Does defendant gender cause differences in criminal sentence length?

W241 Final Project Notebook; Summer 2019, Section 3

Neha Kumar(neha.kumar), Steve Sanders (steve_sanders), Prabhat Tripathi(ptripathi) @berkeley.edu

August 10, 2019

```
# load packages
library(data.table)
library(foreign)
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
```

```
## Warning: package 'sandwich' was built under R version 3.5.2
```

```
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
#library(ggplot2)
```

Research Question and Background

Criminal sentencing is determined by a judge in a trial. Compounding to the subjectivity and interpretability of the law, judges are fallable to biases like any other individual. The purpose of this study is to determine whether the gender of a criminal has an impact on the length of their sentenced jail time post arrest. Observational studies such as those from Professor Starr et al. from the University of Michigan [1] establish a correlation between gender and severity of punishment. These studies suggest that women receive lighter criminal sentences than men, often because they are perceived to be less threatening than their male counterparts.

Here, we conduct an experiment that examines whether gender has an impact on criminal sentences. A similar study was conducted by Desanttes and Kayson of Iona College in 1997 where sentencing severity was measured as a function of atractiveness, race, and sex [2].

Literature Review and Related Work

Before operationalizing the experiment, we looked into existing research and literature in the related area. Through an experiment involving 160 participants, mostly college students pretenting to be potential jurors, DeSantis et. al. [2] supported their hypothesis that attractiveness, race and sex of the defendants have significant effect on the length of the sentencing. In the finding, women received lighter sentences than men for the same crime. One possible reason cited here is that potential jurors held ideas about chivalry and paternalism. In an observation study where Farrington & Morris [3] analyzed criminal statistics to suggest that men receive relatively more severe sentences than women in magistrates' courts in England and Wales. As authors themselves agreed, this observational study results could be biased because of many factors confounded with the gender of the defendants. After controlling for other factors in multivariate analyses, the effects of sex on sentencing were small in comparison with those of other factors such as offense seriousness and prior record, but were *demonstrably present*. In an another validation of the *Chivalry theory*, Cassia Spohn [4] analyzed the data from felony drug offences in Cook county and indicated that females are significantly less likely than males to be sentenced to prison.

While the first study above was experimental, it tries to address several treatment conditions (attractivess, race and sex each with multiple factors) with relatively small sample size. The other two observational studies lack apple-to-apple comparisions and are subject to ommitted variable bias due to confounding factors, such as different likelihoods between women and men to be arrested in the first place or the prevelance of crime between the two genders. The purpose of this study is to validate the *Chivalry theory* through a field experiment. In other words, we conduct a field experiment that tries just to evaluate role of defendent's gender in the criminal sentencing post-arrest.

Proposed Methodology

Experiment Design

We plan to conduct multi-factor factorial design experiment that asseses both treatment-by-covariate and treatment-by-treatment interactions.

Factors: 2 x 2
- Defendant's Gender: Male, Female
- Crime Severity: Felony, Misdemeanor

Covariates: - Respondant's Gender: Male, Female

We have four treatment conditions (2 X 2) here.

Some important design considerations

- Two factors (Defendant's Gender and Crime Severity) are *uncorrelated*, allowing us later to ignore one factor if it proves to be inconsequential.

- We plan to use F-test to compare CATE between subgroups and Bonferroni p-value correction to compensate for multiple comparisions.

- We decided not to include defendant's picture in the survey in the fear that this might introduced new unintended treatments such as race, attractiveness etc.

- We chose the defendant's name that are easy signal for defendant's sex but, at the same time, are race neutral.

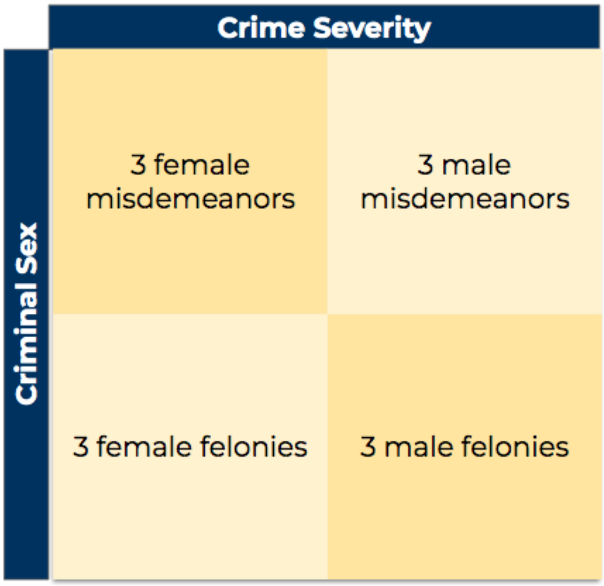


Figure 1: Experimental Matrix

Note that there are nuances in our design when it comes to classifying the study as finding a “within subject” treatment effect versus a “between subjects” treatment effect. Each respondent is observing both male and female profiles. In other words, if we were to say that “seeing a male criminal profile” is the treatment while “seeing a female criminal profile” is the control, each respondent is exposed to both the treatment and control. However, each respondent is not seeing both the treatment and control condition for the same exact crime. In this sense, this design is a between-subjects set up.

Regarding sample size, *a priori* we did not have a good estimate of expected effect size, so we conjectured that we may be looking for an effect on the order of 1 year, and desired a sufficiently large sample size to discern this effect if the standard deviation were twice the effect, or 2 years. Requiring a statistical power of at least 0.80, and at a 5% significance level (an alpha of 0.05), we estimated a need for a minimum of 128 subjects total (64 per group).

```
power.t.test(n = NULL, delta = 1, sd = 2, sig.level = 0.05,
             power =0.8,
             type = "two.sample",
             alternative = "two.sided",
             strict = FALSE, tol = .Machine$double.eps^0.25)
```

```
##
##      Two-sample t test power calculation
##
##               n = 63.76576
##           delta = 1
##              sd = 2
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Operationlizing the Experiment

Similar to DeSantis and Kayson, our team developed a survey comprised of fake criminal profiles. In DeSantis and Kayson's study however, they only took one type of crime (a burglarly case) and varied the race, gender and attractiveness of the defendent, administering one criminal profile per respondent. In our study, we wished to see if different crimes of varying severities and natures experienced different sentencing for males versus females. Thus, we developed a survey of 12 criminal profiles, consisting of 6 different misdemeanors and 6 different felonies. For each survey, half of the criminal profiles were associated with male criminal names, and the other half to female criminal names. Two versions of the survey were administered. A crime associated with a male criminal in one version was associated with a female criminal in the other version and vice versa. This setup allows us to compare male and female average scores for each crime, without the individual survey respondent receiving identical criminal profiles except for the name, which would thereby offer them a moral bias to assign the same sentence length. The complete survey can be accessed via the provided link [5].

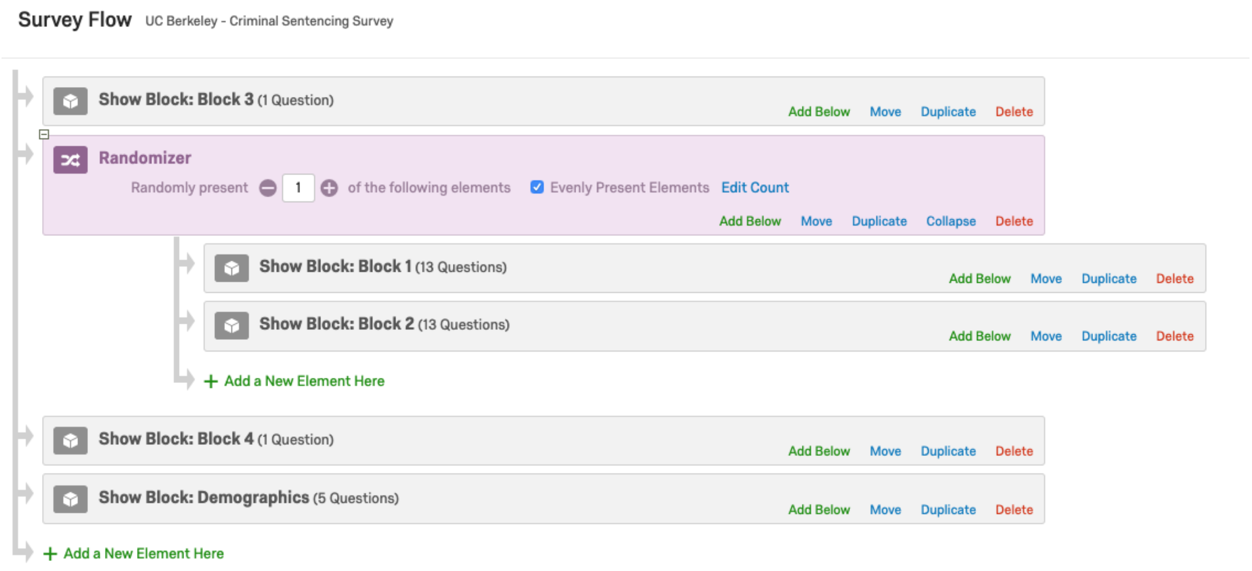
Criminal profiles took the following format.

Name: Naomi A.
Date of Crime: 2019-06-01
Offense Level: FELONY
Offense: ROBBERY/ARMED/FIREARM OR DEADLY WEAPON
Description: The suspect, Naomi A. robbed a woman at gunpoint outside a bar in New York City, taking her purse which contained her wallet and other valuable possessions before fleeing. The woman was not seriously injured, but sustained a sprained wrist from the suspect's grip.

Note that racially ambiguous names were intentionally selected for each criminal profile. We opted to not show photos of the criminals in order to remove the confounding factors of race, attractiveness, and perceived age on criminal sentences.

After showing the respondents the criminal profiles, the survey asks a series of demographic questions: the respondent's gender, zip code, political preference, and whether or not they are in the legal field. While covariates are preferably collected prior to treatment to avoid their results having been impacted from treatment, our group decided to place them following the treatment for the following reasons: (1) Responses to the political preference and legal field question may bias the respondents from responding one way versus another. For example, someone who is not in the legal profession may not take the study as seriously because they may feel underqualified to participate. Similarly, participants may behave different when reminded to consider their political preferences. Some people may intentionally give lighter prison sentences as their political party does not encourage incarceration. (2) we believe the responses to these covariates is robust to the treatment. For example, an individual's gender and zip code are not going to change as a result of taking a short 5 to 10 minute survey. In the event we have a suspicion that our covariates do manage to be influenced by treatment, we can always omit these covariates from our analysis. However, if our covariates were collected before the study and managed to influence our respondent's behavior, then we cannot trust the responses to the durations of jail time for each of the criminal profiles. Since the latter situation is more risky for the integrity of our study, we kept covariate collection at the conclusion of the study.

Initially, our study aimed to survey law students only, as they would be more representative of judges (who are the individuals who actually hand down a criminal sentence). However, from our pilot study, we found it difficult to collect enough responses from law students specifically. Many law professors we emailed were not teaching summer courses, and we had not received responses from any listserves to law students to distribute our survey. Therefore, we decided to open up our responses to the general public to understand if these biases persist in a larger population. We do understand that we are likely increasing our standard errors of the average length of criminal sentencing, as the general public is less likely to be aware of an appropriate sentence length for each "type" of crime. This suggests that we would need to collect a much larger sample to find a treatment effect of being a female criminal, if one does exist. Additionally, we aimed to mitigate the range of answers that could be provided by constraining misdemeanor sentences from 0 to 5 years, and felony sentences from 0 to 20 years.



Survey Flow UC Berkeley - Criminal Sentencing Survey

Show Block: Block 3 (1 Question)

Randomizer

Randomly present 1 of the following elements ☑ Evenly Present Elements Edit Count

Show Block: Block 1 (13 Questions)

Show Block: Block 2 (13 Questions)

+ Add a New Element Here

Show Block: Block 4 (1 Question)

Show Block: Demographics (5 Questions)

+ Add a New Element Here

Figure 2: Survey Flow

We also entertained the idea of blocking versions of the survey by gender of the responsent. However, since each respondent is seeing both male and female criminal profiles, we determined that a large enough sample size should be sufficient to prevent treatment effects being attributed to a specific sub-population. If we only showed one criminal profile to each respondent (similar to Desanttes and Kayson), then we would have opted for blocking. In our final study, we set qualtrics to randomly assign versions of the survey to each respondent.

The survey was distributed by emailing a list of law professors our survey link with a description of the study (without giving away the exact hypothesis we are testing) and requesting that they forward the study to their law students. We also shared our survey on social media channels such as the MIDS Slack channel and public-facing LinkedIn posts. We realize that this suggests that the findings of our study are likely influenced by the types of people who would come across our study through any of these forums. We attempted to diversify the forums used to collect responses; however there are likely underrepresented subpopulations, which challenges the generalizability of our results. We recommend that other studies be replicated to determine whether findings remain consistent with our findings when different populations of respondents are surveyed.

Qualtrics Survey Flow

ROXO Grammar

The survey flow, along with ROXO grammar, can be found in the Figure 3:

The first step was randomization by Qualtrics. This was the “R” notation that makes our experiment results valid as we have randomized covariates between the different survey versions. Following the initial randomization and once participants had been assigned to a survey version, participants were shown 12 criminal profiles in a non random format (6 misdemeanors and 6 felonies, half of which are male or female). This is the N in the ROXO grammar above. Then, for each survey version half of the profiles were male and half were female. Criminal sentences were assigned for both. This is the Xˆ6 and Oˆ12 notation in the above ROXO grammar.

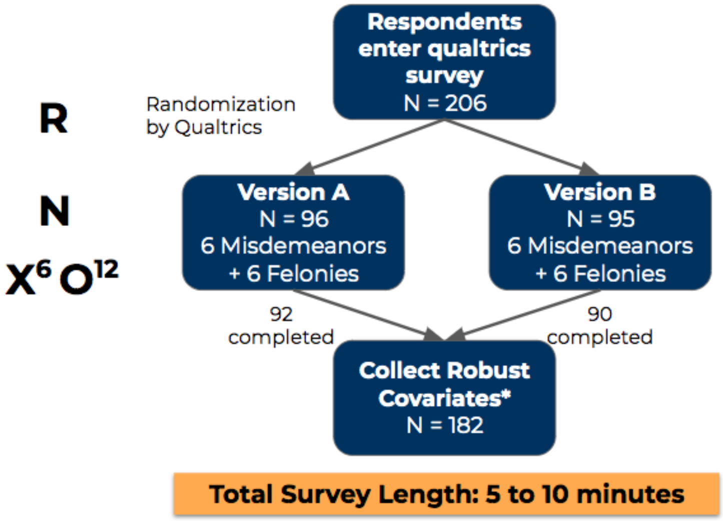

R
N
X6 O12
Randomization by Qualtrics
Respondents enter qualtrics survey N = 206
Version A N = 96 6 Misdemeanors + 6 Felonies
Version B N = 95 6 Misdemeanors + 6 Felonies
92 completed
90 completed
Collect Robust Covariates* N = 182
Total Survey Length: 5 to 10 minutes

Figure 3: Experimental Flow

Analysis Plan

To ensure we do not succumb to selecting a sample size based on "waiting" for a statistically significant result from our data, we decided well in advance that July 30, 2019 will be our last day collecting data in field. The group agreed to close down the survey at 8pm that evening regardless of the sample size to ensure that we are not introducing additional comparisons.

In addition, we have laid out our research plan in order to prevent a fishing expedition. First, we will complete a covariate balance check to ensure that no covariate was over-represented in one version of the survey versus another. Then, we will restructure our dataframe such that each row is an individual crime and sentence. Since each respondent who completed the survey assigned sentences to 12 criminals, this resulting dataframe will have 12 times the number of rows as our original dataframe. Then, we will regress all misdemeanors against whether or not a criminal profile presented a male or female, clustering by the response id of the individual who took the survey. We will also look at whether there is a heterogeneous treatment effect with each of the covariates (respondent gender, respondent location by state, respondent familiarity with the law field, and respondent political leaning).

Pilot Study

Data Prep

First we perform the following steps to clean the data and transform it so it is ready for analysis. Here we first conduct these steps for the pilot study.

