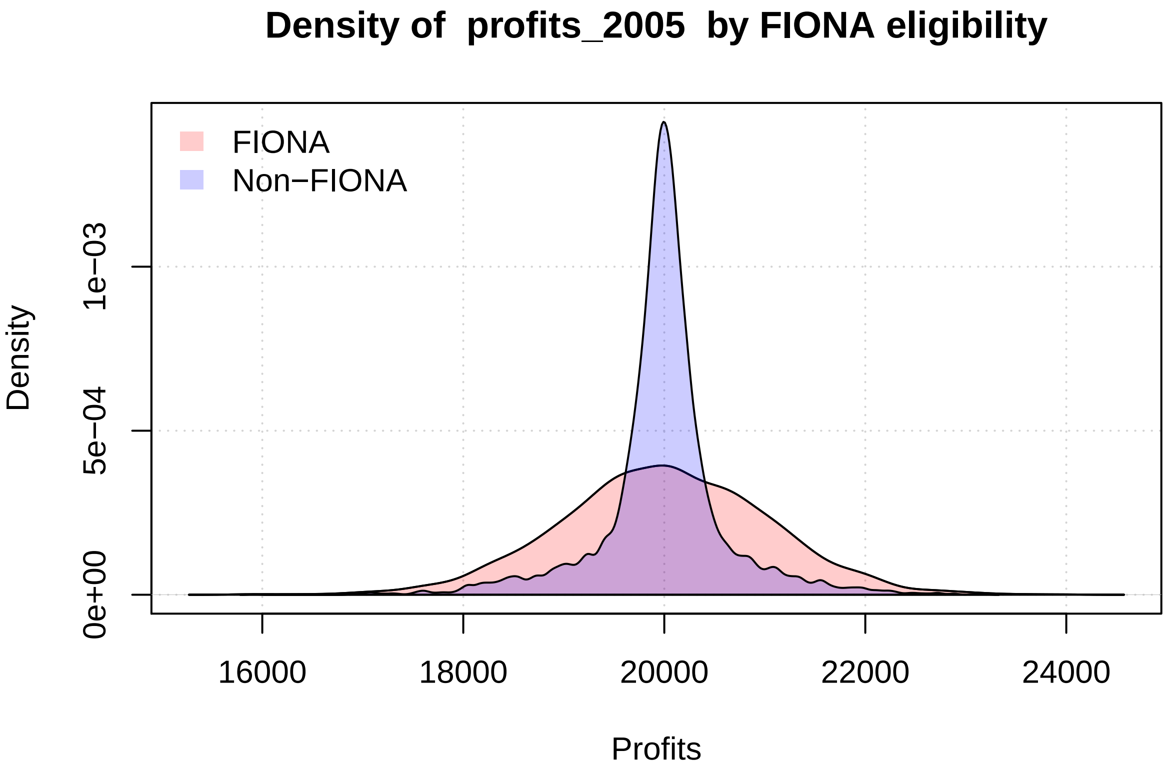
```
plot_densities <- function(x_label, col_number) {
    ## calculate the density - don't plot yet
    dens_fiona = density(fiona[,col_number])
    dens_non_fiona = density(non_fiona[,col_number])
    ## calculate the range of the graph
    xlim <- range(dens_non_fiona$x,dens_fiona$x)
    ylim <- range(0,dens_non_fiona$y, dens_fiona$y)
    #pick the colours
    fionaCol <- rgb(1,0,0,0.2)
    non_fionaCol <- rgb(0,0,1,0.2)
    ## plot the carrots and set up most of the plot parameters
    plot(dens_fiona, xlim = xlim, ylim = ylim, xlab = x_label,
         main = paste('Density of ', colnames(fiona)[col_number],
                      ' by FIONA eligibility'),
         panel.first = grid())
    #put our density plots in
    polygon(dens_fiona, density = -1, col = fionaCol)
    polygon(dens_non_fiona, density = -1, col = non_fionaCol)
    ## add a legend in the corner
    legend('topleft',c('FIONA','Non-FIONA'),
           fill = c(fionaCol, non_fionaCol), bty = 'n',
           border = NA)

}
```

Nest we show barplots for district, crop types and fertilizer use. In each case, we first plot for fiona farmers and then for non fiona farmers.