```
# Download the CSV of the results. We downloaded the fields with full text here
d_raw <- fread('final_data.csv')
d_raw <- data.table(d_raw)

# Anonymizing data
d_raw <- d_raw[, QD5:=NULL]

# Remove field descriptions
```

```
d_raw <- tail(d_raw, -2)
d_raw[,StartDate:= strptime(StartDate, "%Y-%m-%d %H:%M:%S")]
```

```
## Warning in strptime(StartDate, "%Y-%m-%d %H:%M:%S"): POSIXlt column type
## detected and converted to POSIXct. We do not recommend use of POSIXlt at
## all because it uses 40 bytes to store one date.
```

```
d_raw[,EndDate:= strptime(EndDate, "%Y-%m-%d %H:%M:%S")]
```

```
## Warning in strptime(EndDate, "%Y-%m-%d %H:%M:%S"): POSIXlt column type
## detected and converted to POSIXct. We do not recommend use of POSIXlt at
## all because it uses 40 bytes to store one date.
```

```
# Split pilot study results from main study
d_pilot <- d_raw[StartDate < '2019-07-16 22:00:00' & Status == 'IP Address',]
d_main <- d_raw[StartDate > '2019-07-16 22:00:00' & Status == 'IP Address',]

# Capture Attrition
d_dropout <- d_main[Finished == "False",]

# Only take finished records for analysis
d_pilot <- d_pilot[Finished == "True",]
d_main <- d_main[Finished == "True",]

# Removing records that were the reseachers testing out the link from the pilot study
d_pilot <- d_pilot[IPAddress != "96.60.7.228" & IPAddress != "24.5.86.93" & IPAddress != "136.24.143.47
```

Modifications based on the Pilot Study

For the pilot study, we ran some surveys with respondents next to the researchers while the researchers asked the respondents follow up questions probing on whether the criminal profiles were clear, whether the names were indeed racially ambiguous, and whether the gender was clear from the name. Based on these findings, we switched out names where the gender of the criminal was less obvious and the name of the criminal suggested the criminal's race. Running through this exercise gives us additional confidence that the different names only suggested different genders and did not suggest any other differences in the criminals. However, we cannot control for associations certain individuals may have with certain names. Therefore, by having a total of 12 profiles, each with different names, we aim to minimize the effect of a particular name having an unusually strong influence on a respondent.

Additionally, surveys were administered to respondents without the researchers sitting next to them, simulating the flow of the actual survey. This testing ensured that the randomization by qualtrics was indeed working as expected.

Main Study Findings

Data Prep

```
# Now melting the data table so we see data on the question level
d_melt_main = melt(d_main, id.vars = c("RecordedDate", "ResponseId","QD1","QD2","QD3","QD4"),
                measure.vars = c("Q1A_2", "Q1B_2"
                                 , "Q2A_2","Q2B_2"
                                 , "Q3A_2","Q3B_2"
```

```r
                                 , "Q4A_2","Q4B_2"
                                 , "Q5A_2","Q5B_2"
                                 , "Q6A_2","Q6B_2"
                                 , "Q7A_2","Q7B_2"
                                 , "Q8A_2","Q8B_2"
                                 , "Q9A_2","Q9B_2"
                                 , "Q10A_2","Q10B_2"
                                 , "Q11A_2","Q11B_2"
                                 , "Q12A_2","Q12B_2"))
names(d_melt_main) <- c("RecordedDate", "ResponseId","Field", "ZipCode","Gender","Party","Question","Val
d_melt_main[, Version := ifelse(grepl('A',Question),1,0)]
d_melt_main[, male_criminal := ifelse(Question %in% c("Q3A_2", "Q10A_2","Q12A_2", "Q1A_2" ,"Q4A_2", "Q6
d_melt_main[, is_felony := ifelse(Question %in% c("Q1A_2","Q4A_2","Q5A_2","Q6A_2","Q9A_2","Q11A_2", "Q1
head(d_melt_main)
```

```
##           RecordedDate        ResponseId
## 1: 2019-07-16 22:24:28 R_1r7bOULQFOUE8lZ
## 2: 2019-07-16 22:25:08 R_3qCdUK8V0IcdVbD
## 3: 2019-07-16 22:43:04 R_3Jr5X6jmcAYal2b
## 4: 2019-07-16 22:49:15 R_RwN9U7fsxoMME49
## 5: 2019-07-16 22:59:30 R_3Ok9fxZ8L5Mso6H
## 6: 2019-07-16 23:14:33 R_1Cm5rOHjnz5dFwI
##                                             Field ZipCode Gender
## 1: Other or Not Associated with legal profession   95120 Female
## 2: Other or Not Associated with legal profession   95124 Female
## 3: Other or Not Associated with legal profession   94025   Male
## 4: Other or Not Associated with legal profession   94123   Male
## 5: Other or Not Associated with legal profession   94024 Female
## 6: Other or Not Associated with legal profession   94103   Male
##          Party Question Value Version male_criminal is_felony
## 1:    Democrat    Q1A_2             1             1         1
## 2:    Democrat    Q1A_2     5       1             1         1
## 3: Independent    Q1A_2     5       1             1         1
## 4:    Democrat    Q1A_2             1             1         1
## 5:    Moderate    Q1A_2     4       1             1         1
## 6: Independent    Q1A_2             1             1         1
```

```r
# Convert characters in 'Value' column to numeric
strTmp = c('Value')
d_melt_main[, (strTmp) := lapply(.SD, as.numeric), .SDcols = strTmp]
#d_melt_main
```

```r
# Omit rows where 'Value' is missing, since these were not typically assigned to the subjects.
dp_final = na.omit(d_melt_main, cols='Value')
#dp_final
```

```r
sample_size_final <- dp_final[ , .N/12]
#sample_size_final
```

The number of respondents in the final sample is 182.

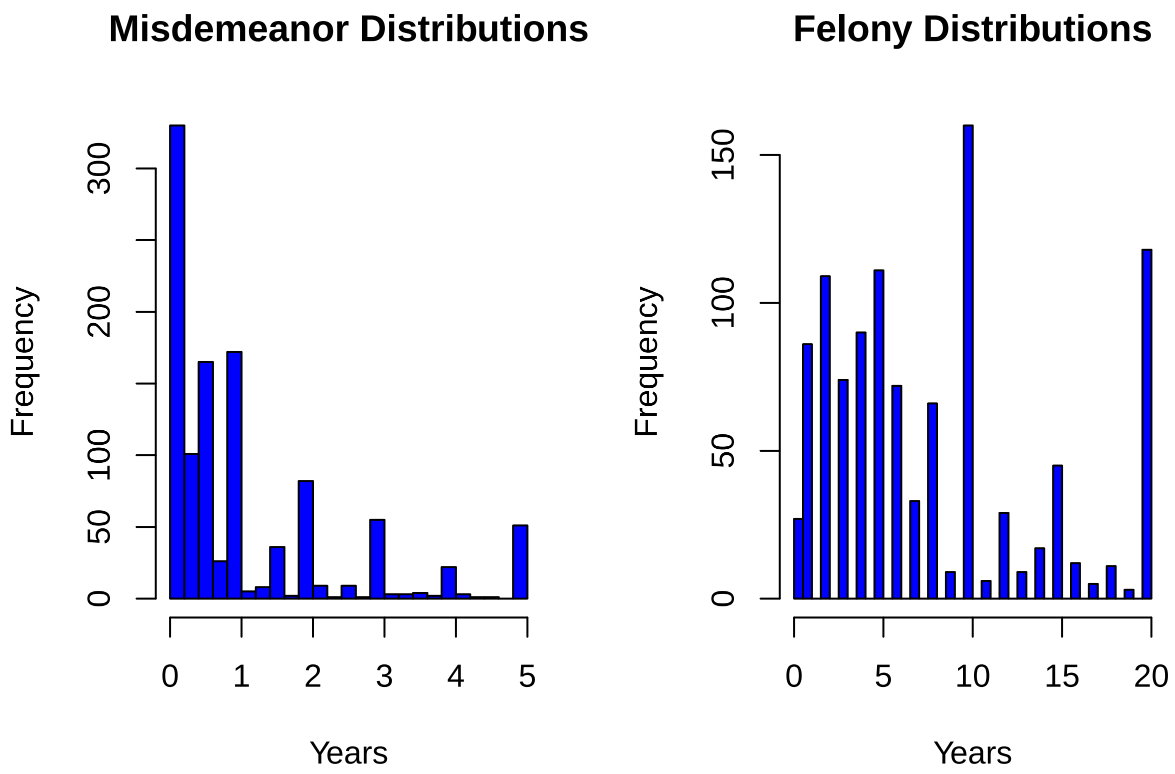


Figure 4: Sentencing in Years distribution by crime type

EDA

Now that the main study data has been prepped, we examine the distribution of the data both to ensure that the study design ran as aniticpated and to better understand the population of respondents that comprise the sample.

```
# Plot histograms of sentences ('Values') for all misdemeanors and all felonies
par(mfrow=c(1,2))
hist(dp_final[is_felony==0, Value], breaks=30, col='blue', main = 'Misdemeanor Distributions', xlab = 'Y
hist(dp_final[is_felony==1, Value], breaks=30, col='blue', main = 'Felony Distributions', xlab = 'Years
```

First, we look at the distributions of the response variable among misdemeanors and felonies. For the misdemeanors, the distribution is quite skewed right, with most values clustered in the 0 to 1 year mark. Peaks are seen at whole number years following one year, aligning with expectations as users are more likely to select whole number years over 1 year than fractions of a year.

The distribution of the felonies also tends to have more data at the bottom half of the distribution than the top half. However, the degree of skew is much less than that of misdemeanors. Additionally, there are noticable peaks at 10 and 20 years, as 10 is a middle-of-the-road metric that users would gravitate toward given that they were provided a scale from 0 to 20. The peak at 20 could be explained by certain crimes being so heinous to a particular respondent that they assign the maximum penalty possible. A smaller peak is observed at the 15 year mark, as this again is an easy benchmark for users (3/4 of the maximum severity).

```
barplot(prop.table(table(dp_final[, Version])), main = 'Split between crimes from each version of the su
```

Next, a bar plot showing how many respondents were assigned one version of the survey versus another was run to verify that half of respondents were indeed assigned to take each version of the survey. Understanding this is critical to validating the experiment ran as expected. Based on the results above, the Qualtrics randomization did indeed put half of respondents into each version of the survey. (Note, as described above,

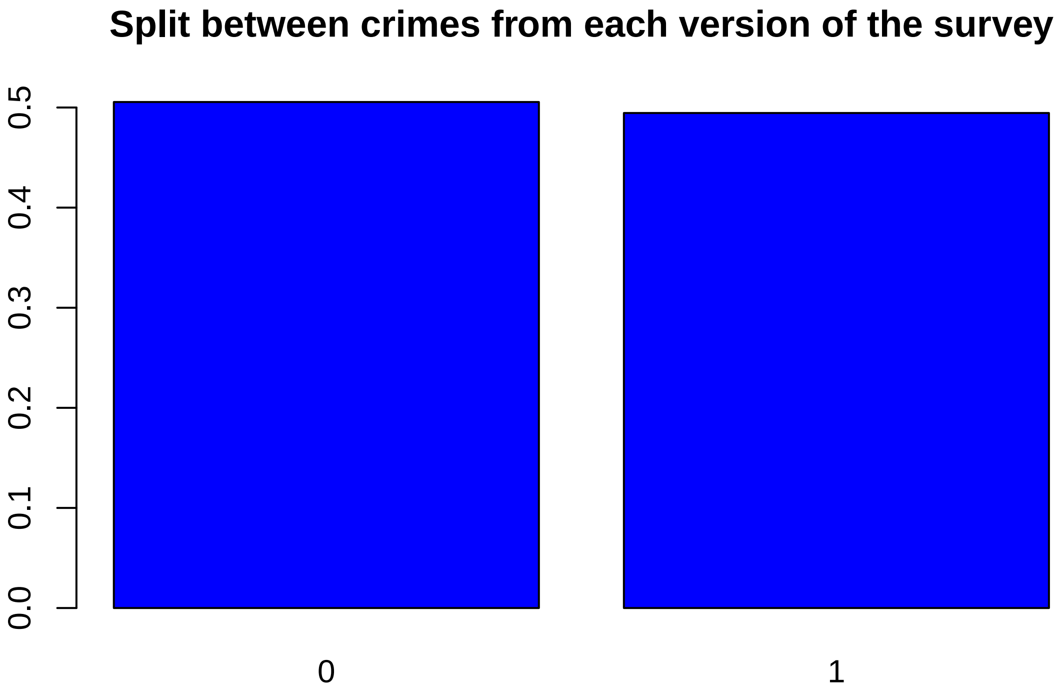


Figure 5: Survey Versions split

both versions of the surveys were identical except the genders for each crime were swapped).

```
barplot(prop.table(table(dp_final[, is_felony])), main = "Misdemeanor / Felony Split (Felony = 1)", col=
```

Again, to validate the set up of the survey, the bar plot above shows that exactly half of criminal sentences corresponded to misdemeanors and the other half corresponds to felonies. This is in line with expectations as each version of the survey was designed with exactly 6 misdeameanors and 6 felonies.

```
barplot(prop.table(table(dp_final[, Gender])), main = "Respondent Gender Breakdown", col='blue')
```

```
prop_males <- dp_final[ Gender == 'Male', .N] / dp_final[, .N]
```

The bar plot above shows the distribution of the **respondent's** gender. From the above, more females took the study than males. That being said, there is still a fairly strong representation of male respondents. In the study, **0.4120879** of respondents were male.

```
barplot(prop.table(table(dp_final[, Field])), main = "Respondents in a Law Field", col='blue')
```

The original design of the study aimed to collect responses only from those in the law field as this would be more representative of the judges who would be handing sentences to criminals. However, from the pilot study, the researchers learned that it would be infeasible to collect sufficient sample given the duration of the study. In particular, emailing criminal law professors to send the survey out to their students was less fruitful than hoped as a data collection method, as most professors were not teaching classes during the summer term.

```
barplot(prop.table(table(dp_final[, Party])), main = "Political Affiliation Split", col='blue')
```

Now observing a split of political leaning among the respondents, we see that the sample was overall left-leaning. While effort was taken to find respondents across a diversity of networks, the online channels that were pursued likely do have more respondents that identify themselves as Democrats. Note this information was captured because individuals' opinions on prison sentences may likely depend on their political preference. For example, Democrats may believe in shorter jail times as they could hold the belief that prison overcrowding,

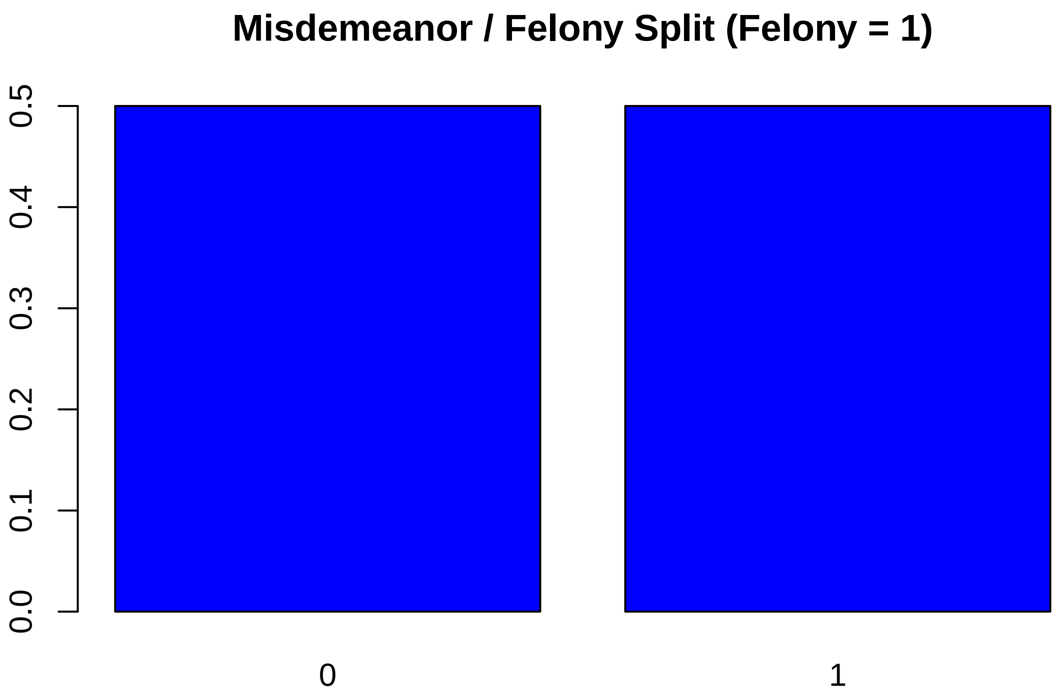


Figure 6: Crime Type split

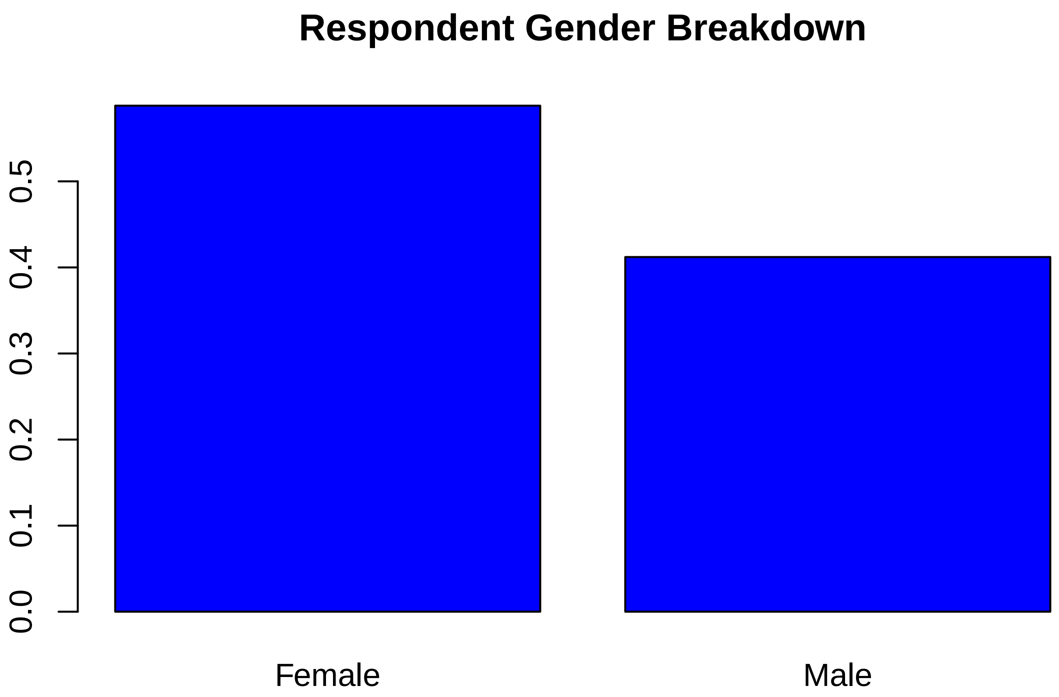


Figure 7: Subjects gender

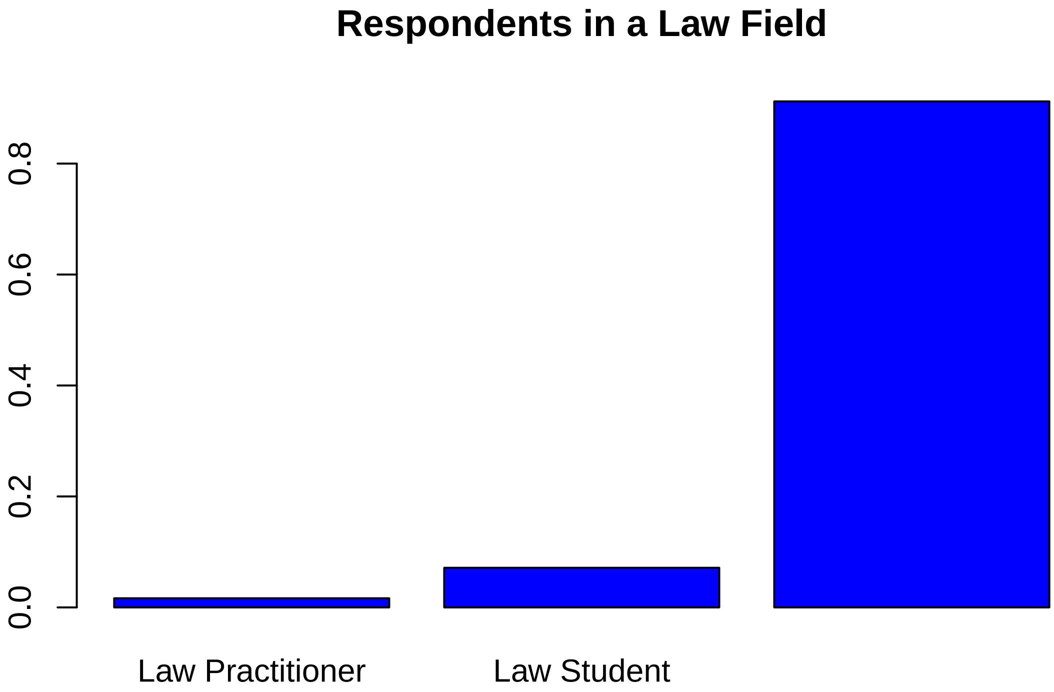


Figure 8: Subjects Law profession association

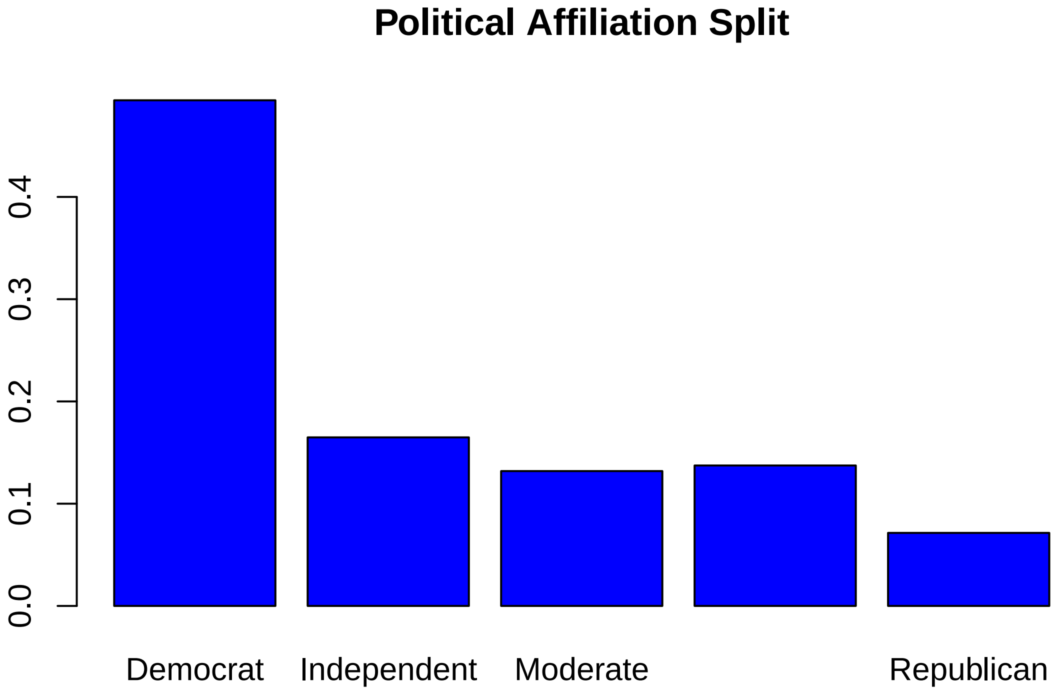


Figure 9: Subjects Political Affiliations

or certain crimes may be a greater offense to people of one political leaning versus another.

As there are more left-leaning respondents in the study, the study results cannot necessarily be generalized to all Americans, since the right-leaning side of the political spectrum is under-represented. Still, even though there is an uneven distribution, assignment of each version of the survey was random, allowing an apples-to-apples comparison between the treatment and control.