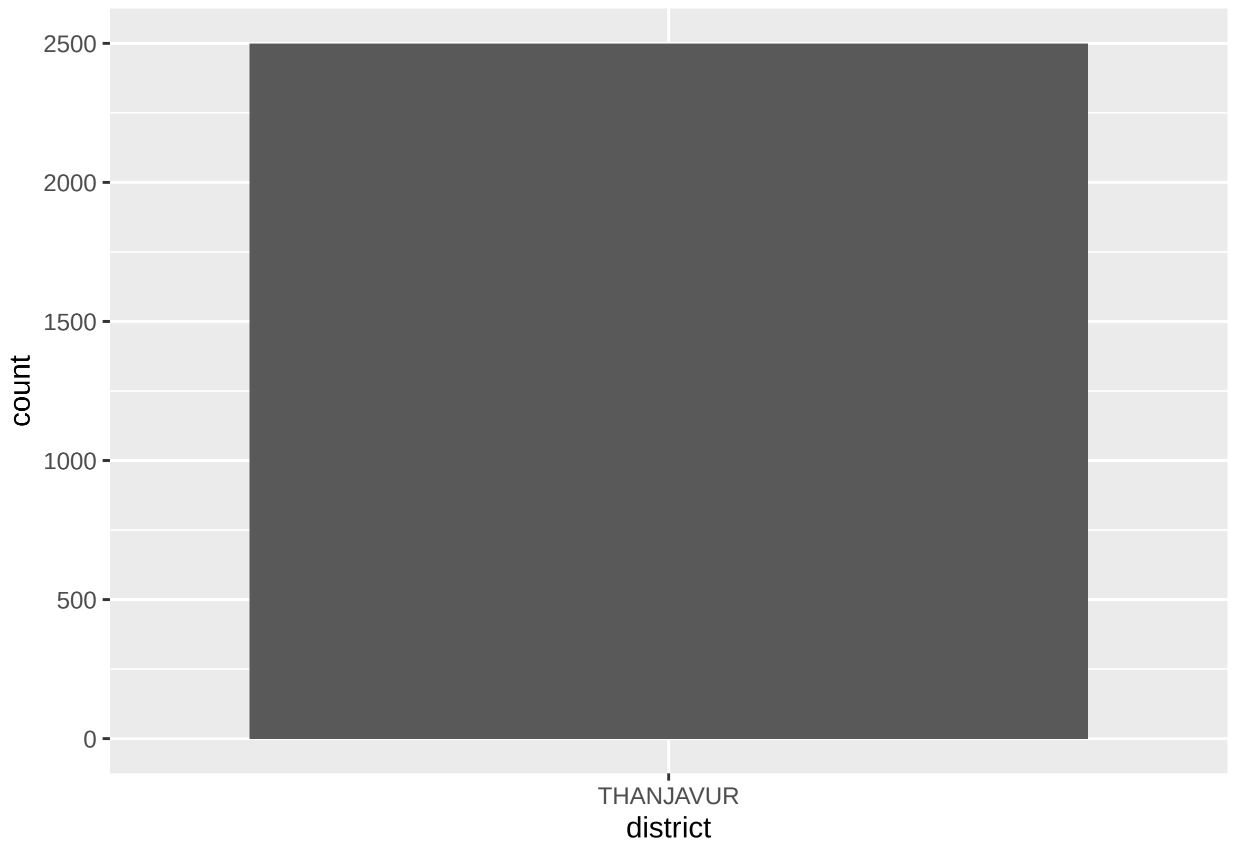
```r
plot_densities("Birth year", 4)
```



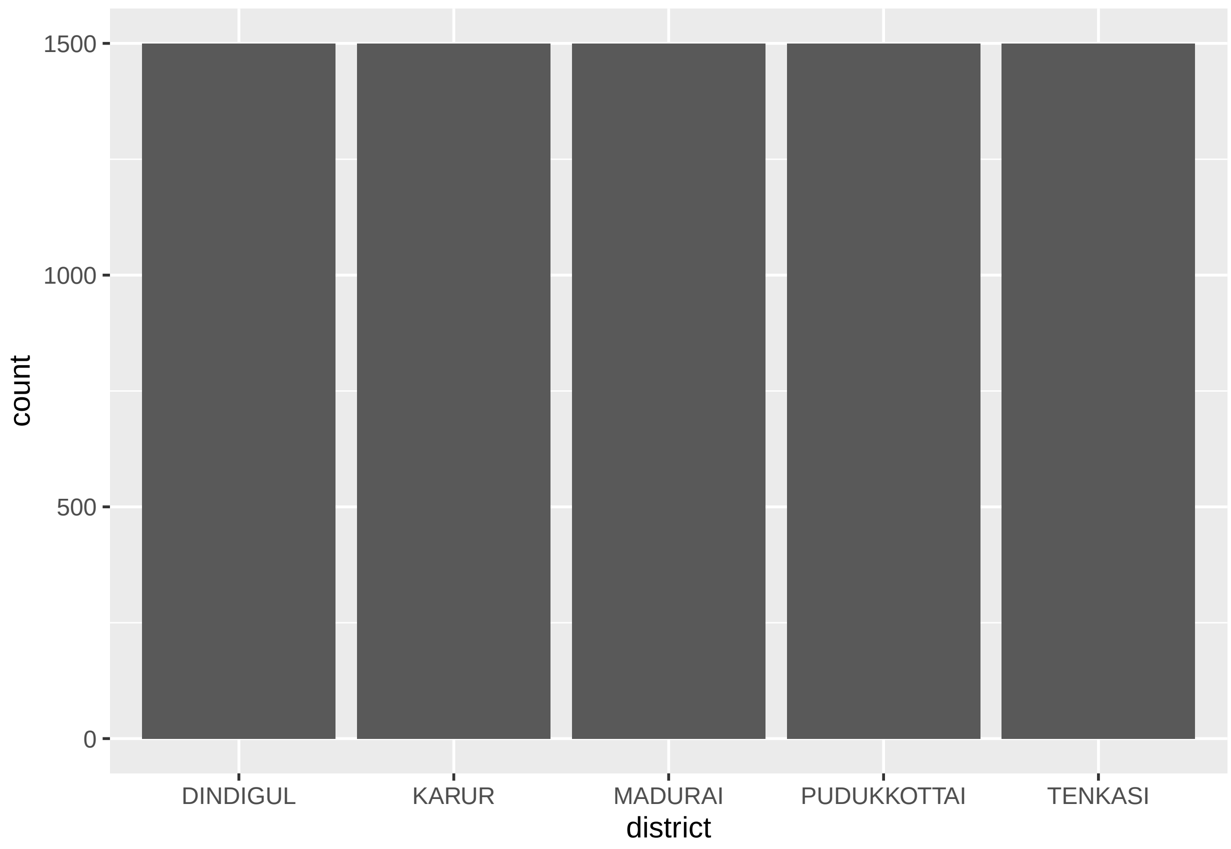
```r
plot_densities("Profits", 6)
```



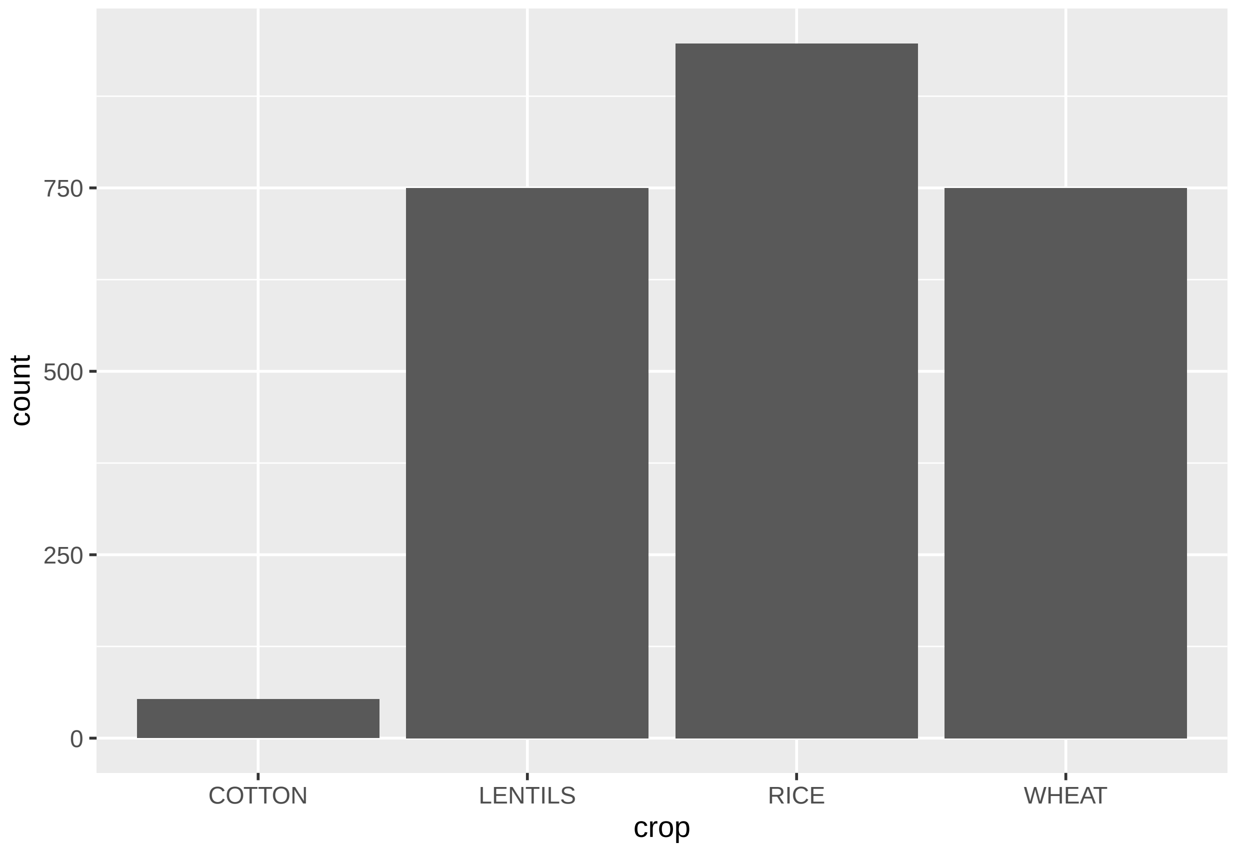
```
ggplot(data.frame(fiona), aes(x=district)) +
  geom_bar()
```