```r
# Download the CSV of the zip codes, found in a publicly accissble database
zips <- fread('zipcodes.csv')
zips <- data.table(zips)
zips <- zips[, c("Zipcode", "State"), with=FALSE]
setnames(zips, old=c("Zipcode","State"), new=c("ZipCode", "State"))
zips <- zips[, ZipCode:=as.character(ZipCode)]
dt_with_location <- merge(dp_final, zips, all.x=TRUE)

state_agg <- dt_with_location[, .(count = .N), by = State]
state_agg[, num_respondents := count / 12]
state_agg
```

```
##     State count num_respondents
##  1:  <NA>   132              11
##  2:    NY   444              37
##  3:    PA    48               4
##  4:    DC    12               1
##  5:    VA    96               8
##  6:    GA   252              21
##  7:    FL    24               2
##  8:    ND    24               2
##  9:    CT    12               1
## 10:    IL    24               2
## 11:    MO    72               6
## 12:    KS    60               5
## 13:    LA    12               1
## 14:    TX    60               5
## 15:    CO   168              14
## 16:    ID    36               3
## 17:    UT    12               1
## 18:    AZ   936              78
## 19:    CA  1584             132
## 20:    OR   120              10
```

Lastly, we examined the distribution of respondents by state. The above information was gathered from matching the Zip Code data to a publicly-available dataset of zipcodes with location data [7]. From the above analysis, we see that most respondents (over half) were from California, explaining the left-leaning skewness. This could be explained by a higher response rate within surveyors' professional networks (2 of the 3 researchers of this study are based in California) even though attempts were made to bring in respondents outside of the researchers' networks. This knowledge allows us to contextualize the findings as they may be more applicable to California residents than the broader United States.

Notes on Attrition and Non-compliance

In our study, there were 24 individuals who did not complete the survey. This was classified as attrition. For the most part, attrition occurred when respondents reached the landing page of the survey but did not even start the survey to begin with. As this attrition happened before respondents were randomized into one of two groups, we are not concerned that it biased our results. There were a total of 10 respondents who started the survey but did not complete it.

```
version_a_drop <- d_dropout[Q2A_2 >= 0 , .N]
version_b_drop <- d_dropout[Q2B_2 >= 0 , .N]
```

Of these respondents, 4 of them were assigned to one version and 6 of them were assigned to another version. This doesn't suggest a strong attrition for one version of the survey versus another, keeping our overall sample size in mind. Additionally, since each version of the survey had some mix of treatment and control, we are less likely to attribute the attrition to differential effects of the treatment condition (seeing a criminal profile of one gender versus another).

Regarding compliance, with the study set up there is no concept of a user refusing treatment other than dropping out of the study / attrition. Therefore we do not need to be concerned of non-compliance effects.

Covariate Balance Check

```
cov_balance_mod <- dp_final[, lm(male_criminal ~ Field)]
robust_cov_balance_mod <- coeftest(cov_balance_mod,vcovCL(cov_balance_mod, cluster=dp_final$ResponseId)
robust_cov_balance_mod
```

```
##
## t test of coefficients:
##
##                                                      Estimate Std. Error
## (Intercept)                                        5.0000e-01 1.8051e-14
## FieldLaw Student                                   4.2242e-14 1.8051e-14
## FieldOther or Not Associated with legal profession 4.1644e-14 1.8054e-14
##                                                       t value Pr(>|t|)
## (Intercept)                                        2.7699e+13  < 2e-16 ***
## FieldLaw Student                                   2.3401e+00  0.01937 *
## FieldOther or Not Associated with legal profession 2.3066e+00  0.02117 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cov_balance_mod2 <- dp_final[, lm(male_criminal ~ Gender)]
robust_cov_balance_mod2 <- coeftest(cov_balance_mod2,vcovCL(cov_balance_mod2, cluster=dp_final$Response
robust_cov_balance_mod2
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error     t value Pr(>|t|)
## (Intercept)  5.0000e-01  5.4490e-16  9.1761e+14   <2e-16 ***
## GenderMale  -5.3333e-16  5.4676e-16 -9.7540e-01   0.3295
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cov_balance_mod3 <- dp_final[, lm(male_criminal ~ Party)]
robust_cov_balance_mod3 <- coeftest(cov_balance_mod3,vcovCL(cov_balance_mod3, cluster=dp_final$Response
robust_cov_balance_mod3
```

```
##
## t test of coefficients:
##
##                                Estimate  Std. Error     t value Pr(>|t|)
## (Intercept)                  5.0000e-01  6.4580e-16  7.7423e+14  < 2e-16
## PartyIndependent            -1.8467e-15  6.5599e-16 -2.8151e+00  0.00492
## PartyModerate                1.8084e-15  7.1618e-16  2.5250e+00  0.01164
## PartyPrefer not to disclose -1.4205e-15  6.4925e-16 -2.1879e+00  0.02878
## PartyRepublican             -1.5738e-15  6.9151e-16 -2.2759e+00  0.02295
##
## (Intercept)                 ***
## PartyIndependent            **
## PartyModerate               *
## PartyPrefer not to disclose *
## PartyRepublican             *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
stargazer(robust_cov_balance_mod,robust_cov_balance_mod2,robust_cov_balance_mod3, title = "Covariate Ba
```

Table 1: Covariate Balance Check

	Dep. variable: Criminal Gender		
	(1)	(2)	(3)
FieldLaw Student	0.000^{**}		
	(0.000)		
FieldOther or Not Associated with legal profession	0.000^{**}		
	(0.000)		
GenderMale		−0.000	
		(0.000)	
PartyIndependent			-0.000^{***}
			(0.000)
PartyModerate			0.000^{**}
			(0.000)
PartyPrefer not to disclose			-0.000^{**}
			(0.000)
PartyRepublican			-0.000^{**}
			(0.000)
Constant	0.500^{***}	0.500^{***}	0.500^{***}
	(0.000)	(0.000)	(0.000)
Note:	*p<0.1; **p<0.05; ***p<0.01		

From the table above, the predictive power of each of each covariate is close to 0, with the y intercept of each covariate individually a clean 0.5. While some treatment effects are "statistically significant," note there is barely any practical significance to the above coefficients. Going to 3 decimal places, the standard error in all cases is till 0.

In other words, there is no actual effect of being of one covariate and being assigned to more male criminal profiles. This is expected, as each survey had the same number of male and female criminal profiles. All respondents who completed the survey saw exactly 6 male and 6 female criminal profiles. It does not matter what other covariates are true, the treatment will still be administered exactly hafl the time, in line with the study design.

To test our randomization, we can look at our covariates measured against the survey version on the dataframe prior to melting the information to the individual crime level.

```r
# Creating a "Version"" column that represents the survey version
d_main[, Version := ifelse(Q2A_2 >= 0 , 1 ,0)]

survey_version_mod1 <- d_main[, lm(Version ~ QD1)]
robust_survey_version_mod1 <- coeftest(survey_version_mod1,vcovHC(survey_version_mod1))
robust_survey_version_mod1
```

```
##
## t test of coefficients:
##
##                                                     Estimate  Std. Error
## (Intercept)                                      -1.7735e-15  2.5810e-08
## QD1Law Student                                    4.6154e-01  1.4979e-01
## QD1Other or Not Associated with legal profession  5.0602e-01  3.9040e-02
##                                                  t value  Pr(>|t|)
## (Intercept)                                       0.0000  1.000000
## QD1Law Student                                    3.0813  0.002387 **
## QD1Other or Not Associated with legal profession 12.9617 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
survey_version_mod2 <- d_main[, lm(Version ~ QD3)]
robust_survey_version_mod2 <- coeftest(survey_version_mod2,vcovHC(survey_version_mod2))
robust_survey_version_mod2
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.532710   0.048688 10.9412   <2e-16 ***
## QD3Male     -0.092710   0.075798 -1.2231   0.2229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
survey_version_mod3 <- d_main[, lm(Version ~ QD4)]
robust_survey_version_mod3 <- coeftest(survey_version_mod3,vcovHC(survey_version_mod3))
robust_survey_version_mod3
```

```
##
## t test of coefficients:
##
##                             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                 0.477778   0.053244  8.9733 4.105e-16 ***
## QD4Independent              0.055556   0.108228  0.5133    0.6084
## QD4Moderate                 0.063889   0.118736  0.5381    0.5912
## QD4Prefer not to disclose -0.077778   0.115116 -0.6756    0.5001
## QD4Republican               0.137607   0.155571  0.8845    0.3776
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
stargazer(robust_survey_version_mod1,survey_version_mod2,survey_version_mod3, title = "Randomization Ch
```

Table 2: Randomization Check

	Dep. variable:Survey Version		
		Version	
	coefficient test	*OLS*	
	(1)	(2)	(3)
QD1Law Student	0.462^{***}		
	(0.150)		
QD1Other or Not Associated with legal profession	0.506^{***}		
	(0.039)		
QD3Male		−0.093	
		(0.075)	
QD4Independent			0.056
			(0.106)
QD4Moderate			0.064
			(0.116)
QD4Prefer not to disclose			−0.078
			(0.114)
QD4Republican			0.138
			(0.150)
Constant	−0.000	0.533^{***}	0.478^{***}
	(0.00000)	(0.048)	(0.053)
Observations		182	182
R^2		0.008	0.012
Adjusted R^2		0.003	−0.011
Residual Std. Error		0.501 (df = 180)	0.504 (df = 177)
F Statistic		1.512 (df = 1; 180)	0.529 (df = 4; 177)
Note:			*p<0.1; **p<0.05; ***p<0.01

Based on the regression results here, we see that for Regression 1, the baseline calculation assumes the respondent is a law practitioner. There were 3 such respondents in our survey and all of them happened to be assigned to one version of the survey. As such, we cannot assume the likelihood of being a Law student or unaffiliated with the law profession is significant. In fact, these are each roughly 50%, suggesting the randomization did indeed work correctly.

Regressions 2 and 3 both demonstrate that there is no statistically significant difference in the likelihood of being assigned one version of the survey versus another based on the respondent's political party and gender.

Regression Results for Sentence Length

In our treatment effect analysis, we built and analyzed three regression models with the sentence length as the dependent variable.

In the *first* model, we checked the effect of the criminal gender in this experiment as a standalone regressor. Note that with good randomization, we would not expect the treatment effect to change appreciably as covariates are added in subsequent models. In this study, however, we have two treatment variables, so in theory we would baseline with both treatment variables. For simplicity, we ran an initial baseline using only criminal gender as a baseline. In order to estimate robust standard errors, we use clustered standard errors with the ResponseID as the clustering variable; the rationale for this is that each individual is answering multiple survey questions (12 in the present study).

```
# Model the criminal gender treatment effect by itself
mod1 = dp_final[ , lm(Value ~ male_criminal)]
#summary(mod1)
mod1_robust_results <- coeftest(mod1,vcovCL(mod1, cluster=dp_final$ResponseId))
mod1_robust_results
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.29973    0.19726 21.7971  < 2e-16 ***
## male_criminal  0.40934    0.21204  1.9305  0.05368 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Examining the results of the first regression model, we observe that, overall, male criminals receive sentences that are 0.41 years longer on average than female criminals, although the p-value is 0.05368, so we do not have statistical significance at the 5% level.

In the *second* model, which we called the “kitchen sink model”, we combined treatments, covariates and their interactions. We modeled the effects of the two treatment variables, criminal gender and crime type, along with their interaction, plus respondent gender covariate and the interaction between respondent gender and criminal gender. Note that given the EDA results, we did not include covariates for location, political leaning, or law field, since these covariates are dominated by a single level, and the non-dominant subgroups have very small sample sizes. Thus, we prefer not to add multiple comparisons given the low likelihood of useful information with the splintered subgroups.

```
# Model treatments and Gender covariate, plus interactions
mod2 = dp_final[ , lm(Value ~ male_criminal + is_felony + male_criminal*is_felony + Gender + male_crimin
mod2_robust_results <- coeftest(mod2,vcovCL(mod2, cluster=dp_final$ResponseId))
mod2_robust_results
```

```
##
## t test of coefficients:
##
##                          Estimate Std. Error t value  Pr(>|t|)
## (Intercept)               1.31278    0.18597  7.0592 2.242e-12 ***
## male_criminal            -0.13873    0.21609 -0.6420   0.52094
## is_felony                 6.48846    0.29784 21.7849 < 2.2e-16 ***
## GenderMale               -0.62434    0.39946 -1.5630   0.11820
## male_criminal:is_felony   0.68681    0.30650  2.2409   0.02514 *
## male_criminal:GenderMale  0.49666    0.41749  1.1896   0.23432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the second regression model, we observe that, as expected, the crime type (misdemeanor vs. felony) is the dominant factor influencing the sentence length, with felony sentences averaging 6.5 years longer than misdemeanor sentences. By adding the crime type treatment variable to the regression, the average treatment effect for the criminal gender is reduced to nearly zero (-0.14) and is clearly statistically insignificant. However, we do observe a *statistically significant* interaction effect between the criminal gender and the crime type, with an interaction coefficient of 0.69, meaning longer sentences for male criminals relative to female criminals. In this model, the respondent gender is not significant, and there is no heterogeneous treatment effect based on *respondent* gender.

In the *third* and final linear interaction model, we run a regression in the "middle" of the specification spectrum by including only the two treatment variables and their interaction.

```
mod3 = dp_final[ , lm(Value ~ male_criminal + is_felony + male_criminal*is_felony)]
mod3_robust_results <- coeftest(mod3,vcovCL(mod3, cluster=dp_final$ResponseId))
mod3_robust_results
```

```
##
## t test of coefficients:
##
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.055495   0.076846 13.7353  < 2e-16 ***
## male_criminal           0.065934   0.092010  0.7166  0.47370
## is_felony               6.488462   0.297706 21.7949  < 2e-16 ***
## male_criminal:is_felony 0.686813   0.306356  2.2419  0.02507 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results of the third model indicate that the average treatment effects for the two treatment variables as well as the interaction between them do not change appreciably after removing the respondent gender covariate. This is additional confirmation that our treatment assignment randomization is *robust*. One may remark that it may be surprising that the coefficient for the interaction term for the is_felony and male_criminal term did not change at all. However, when considering that our study design forced that each respondent see the exact same number of male vs female profiles and felony vs misdemeanor profiles, we can reasonably expect that respondent covariates will have a minimal impact on our coefficients.

Next, we perform an F-test to compare model 2 and model 3 to gauge if one model is significantly more performant relative to the other.

```
mod_compare = anova(mod2, mod3, test='F')
mod_compare
```

```
## Analysis of Variance Table
##
## Model 1: Value ~ male_criminal + is_felony + male_criminal * is_felony +
##     Gender + male_criminal * Gender
## Model 2: Value ~ male_criminal + is_felony + male_criminal * is_felony
##   Res.Df   RSS Df Sum of Sq     F  Pr(>F)
## 1   2178 39891
## 2   2180 39999 -2   -107.44 2.933 0.05345 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-test results in a p-value of 0.05345, so we cannot reject the null hypothesis that the two models perform equally well.

Finally, we summarize the three regression models in the following table.

```
stargazer(mod1_robust_results, mod2_robust_results, mod3_robust_results, title = "Regressions for Sente
```

Table 3: Regressions for Sentence Length (Years)

	Dependent variable: Sentence Length (Years)		
	(1)	(2)	(3)
male_criminal	0.409^{*}	-0.139	0.066
	(0.212)	(0.216)	(0.092)
is_felony		6.488^{***}	6.488^{***}
		(0.298)	(0.298)
GenderMale		-0.624	
		(0.399)	
male_criminal:is_felony		0.687^{**}	0.687^{**}
		(0.306)	(0.306)
male_criminal:GenderMale		0.497	
		(0.417)	
Constant	4.300^{***}	1.313^{***}	1.055^{***}
	(0.197)	(0.186)	(0.077)
Note:			*p<0.1; **p<0.05; ***p<0.01

Regarding multiple comparisons, assuming that we made two comparisons in model 2 and model 3 directed toward the null hypothesis that the interaction between the male_criminal and is_felony treatment variatibles is zero, then we can examine the consequence of applying the Bonferroni correction or the Holm-Bonferroni correction.[7] The Bonferroni correction, which is the most conservative approach, reduces the significance level, α by the number of comparisons, m. Thus, $\alpha^* = \alpha/m$. Given two comparisons, therefore, we would need our p-value to be less than 0.025 in order to reject the null hypothesis. We see, for both models, that the p-value for this interaction term is slightly greater than this (0.02514 and 0.02507 for models 2 and 3, respectively). Therefore, under this most conservative method, we may have to concede the "strict" proclamation of statistical significance. We note, however, that this strict binary approach is less desireable in modern applied statistics.

The Holm-Bonferroni method is less conservative in its approach relative to the Bonferroni adjustment. In this method, $\alpha^* = \alpha/(m+1-k)$, where k is the minimal index such that $P_{(k)} > \frac{\alpha}{m+1-k}$. In this study, the second of the two comparisons, with $k = 2$, will result in a rejection of the null hypothesis. (In this particular case, the Hochberg adjustment, which uses the same α^* calculation as Holm, but inverts the ranking order, also leads to rejecting the null.)

The interaction plot for the two treatment variables is shown in Figure 10. This illustrates the lack of effect of criminal gender for misdemeanor crimes, and the statistically significant difference in sentence length between male and female criminals for felony crimes. On average, male felons received sentences that were 0.75 years longer than sentences for female felons. (The 0.75 year difference is comprised of the 0.687 year interaction effect plus the 0.066 year main effect of male criminal gender.) Note that error bars in this plot correspond to the confidence intervals obtained by using the "predict" method on model 3, with interval="confidence". This results in a 95% confidence interval span of 0.72 years.