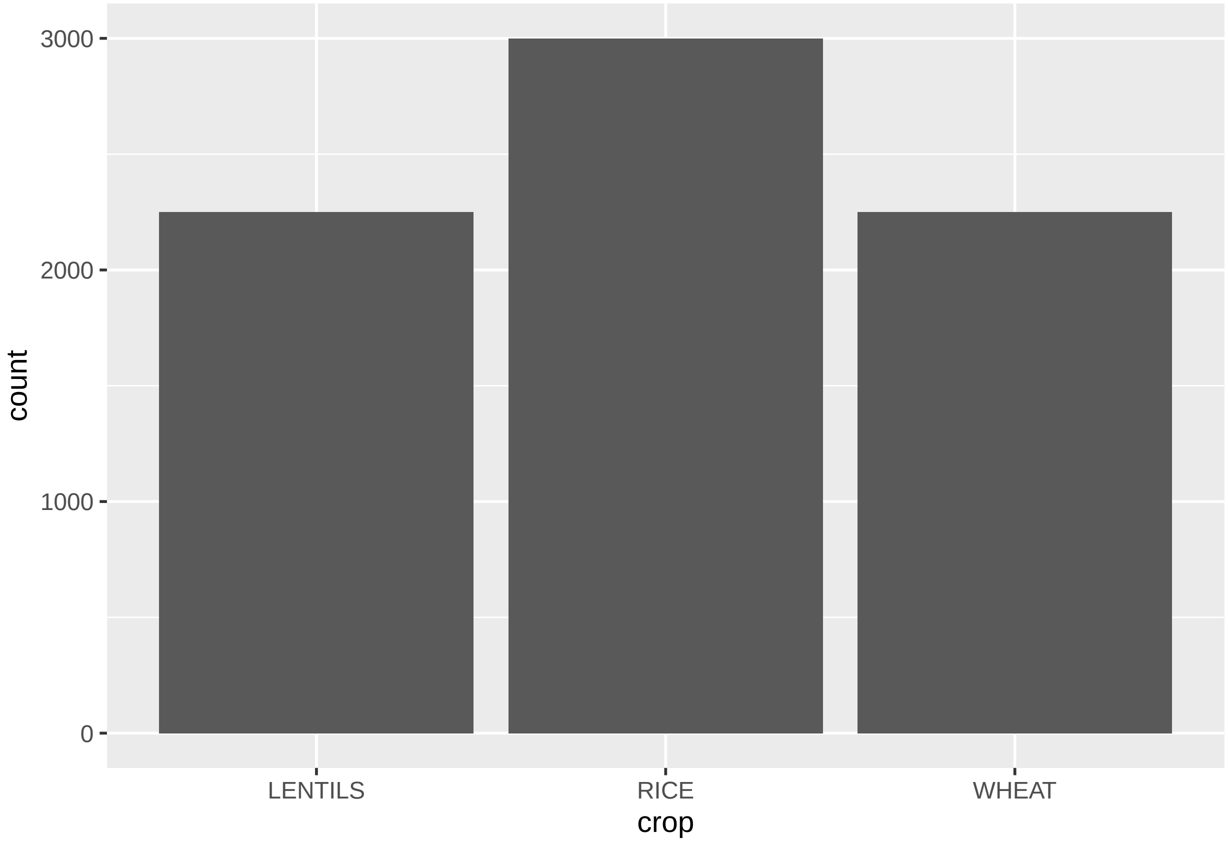
```
ggplot(data.frame(non_fiona), aes(x=district)) +
  geom_bar()
```