```
newdata = data.frame(c(male_criminal=0, is_felony=0), c(male_criminal=0, is_felony=1), c(male_criminal=
plot_pts = unique(predict(mod3, interval='confidence', newdata = newdata))
```

```
## Warning: 'newdata' had 2 rows but variables found have 2184 rows
```

```
plot_pts
```

```
##          fit      lwr      upr
## 1   8.296703 7.937213 8.656194
## 91  7.543956 7.184465 7.903447
## 183 1.055495 0.696004 1.414985
## 273 1.121429 0.761938 1.480919
```

```
male_felon = plot_pts[1,1]
#male_felon  #8.296703
female_felon = plot_pts[2,1]
#female_felon  #7.543956
male_misd = plot_pts[4,1]
#male_misd  #1.121429
female_misd = plot_pts[3,1]
#female_misd  #1.055495

d2 = data.table(id=1:2)
d2[ , is_felony := c(as.integer(0), as.integer(1))]
#d2[ , Crime := c("Misdemeanor", "Felony")]
d2[ , male := c(male_misd, male_felon)]
d2[ , female := c(female_misd, female_felon)]
d2[ , lower_ci_male := c(plot_pts[4,2], plot_pts[1,2])]
d2[ , upper_ci_male := c(plot_pts[4,3], plot_pts[1,3])]
d2[ , lower_ci_female := c(plot_pts[3,2], plot_pts[2,2])]
d2[ , upper_ci_female := c(plot_pts[3,3], plot_pts[2,3])]
d2
```

```
##    id is_felony     male   female lower_ci_male upper_ci_male
## 1:  1         0 1.121429 1.055495      0.761938      1.480919
## 2:  2         1 8.296703 7.543956      7.937213      8.656194
##    lower_ci_female upper_ci_female
## 1:        0.696004        1.414985
## 2:        7.184465        7.903447
```

```
x = d2[ , is_felony]
y1 = d2[ , male]
y2 = d2[ , female]

matplot(x,cbind(y1, y2),pch=c(1, 16),xlab="is_felony",ylab="Sentence Length (Years)", main="Average Resu
axis(side = 1, at = c(0,1))
legend(0,9,legend=c("male","female"),pch=c(1,16), col=c(2,1))
arrows(x, d2[ ,lower_ci_male], x, d2[ ,upper_ci_male], length=0.05, angle=90, code=3, col=2)
arrows(x, d2[ ,lower_ci_female], x, d2[ ,upper_ci_female], length=0.05, angle=90, code=3, col=1)
```

Discussion

The experimental design and survey design in this study were conducive to effective randomization, balanced treatments, and minimal non-compliance and attrition. This enabled the subsequent analysis to be conducted

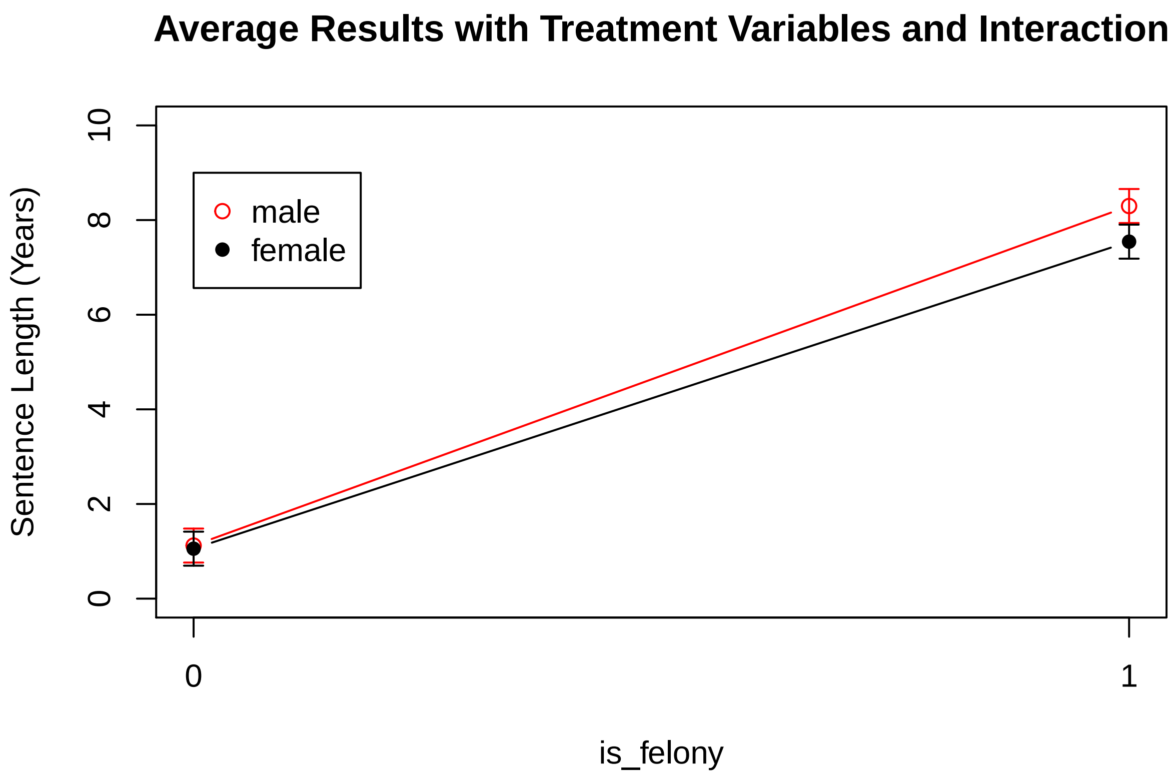


Figure 10: Interaction Plot

efficiently.

We observed results that were statistically significant and support our hypothesis. From the regression analysis, we see a clear interaction effect of criminal gender with crime type. Males convicted of felony crimes received sentences that were 0.75 years (9 months, or approximately 10%) longer on average relative to female felons. It is possible that for this study, the durations of the misdemeanor sentences were so short that they obscure the signal due to the criminal gender. That is, the EDA showed that the vast majority of all misdemeanor sentences were less that one year in length, so it's possible that any signal that may exist in misdemeanor sentences is obscured by the additional noise; it's also possible that the survey respondents were generally indifferent to criminal genders in the case of misdemeanor crimes. For the felony sentences, extending the range beyond 20 years may have resulted in even more distinction between male and female criminal sentences, since the EDA histogram indicated a peak at the 20-year sentence level, where the input was truncated. Future studies may explore these questions in more detail.

It is important to note that our study only distinguished criminal gender through names. We did not include a photo to draw attention to the gender of the criminal in order to avoid confounding variables such as perceived aged, race and attractiveness. As there was no photo displayed however, it is still likely that a respondent could have completed the survey and barely noticing the gendered names. Even still, we saw there was a noticeable effect with criminal sentences based on names alone. Future studies could include photographs and we would expect the effect of gender on sentences to be even stronger in those cases.

Overall, the effect that we observe for male felons in this study is substantial, and should be considered a cautionary signal for those in the justice system, and in society, concerned with fair, unbiased sentences lengths in criminal justice proceedings. While the demographic representation of our sample was limited, it nonetheless represents a sizable segment of society, and has important implications.

Areas of Improvement and Future Work Proposed

Based on the lessons learned in the current study, future research in this area should consider: (1) capturing more diverse respondents to broaden the demographics represented, such as location, field of study/work, and political affiliations, (2) extending the range of choices for felony sentences, (3) recruiting law students for surveys during the fall or spring semesters when classes are in session, and (4) including representative gender-matched photographs of the criminals in the survey's crime scenarios.

References

[1] Prof. Starr's Research Shows Large Unexplained Gender Disparities In Federal Criminal Cases. http://www.law.umich.edu/newsandinfo/features/Pages/starr_gender_disparities.aspx. Michigan Law, University of Michigan. Nov 2012.

[2] DeSantis, A., & Kayson, W. A. (1997). Defendants' characteristics of attractiveness, race, and sex and sentencing decisions. Psychological Reports, 81(2), 679-683.

[3] DAVID P. FARRINGTON, ALLISON M. MORRIS, SEX, SENTENCING AND RECONVIGTION, The British Journal of Criminology, Volume 23, Issue 3, July 1983, Pages 229–248, https://doi.org/10.1093/oxfordjournals.bjc.a047377

[4] Spohn, C. (1999). Gender and Sentencing of Drug Offenders: Is Chivalry Dead? Criminal Justice Policy Review, 9(3–4), 365–399. https://doi.org/10.1177/088740349900900305

[5] Survey Link for team project: https://berkeley.qualtrics.com/jfe/form/SV_9HqoecNRj4tjaYJ

[6] A Free Zip Code Database. http://federalgovernmentzipcodes.us/download.html

[7] Shi-Yi Chen, Zhe Feng, and Xiaolian Yi, A general introduction to adjustment for multiple comparisons, J Thorac Dis. 2017 Jun; 9(6): 1725–1729. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5506159/

Appendices

Appendix 1: Initial exploration with simulated data

NOTE: When designing the experiment, we created simulated data, then went through data processing and analysis steps to ensure we had a solid plan prior to executing the pilot study.
### Read file with simulated data

```
#d <- fread("./Experimental_matrix_W241_Final_Project.csv") # simulation; contains nominal effects crim
# d <- fread("./Experimental_matrix_exploration_W241_Final_Project.csv")  # simulation; contains nomina
# d <- data.table(d)
# head(d)
#d
```

Spot check a few cases

```
# d[ , .(group_mean = mean(Misd1)), keyby = .(gender_misd1)]
#d[ , .(grp_mean = mean(Misd1))]
```

Here, based on simulated data we expect an offset of ~ 0.15 + 0.5*0.20 = 0.25 (sentence, in years), with males (gender_misd1==0) getting a larger sentence than the female treatment group. This result is expected.

```
# d[ , .(group_mean = mean(Felony1)), keyby = .(gender_felony1)]
```

Here, based on simulated data we expect an offset of ~ 1.5 + 0.5*2 = 2.5, with males (‘0’) getting a larger sentence than the femal treatment group. This result is expected.

```
# ATE_F1 = d[ , .('grp_avg' = mean(Felony1)), keyby = gender_felony1][, .('ATE_F1' = diff(grp_avg))]
# ATE_F1  # -2.394936
```

```
# NSIM = 1000
# results = rep(NA, NSIM)

# for(i in 1:NSIM){
#   results[i] = d[ , .('grp_avg' = mean(Felony1)), keyby = sample(gender_felony1)][ , diff(grp_avg)]
# }

# hist(results, main='Randomization Inference ATE Distribution', col='blue', breaks=30, xlim=c(-5, 5))
# abline(v=-2.394936, col='red', lwd=4)  #



# p_value_two_tailed <- mean( (results <= -2.394936) | (results >= 2.394936) )
# p_value_two_tailed
```

```
# hist(d[ , Misd1], breaks=30, col='blue')
```

```
# hist(d[ , Felony1], breaks=30, col='blue')
```

```
# model1 = d[ , lm(Misd1 ~ gender_misd1)]
# summary(model1)
# model1$vcovHC_ <- vcovHC(model1)
# coeftest(model1, vcov. = model1$vcovHC_)
```

Here, based on simulated data we expect an offset of 0.25 (sentence, in years), with males (gender_misd1==0) getting a larger sentence than the female treatment group. This result

is expected.

```
# model2 = d[ , lm(Felony1 ~ gender_felony1)]
# summary(model2)
# model2$vcovHC_ <- vcovHC(model2)
# coeftest(model2, vcov. = model2$vcovHC_)
```

Here, based on simulated data we expect an offset of 2.5, with males ('0') getting a larger sentence than the femal treatment group. This result is expected.

Check effect of Subject_Gender. Since we didn't build any offset into the simulation, we shouldn't see any significance with this factor.

```
# model3 = d[ , lm(Misd1 ~ gender_misd1 + Subject_Gender)]
# summary(model3)
# model3$vcovHC_ <- vcovHC(model3)
# coeftest(model3, vcov. = model3$vcovHC_)
```

This result is expected. In addition to the significant effect of the criminal's gender that was built into the simulated data, we now also have an additional effect of the Subject's (survey participant's) gender, where we made males have an additional positive sentence bias toward male criminals, but not female criminals. Since, for male subjects, we added 0.2 years to the misdemeanor sentence lengths for male criminals, but not female criminals, we see an average additional contribution of ~ 0.1 years, as expected.

```
# model4 = d[ , lm(Misd1 ~ gender_misd1 + Subject_Gender + Subject_Gender*gender_misd1)]
# summary(model4)
# model4$vcovHC_ <- vcovHC(model4)
# coeftest(model4, vcov. = model4$vcovHC_)
```

This captures the simulated situation. Overall, misdemeanor sentences are 0.15 years shorter for female 'criminals'. In addition, Male participants impose an additional 0.2 years for male 'criminals', but not for female criminals.

```
# model5 = d[ , lm(Felony1 ~ gender_felony1 + Subject_Gender + Subject_Gender*gender_felony1)]
# summary(model5)
# model5$vcovHC_ <- vcovHC(model5)
# coeftest(model5, vcov. = model5$vcovHC_)
```

Analogous to model4, except here for Felony1, this captures the simulated situation.

Try aggregating the 6 misdemeanor crimes. Can use the "sequence" variable as a check.

```
# d[ , .(sequence_sum = sum(Misd1, Misd2, Misd3, Misd4, Misd5, Misd6)), keyby = .(sequence)]
```

The average difference of (148-130)/100 = 0.18 is reasonable given the simulated data.

Group crimes together, keeping female treatment subgroups together, and using 'sequence' as the proxy for that treatment.

```
# d[ , Misd_A := (Misd1 + Misd3 + Misd5)]  # group crime subgroups with equal treatment assignments (fe
# d[ , Misd_B := (Misd2 + Misd4 + Misd6)]  # group crime subgroups with equal treatment assignments (fe
# d[ , Felony_A := (Felony1 + Felony3 + Felony5)]  # group crime subgroups with equal treatment assignm
# d[ , Felony_B := (Felony2 + Felony4 + Felony6)]  # group crime subgroups with equal treatment assignm
# head(d)
```

```
# model6 = d[ , lm(Misd_A ~ sequence + Subject_Gender + Subject_Gender*sequence)]
# summary(model6)
# model6$vcovHC_ <- vcovHC(model6)
# coeftest(model6, vcov. = model6$vcovHC_)
```

Results are about as expected (3 * 0.1) for the Subject_Gender ATE. (Everything scaled by a factor of ~ 3, or 1/3 relative to model4.) It is probably better to calculate weighted ATE rather than pooling the results together, even though standard error has scaled downward.

Appendix 2: Results of pilot study

```
# Now melting the data table so we see data on the question level
d_melt_pilot = melt(d_pilot, id.vars = c("RecordedDate", "ResponseId","QD1","QD2","QD3","QD4"),
                measure.vars = c("Q1A_2", "Q1B_2"
                                 , "Q2A_2","Q2B_2"
                                 , "Q3A_2","Q3B_2"
                                 , "Q4A_2","Q4B_2"
                                 , "Q5A_2","Q5B_2"
                                 , "Q6A_2","Q6B_2"
                                 , "Q7A_2","Q7B_2"
                                 , "Q8A_2","Q8B_2"
                                 , "Q9A_2","Q9B_2"
                                 , "Q10A_2","Q10B_2"
                                 , "Q11A_2","Q11B_2"
                                 , "Q12A_2","Q12B_2"))
names(d_melt_pilot) <- c("RecordedDate", "ResponseId","Field", "ZipCode","Gender","Party","Question","V
d_melt_pilot[,Version := ifelse(grepl('A',Question),1,0)]
d_melt_pilot[, male_criminal := ifelse(Question %in% c("Q3A_2", "Q10A_2","Q12A_2", "Q1A_2" ,"Q4A_2", "Q
 d_melt_pilot[, is_felony := ifelse(Question %in% c("Q1A_2","Q4A_2","Q5A_2","Q6A_2","Q9A_2","Q11A_2", "
head(d_melt_pilot)
```

```
##           RecordedDate        ResponseId
## 1: 2019-07-03 19:32:54 R_3oTujvr3EvlnfBf
## 2: 2019-07-04 19:54:33 R_BM9fDhyQB4GMBNf
## 3: 2019-07-04 20:14:04 R_vfffaRm7LRYbd8R
## 4: 2019-07-11 07:43:16 R_1nTli7A5MFIHABi
## 5: 2019-07-11 11:20:38 R_2DOlvcKaIc7BsQc
## 6: 2019-07-11 13:19:32 R_3mf7we9BZMNGFIg
##                                             Field ZipCode Gender
## 1:                                                 95035   Male
## 2:                                                 95120   Male
## 3:                                                 92603   Male
## 4: Other or Not Associated with legal profession   94539   Male
## 5: Other or Not Associated with legal profession   14604   Male
## 6: Other or Not Associated with legal profession   47906   Male
##                      Party Question Value Version male_criminal is_felony
## 1: Prefer not to disclose    Q1A_2             1             1         1
## 2:               Democrat    Q1A_2     8       1             1         1
## 3:               Democrat    Q1A_2     2       1             1         1
## 4:               Democrat    Q1A_2     4       1             1         1
## 5:               Democrat    Q1A_2             1             1         1
## 6:               Democrat    Q1A_2             1             1         1
```

```r
#d_melt_pilot
```

```r
summary(d_melt_pilot)
```

```
##  RecordedDate        ResponseId           Field          
##  Length:216         Length:216         Length:216        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                          
##                                                          
##                                                          
##                                                          
##    ZipCode             Gender              Party               Question  
##  Length:216         Length:216         Length:216         Q1A_2  :  9  
##  Class :character   Class :character   Class :character   Q1B_2  :  9  
##  Mode  :character   Mode  :character   Mode  :character   Q2A_2  :  9  
##                                                           Q2B_2  :  9  
##                                                           Q3A_2  :  9  
##                                                           Q3B_2  :  9  
##                                                           (Other):162  
##     Value              Version     male_criminal   is_felony  
##  Length:216         Min.   :0.0   Min.   :0.0   Min.   :0.0  
##  Class :character   1st Qu.:0.0   1st Qu.:0.0   1st Qu.:0.0  
##  Mode  :character   Median :0.5   Median :0.5   Median :0.5  
##                     Mean   :0.5   Mean   :0.5   Mean   :0.5  
##                     3rd Qu.:1.0   3rd Qu.:1.0   3rd Qu.:1.0  
##                     Max.   :1.0   Max.   :1.0   Max.   :1.0  
## 
```

```r
# Convert characters in 'Value' column to numeric
strTmp = c('Value')
d_melt_pilot[, (strTmp) := lapply(.SD, as.numeric), .SDcols = strTmp]
#d_melt_pilot
```

```r
# Omit rows where 'Value' is missing, since these were not typically assigned to the subjects.
dp = na.omit(d_melt_pilot, cols='Value')
#dp
```

Felony Questions: Q1, Q4, Q5, Q6, Q9, Q11

Misdemeanor Questions: 2, 3, 7, 8, 10, 12

```r
summary(dp)
```

```
##  RecordedDate        ResponseId           Field          
##  Length:108         Length:108         Length:108        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                          
##                                                          
##                                                          
##                                                          
##    ZipCode             Gender              Party               Question 
##  Length:108         Length:108         Length:108         Q1A_2  : 5  
##  Class :character   Class :character   Class :character   Q2A_2  : 5
##  Mode  :character   Mode  :character   Mode  :character   Q3A_2  : 5
##                                                            Q4A_2  : 5
##                                                            Q5A_2  : 5
##                                                            Q6A_2  : 5
##                                                            (Other):78
##      Value           Version       male_criminal   is_felony
##  Min.   : 0.000   Min.   :0.0000   Min.   :0.0   Min.   :0.0
##  1st Qu.: 0.500   1st Qu.:0.0000   1st Qu.:0.0   1st Qu.:0.0
##  Median : 2.000   Median :1.0000   Median :0.5   Median :0.5
##  Mean   : 4.202   Mean   :0.5556   Mean   :0.5   Mean   :0.5
##  3rd Qu.: 6.000   3rd Qu.:1.0000   3rd Qu.:1.0   3rd Qu.:1.0
##  Max.   :20.000   Max.   :1.0000   Max.   :1.0   Max.   :1.0
##
```

```
dp[ , .N, keyby=Party]
```

```
##                      Party  N
## 1:               Democrat 96
## 2: Prefer not to disclose 12
```

```
dp[ , .N, keyby=Gender]
```

```
##    Gender  N
## 1: Female 36
## 2:   Male 72
```

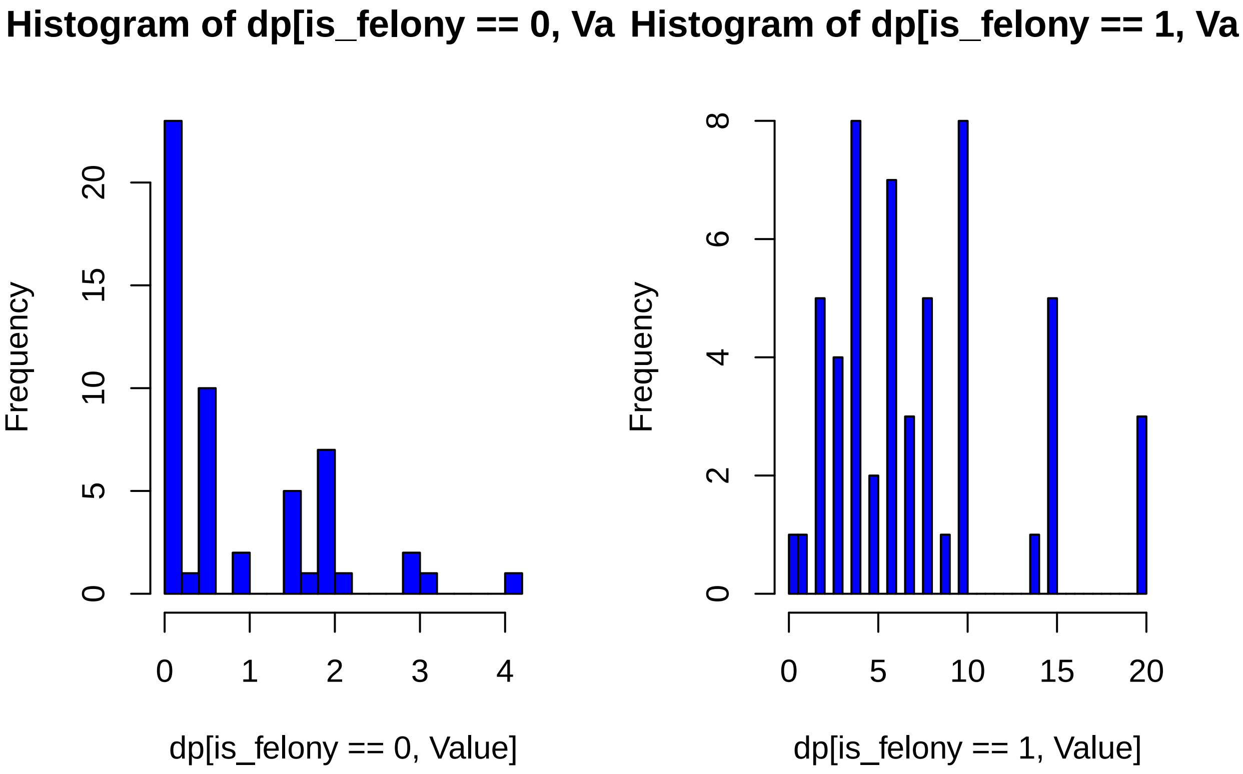
```
dp[ , .N, keyby=Field]
```

```
##                                             Field  N
## 1:                                                36
## 2: Other or Not Associated with legal profession 72
```

```
dp[ , .N, keyby=Version]
```

```
##    Version  N
## 1:       0 48
## 2:       1 60
```

```
# Plot histograms of sentences ('Values') for all misdemeanors and all felonies
par(mfrow=c(1,2))
hist(dp[is_felony==0, Value], breaks=30, col='blue')
hist(dp[is_felony==1, Value], breaks=30, col='blue')
```



Aggregate ALL felonies:

```
# For ALL felony crimes, compare means for male vs. female criminal
dp[is_felony==1, .(group_mean = mean(Value), group_sd = sd(Value)), keyby = .(male_criminal)]
```

```
##    male_criminal group_mean group_sd
## 1:             0   7.925926 5.060601
## 2:             1   7.111111 4.917264
```

NOTE: We see that the overall difference in the means when comparing male vs. female criminals for ALL felonies is very small relative to the variation.