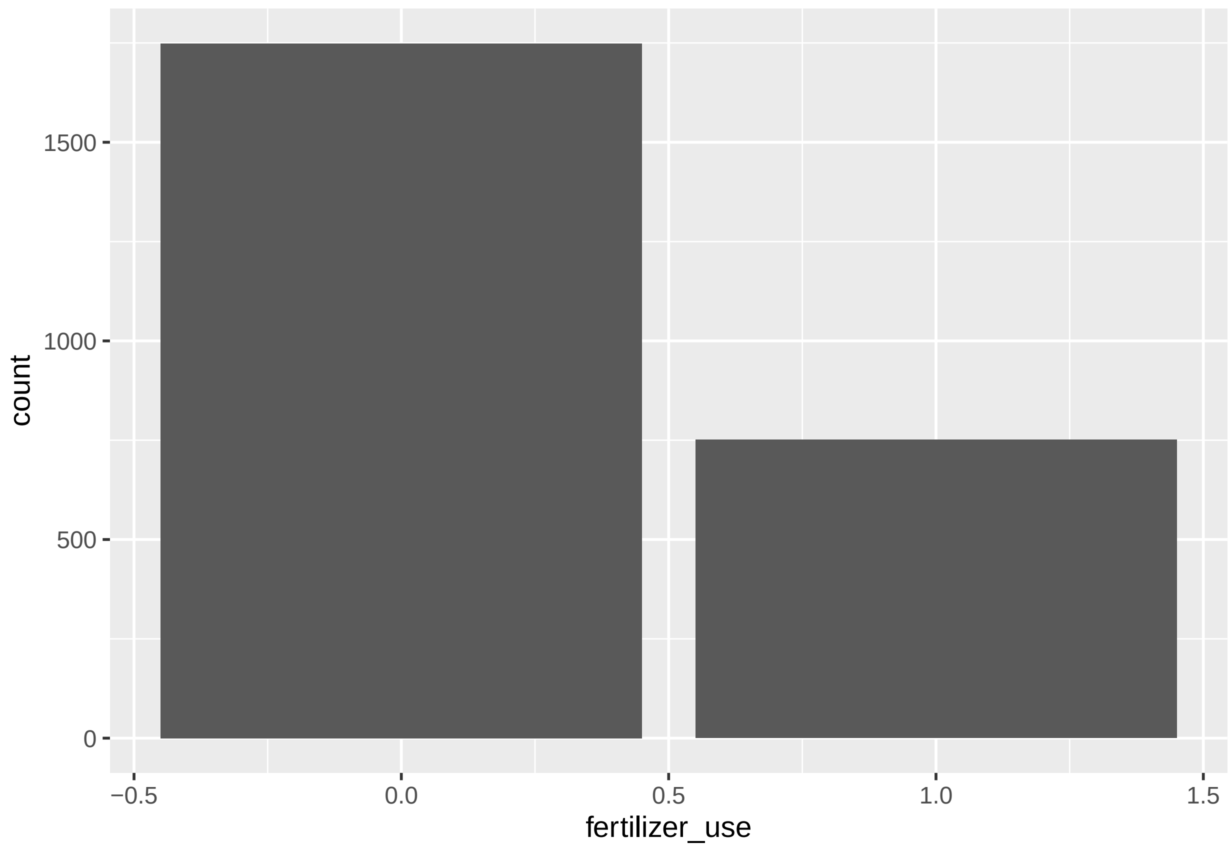
```
ggplot(data.frame(fiona), aes(x=crop)) +
  geom_bar()
```



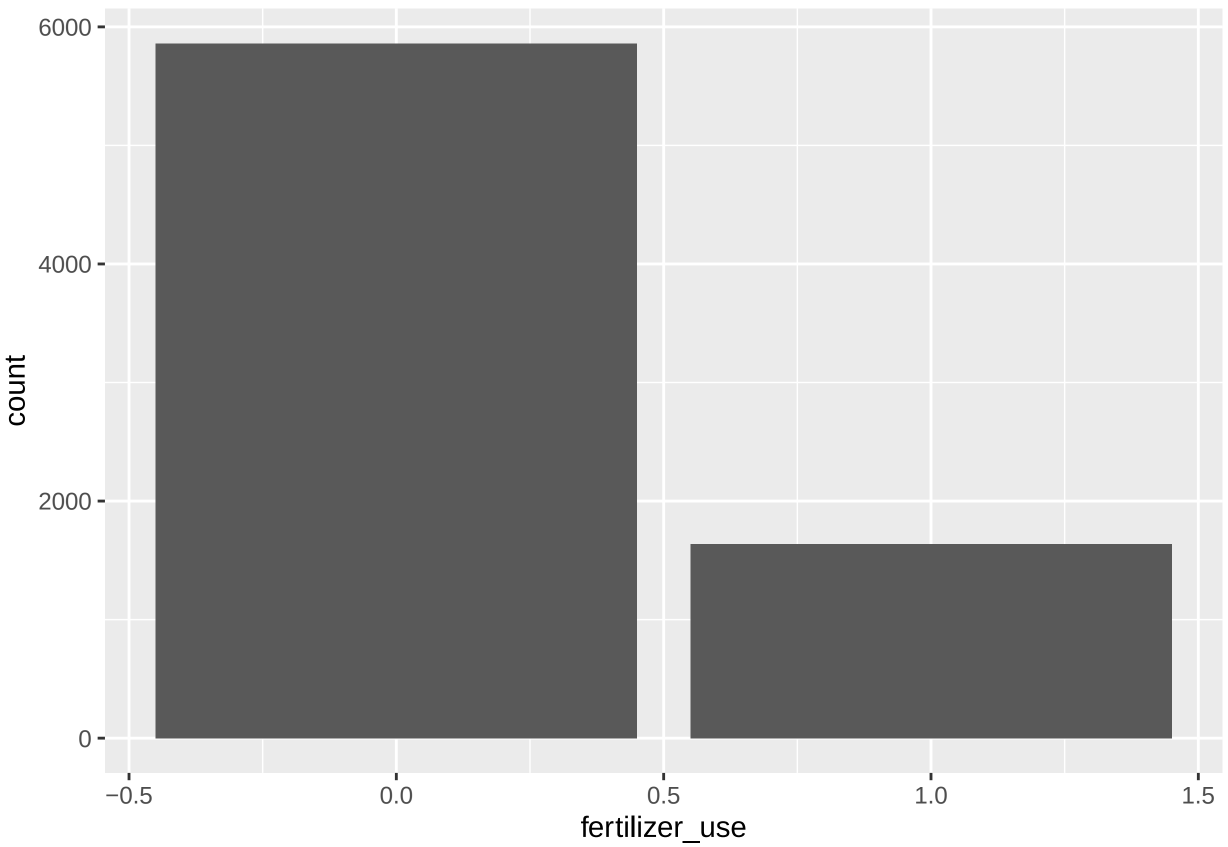
```
ggplot(data.frame(non_fiona), aes(x=crop)) +
  geom_bar()
```



```
ggplot(data.frame(fiona), aes(x=fertilizer_use)) +
  geom_bar()
```



```r
ggplot(data.frame(non_fiona), aes(x=fertilizer_use)) +
  geom_bar()
```



We can make a few observations from the data.

1 - There is the same number of observations for every district.
2 - Every farmer in district THANJAVUR has FIONA = 1, while all the farmers from any other district have FIONA = 0.
3 - Cotton farmers are only present in the FIONA group.
4 - The distributions of profits in 2005 for FIONA and non-FIONA farmers are different. The distributions of profits for FIONA farmers has a much larger variance.
5 - Fertilizer is used more by FIONA farmers.
6 - All the other variables are evenly distributed between the groups, except the outcome variable, profits 2015, that we don't check.

We have a problem since an RCT was not performed to implement FIONA and both groups are fundamentally different in certain variables. To estimate the effect of the treatment we need a selection-on-observables approach (SOO). Two options are available, regression adjustment and exact matching. We will assume that, conditional on observables, FIONA participation is independent of potential profits. If we control for the pre-treatment variables that differentiate both groups, we can obtain a non biased estimator of the ATE. In our case, without looking at balance tables yet, we should include variables on all the observables that seem to differ in distribution between groups as well as covariates of profits. This includes crop, profits in 2005, and age. We can't include district as it is highly correlated with the FIONA treatment assignment (all FIONA farmers are from the same district). The other approach we can use is exact matching, which will consist in grouping the farmers in different cells depending on the same observables (crop, farmer_birth_year), and then estimating the difference in profits 2016 of the FIONA with the non-FIONA ones. Afterwards we will take the weighted average, which will give us an estimand for the ATE. For the same reason as before, we can't include district. As mentioned by HARRIS, insurance permits farmers to invest in riskier up-front inputs like fertilizer, which is why fertilizer must be a post treatment outcome rather than a variable we might use as a pre-treatment control. Therefore, we can't include $fertilizer_use$ in either our regression or matching estimator as we can't control for post-treatment outcomes.

5 - Produce a balance table which displays the differences between FIONA and non-FIONA farmers on observable characteristics. Interpret this table. Does this table make you feel better or worse about your concerns in (3)?

Answer: The balance table is shown next. The district and crop variable were splitted in dummies to perform a regression on the treatment assignment. For the reasons mentioned in 4 ($fertilizer_use$ being a post-treatment outcome), we don't check for balance in $fertilizer_use$.

```r
#install.packages('fastDummies')
library(fastDummies)
ps2_data <- fastDummies::dummy_cols(ps2_data, select_columns = "crop")
ps2_dummies <- fastDummies::dummy_cols(ps2_data, select_columns = "district")
pre_ra_rest<-c("farmer_birth_year", "profits_2005", "crop_RICE", "crop_WHEAT", "crop_LENTILS", "crop_CO
balance_test <- cbind(e = ps2_dummies$fiona_farmer,
                      ps2_dummies[, pre_ra_rest]) %>%
  {map(.[, -1], function(i) lm(i~.$e))} %>%
  {map(., summary)} %>%
  sapply(`[`, "coefficients") %>%
  sapply(function(x) x[2,])%>%
  t()%>%
  data.frame()%>%
  rownames_to_column()%>%
  mutate_if(is.character, ~gsub(".coefficients", "", x=.))%>%
  `colnames<-`(c("var","Estimate","standard_error","t-statistics","Pr(>|t|)"));

balance_test %>% kable()
```

var	Estimate	standard_error	t-statistics	Pr(>|t|)
farmer_birth_year	-0.1118667	0.1649225	-6.782985e-01	0.4975982
profits_2005	11.8348668	17.1785837	6.889315e-01	0.4908824
crop_RICE	-0.0212000	0.0112872	-1.878239e+00	0.0603776
crop_WHEAT	0.0000000	0.0105841	0.000000e+00	1.0000000
crop_LENTILS	0.0000000	0.0105841	0.000000e+00	1.0000000
crop_COTTON	0.0212000	0.0016635	1.274407e+01	0.0000000
district_DINDIGUL	-0.2000000	0.0080008	-2.499750e+01	0.0000000
district_KARUR	-0.2000000	0.0080008	-2.499750e+01	0.0000000
district_MADURAI	-0.2000000	0.0080008	-2.499750e+01	0.0000000
district_PUDUKKOTTAI	-0.2000000	0.0080008	-2.499750e+01	0.0000000
district_TENKASI	-0.2000000	0.0080008	-2.499750e+01	0.0000000
district_THANJAVUR	1.0000000	0.0000000	8.577578e+15	0.0000000

As expected from the observations made in 4, the district dummies are unbalanced. This had to be the case since belonging to THANJAVUR is indistinguishible from being a FIONA farmer. The variable $district\_THANJAVUR$ is perfectly correlated with $fiona\_farmer$, and the ones from other districts as expected are also correlated, and should not be included in regressions to explain observable outcomes.

We can also notice that all other variables are balanced at a $p-value$ of 5%, except being a cotton farmer. This was also expected since this group is only represented in the treated group.

According to the table, we feel worse about what was mentioned in 3. Treatment assignment happened in only one district and this means that farmers from each group can be fundamentally different.

6 - HARRIS are interested in your approach in (4), but would like to know a bit more about how much they should believe your proposal. Describe the assumptions required for these designs to be valid in math and in words. To the extent possible, assess the validity of these assumptions using the provided data. Discuss whether you think you will be able to obtain a credible estimate of the answer to the questions described in (1) and (2) based on the data, and use concrete examples to explain why or why not.

Answer: As mentioned before, we require that, conditional on observables, FIONA participation is independent of potential profits. The central assumption is that profits in 2016, are independent of the assignment (D_i) conditional on covariates. In our case, conditional on crop type, profits on 2015, and birth year. We also need a second assumption, which is that both FIONA and non-FIONA farmers are represented in for different values of the covariates. This means that the probability of $D_i = 1$ for all levels of X_i is berween 0 and 1.

As we found in part (4), all cotton farmers were assigned treatment in the FIONA program, which is a direct violation of the second assumption. Given $crop_COTTON = 1$, we know that $D_i = 1$ with probability equal to 1. This is also the case with the district variable, for THANJAVUR, the probability of FIONA assignment is 1 and for every other district is 0. Given this problems it is not likely that we find a credible estimate of the ATE from (1) or the ATT from (2). As mentioned before, we could never distinguish the treatment from belonging to the THANJAVUR district.

7 - Use a regression-based approach to estimate the effect of FIONA on farmer profits. Describe which variables you chose to include in your regression, and explain why you chose these. Did you leave any variables out? If yes, explain why. Interpret your results. What are the strengths and weaknesses of this approach? How do your results differ from what you find if you instead use the naive estimator?

Answer: Let's calculate the naive estimator first. As we know, it is the difference between the means of the observed outcomes:

$$\tau^N = \overline{Y(1)} - \overline{Y(0)}$$

$$\tau^N = 2369.56$$

We need to control for covariates so that the treatment assignment is independent of the profits in 2016 (outcome) conditional on the covariates. This method is called regression adjustment, and in our case, we will include the following variables:

- $profits$_2005: It is expected that farmers that have higher profits in 2005 have a higher chance of having higher profits in 2016. In other words, there should be some persistance of profits in time.
- $crop\_LENTILS$, $crop\_WHEAT$, $crop\_RICE$: These are dummies for each crop type. These are covariates of profits since different crops have different prices and likely require different amounts of capital and labor to grow.
- $farmer_birth_year$: Although a mechanism is not clear, more experienced farmers might be more productive since they might be better able to predict the seasonal changes and weather patterns.
- $fionar\_farmer$: The treatment assignment variable. Coefficient which we are interested in estimating.

Left-out variables:

- $district$: As observed in part (4), all the fiona farmers belong to the same district. That is, $district\_THANJAVUR$ is perfectly correlated with $fiona\_farmer$, and the other district dummies are also correlated for the same reason.
- $crop\_COTTON$: We can't include $crop\_COTTON$ since only fiona farmers are cotton farmers, which means that the second assumption in our SOO design would be violated.

- $fertilizer_use$: As insurance permits farmers to invest in riskier up-front inputs like fertilizer, fertilizer must be a post treatment outcome. Therefore, we can't include $fertilizer_use$ as a covariate.

The main weakness of this approach is that the seconod assumption on our SOO design is broken. This means that the probability of $D_i = 1$ for some levels of X_i of some variables is either 0 or 1. The strengths of this approach is that the observable part of the selection problem is controlled for, and therefore the estimand is unbiased as long as the treatment assignment is uncorrelated with the unobservable variables.

The results of the regression are shown next:

```r
linearMod <- lm(profits_2016 ~ fiona_farmer +profits_2005+ farmer_birth_year+
                    crop_RICE+ crop_WHEAT+ crop_LENTILS, data=ps2_data)  # build linear regression mode

summary(linearMod)
```

```
FALSE
FALSE Call:
FALSE lm(formula = profits_2016 ~ fiona_farmer + profits_2005 + farmer_birth_year +
FALSE     crop_RICE + crop_WHEAT + crop_LENTILS, data = ps2_data)
FALSE
FALSE Residuals:
FALSE     Min      1Q  Median      3Q     Max
FALSE -3787.2  -702.1     2.3   723.9  3839.1
FALSE
FALSE Coefficients:
FALSE                    Estimate Std. Error t value Pr(>|t|)
FALSE (Intercept)       1.239e+03  2.900e+03   0.427  0.66916
FALSE fiona_farmer      2.358e+03  2.435e+01  96.867  < 2e-16 ***
FALSE profits_2005      1.004e+00  1.406e-02  71.404  < 2e-16 ***
FALSE farmer_birth_year 1.933e-01  1.464e+00   0.132  0.89499
FALSE crop_RICE         3.120e+01  1.458e+02   0.214  0.83056
FALSE crop_WHEAT        4.283e+02  1.461e+02   2.932  0.00338 **
FALSE crop_LENTILS      3.594e+02  1.461e+02   2.460  0.01391 *
FALSE ---
FALSE Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
FALSE
FALSE Residual standard error: 1046 on 9993 degrees of freedom
FALSE Multiple R-squared:  0.6009,  Adjusted R-squared:  0.6006
FALSE F-statistic:  2507 on 6 and 9993 DF,  p-value: < 2.2e-16
```

As can be seen in the coefficients, the one for fiona_farmer, the τ^{ATE}, is 2358, which is measured in dollars and is about 11.5 smaller than the Naive estimator. This means that the average effect of participating in FIONA is an increase in profits of $2358 dollars. As expected, $profits_2 005$ explain part of the profits in 2016 as seen by the statistical significance at the 1% level. The dummies for wheat and lentils are also statistically significant at the 1% and 5% level respectively. Having wheat crops increases profits in 2016 in respect to the base case in $428, lentils crops increase it in $359. Last, for every dollar extra of profit in 2005, we expect profits in 2016 to be 1 dollar higher.

8 - Use an exact matching approach to estimate the effect of FIONA on farmer profits. What variables should you include in the matching procedure? Begin by estimating the answer to the question in (1). Then, estimate the answer to the question in (2). Are these meaningfully different? Would you have expected these results to be the same? Why or why not? What are the strengths and weaknesses of this approach? How do your results differ from what you find if you instead use the naive estimator? From what you found in (8)? Did you run into the Curse of Dimensionality with this analysis? If yes, describe how it affected your approach. If not, describe how the Curse could have generated problems in this setting.

Answer: We need to match fiona with non-fiona farmers in covariates so that we compare units that are equal in both observables. Since we are performing exact matching we don't include continuous variables as profits in 2005. We include crop type and farmer birth year. We can't include district since we have only one group present in each one. Let's consider that the matching estimator will only be able to estimate the ATE from question (1). After dividing the data into cells, we calculate the average profit in 2016 for each treated and untreated group in each cell, then we calculate the weighted average of differences.

We are looking for:

$$\hat{\tau}^{ATE} = \sum_{k=1}^{N} \frac{N_k}{N} \hat{\tau}_k$$

* When estimating ATT, the weight should be the counts of treated in a cell over the total counts of treated individuals.

$$\hat{\tau}^{ATT} = \sum_{k=1}^{N_T} \frac{N_{k,T}}{N_T} \hat{\tau}_k$$

We can calculate the effect of FIONA offering on profits. Since we don't know which farmers are non compliers, we won't be able to calculate the ATT. The results from the matching method are showed next (We use the MatchIt package in R to apply the method):

```
library(MatchIt) #


exact.match <- matchit(formula= fiona_farmer ~ farmer_birth_year + crop_LENTILS + crop_RICE + crop_WHEAT

exact.data <- match.data(exact.match)

y.treat <-
weighted.mean(exact.data$profits_2016[exact.data$fiona_farmer == 1],
exact.data$weights[exact.data$fiona_farmer == 1])

y.cont <-
weighted.mean(exact.data$profits_2016[exact.data$fiona_farmer == 0],
exact.data$weights[exact.data$fiona_farmer == 0])

y.treat - y.cont
```

```
[1] 2384.022
```

The matching ATE estimator is 2384, which is higher than the naive estimator (2369.56) by $14.6 dollars.

The strength of the exact matching method is that we match units that are equal in covariates. We still rely on the assumptions on SOO from before, but we don't need to specify a functinal form since the method is non-parametric. We didn't run into the curse of dimentionality since we didn't use enough variables to stop finding matches, but we could have if we would have included variables such as profits in 2005. Also, including the district would have meant not having any matches and the estimand could not have been calculated.

9 - Based on your results in (8), explain to HARRIS whether or not they should implement a FIONA-like program in Bangladesh. Be sure to tell them the reasoning behind your recommendation.

Answer: The fact that no RCT was performed makes it difficult to obtain unbiased estimators for the ATE. According to the data, the treated units all belong to one district and therefore the effect of the treatment can also be attributed to a farmer being located in THANJAVUR. The problem with the matching estimator, or any SOO design, is that we can't correct for this problem, since if we include district as a covariate for matching we would not have any matches. Based solely on 8, if we assume no unobservables are correlated with the offering of FIONA, we can still recomend the program to HARRIS, since the effect in profits is $2384.