```
# Model ALL Felonies
model0 = dp[Question=='Q1A_2' | Question=='Q1B_2', lm(Value ~ male_criminal)]
summary(model0)
```

```
##
## Call:
## lm(formula = Value ~ male_criminal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -1.80  -1.75   0.20   0.25   4.20
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.750      1.027   2.677   0.0317 *
## male_criminal    1.050      1.378   0.762   0.4710
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.055 on 7 degrees of freedom
## Multiple R-squared:  0.07656,    Adjusted R-squared:  -0.05536
## F-statistic: 0.5804 on 1 and 7 DF,  p-value: 0.471
```

```
model0$vcovHC_ <- vcovHC(model0)
coeftest(model0, vcov. = model0$vcovHC_)
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.75000    0.72648  3.7854 0.006844 **
## male_criminal   1.05000    1.44145  0.7284 0.489994
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

***Not stat sig. for male_criminal across ALL felonies.*

Look at just the first felony:

```
# For the first felony crime, compare means for male vs. female criminal
dp[Question=='Q1A_2' | Question=='Q1B_2', .(group_mean = mean(Value), group_sd = sd(Value)), keyby = .(
```

```
##    male_criminal group_mean group_sd
## 1:             0       2.75 1.258306
## 2:             1       3.80 2.489980
```

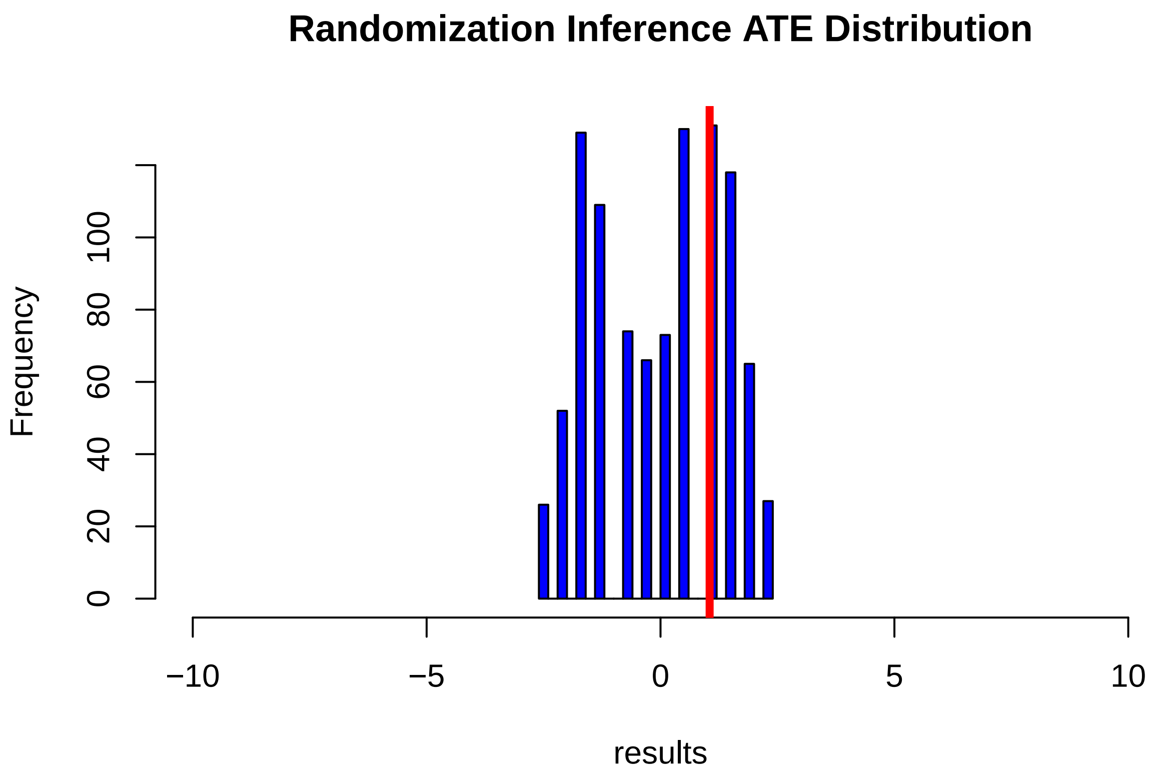
```
ATE_F1 = dp[Question=='Q1A_2' | Question=='Q1B_2', .(group_mean = mean(Value)), keyby = .(male_criminal
ATE_F1  # 1.05
```

```
##    ATE_F1
## 1:   1.05
```

```
NSIM = 1000
results = rep(NA, NSIM)

for(i in 1:NSIM){
  results[i] = dp[Question=='Q1A_2' | Question=='Q1B_2', .(group_mean = mean(Value)), keyby = sample(mal
}

hist(results, main='Randomization Inference ATE Distribution', col='blue', breaks=30, xlim=c(-10, 10))
abline(v=1.05, col='red', lwd=4)   #
```



```
p_value_two_tailed <- mean( (results <= -1.05) | (results >= 1.05) )
p_value_two_tailed
```

```
## [1] 0.526
```

```
# Model the first felony
model1 = dp[Question=='Q1A_2' | Question=='Q1B_2', lm(Value ~ male_criminal)]
summary(model1)
```

```
##
## Call:
## lm(formula = Value ~ male_criminal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -1.80  -1.75   0.20   0.25   4.20
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.750      1.027   2.677   0.0317 *
## male_criminal     1.050      1.378   0.762   0.4710
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.055 on 7 degrees of freedom
## Multiple R-squared:  0.07656,    Adjusted R-squared:  -0.05536
## F-statistic: 0.5804 on 1 and 7 DF,  p-value: 0.471
```

```
model1$vcovHC_ <- vcovHC(model1)
coeftest(model1, vcov. = model1$vcovHC_)
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.75000    0.72648  3.7854 0.006844 **
## male_criminal   1.05000    1.44145  0.7284 0.489994
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Model the first felony
model2 = dp[Question=='Q1A_2' | Question=='Q1B_2', lm(Value ~ male_criminal + Gender)]
summary(model2)
```

```
##
## Call:
## lm(formula = Value ~ male_criminal + Gender)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6923 -0.4615 -0.3077  0.5385  3.3077
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.077      1.335   0.807    0.450
## male_criminal    1.385      1.236   1.121    0.305
## GenderMale       2.231      1.302   1.713    0.138
##
## Residual standard error: 1.819 on 6 degrees of freedom
## Multiple R-squared:  0.3798, Adjusted R-squared:  0.1731
## F-statistic: 1.837 on 2 and 6 DF,  p-value: 0.2385
```

```
model2$vcovHC_ <- vcovHC(model2)
coeftest(model2, vcov. = model2$vcovHC_)
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.07692    0.98322  1.0953   0.3154
## male_criminal   1.38462    1.42545  0.9713   0.3689
## GenderMale      2.23077    1.29427  1.7236   0.1356
```

```
# Model the first felony
model3 = dp[Question=='Q1A_2' | Question=='Q1B_2', lm(Value ~ male_criminal + Gender + male_criminal*Ge
summary(model3)
```

```
##
## Call:
## lm(formula = Value ~ male_criminal + Gender + male_criminal *
##     Gender)
##
## Residuals:
##          1          2          3          4          5          6
##  3.333e+00 -2.667e+00 -6.667e-01 -5.000e-01  5.000e-01  6.667e-01
##          7          8          9
## -3.333e-01 -3.333e-01 -6.661e-16
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.0000     1.9916   0.502    0.637
## male_criminal              1.5000     2.4393   0.615    0.565
## GenderMale                 2.3333     2.2998   1.015    0.357
## male_criminal:GenderMale  -0.1667     2.9316  -0.057    0.957
##
## Residual standard error: 1.992 on 5 degrees of freedom
## Multiple R-squared:  0.3802, Adjusted R-squared:  0.008333
## F-statistic: 1.022 on 3 and 5 DF,  p-value: 0.4569
```

```
model3$vcovHC_ <- vcovHC(model3)
coeftest(model3, vcov. = model3$vcovHC_)
```

```
##
## t test of coefficients:
##
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.00000         NA      NA       NA
## male_criminal             1.50000         NA      NA       NA
## GenderMale                2.33333         NA      NA       NA
## male_criminal:GenderMale -0.16667         NA      NA       NA
```

Quick check of the other felonies: Q4,5,6,9,11

```
# For the first felony crime, compare means for male vs. female criminal
dp[Question=='Q4A_2' | Question=='Q4B_2', .(group_mean = mean(Value), group_sd = sd(Value)), keyby = .(
```

```
##    male_criminal group_mean group_sd
## 1:             0          7 6.164414
## 2:             1          6 3.741657
```

```
dp[Question=='Q5A_2' | Question=='Q5B_2', .(group_mean = mean(Value), group_sd = sd(Value)), keyby = .(
```

```
##    male_criminal group_mean group_sd
## 1:             0        9.8 5.263079
## 2:             1        8.5 7.895146
```

```
dp[Question=='Q6A_2' | Question=='Q6B_2', .(group_mean = mean(Value), group_sd = sd(Value)), keyby = .(
```

```
##    male_criminal group_mean group_sd
## 1:             0        7.0 2.581989
## 2:             1        9.6 7.503333
```

```
dp[Question=='Q9A_2' | Question=='Q9B_2', .(group_mean = mean(Value), group_sd = sd(Value)), keyby = .(
```

```
##    male_criminal group_mean group_sd
## 1:             0      10.60 3.847077
## 2:             1       8.25 1.258306
```

```
dp[Question=='Q11A_2' | Question=='Q11B_2', .(group_mean = mean(Value), group_sd = sd(Value)), keyby =
```

```
##    male_criminal group_mean group_sd
## 1:             0          9  6.78233
## 2:             1          7  2.94392
```

Felony Q9 appears to have the best chance at a significant difference. Run a test. (Note: here we observe longer average sentence for *female* criminals; out of 6 felonies, we observed larger mean sentences for females in 4 of the cases, although they are not statistically significant differences.)

```
# Model the Felony in Q9
model4 = dp[Question=='Q9A_2' | Question=='Q9B_2', lm(Value ~ male_criminal)]
summary(model4)
```

```
##
## Call:
## lm(formula = Value ~ male_criminal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4.60  -1.25  -0.25   1.75   4.40
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     10.600      1.352   7.842 0.000104 ***
## male_criminal   -2.350      2.028  -1.159 0.284454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.023 on 7 degrees of freedom
## Multiple R-squared:  0.161,  Adjusted R-squared:  0.04115
## F-statistic: 1.343 on 1 and 7 DF,  p-value: 0.2845
```

```
model4$vcovHC_ <- vcovHC(model4)
coeftest(model4, vcov. = model4$vcovHC_)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    10.6000     1.9235  5.5107 0.0008963 ***
## male_criminal  -2.3500     2.0562 -1.1429 0.2906549
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Not stat sig.

```
# Model the Felony in Q9
model5 = dp[Question=='Q9A_2' | Question=='Q9B_2', lm(Value ~ male_criminal + Gender + male_criminal*Ger
summary(model5)
```

```
##
## Call:
## lm(formula = Value ~ male_criminal + Gender + male_criminal *
##     Gender)
##
## Residuals:
##          1          2          3          4          5          6
##  4.000e+00 -4.000e+00  2.220e-16  3.500e+00 -3.500e+00 -6.667e-01
##          7          8          9
##  3.333e-01  3.333e-01 -3.331e-16
##
```

```
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                11.5000     2.3910   4.810  0.00484 **
## male_criminal              -1.5000     4.1413  -0.362  0.73200
## GenderMale                 -1.5000     3.0867  -0.486  0.64755
## male_criminal:GenderMale  -0.8333     4.9772  -0.167  0.87359
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.381 on 5 degrees of freedom
## Multiple R-squared:   0.25,  Adjusted R-squared:   -0.2
## F-statistic: 0.5556 on 3 and 5 DF,  p-value: 0.6667
```

```
model5$vcovHC_ <- vcovHC(model5)
coeftest(model5, vcov. = model5$vcovHC_)
```

```
##
## t test of coefficients:
##
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              11.50000         NA      NA       NA
## male_criminal            -1.50000         NA      NA       NA
## GenderMale               -1.50000         NA      NA       NA
## male_criminal:GenderMale -0.83333         NA      NA       NA
```

Look at the misdemeanors (Questions 2, 3, 7, 8, 10, 12):

```
# For the first misdemeanor crime, compare means for male vs. female criminal
dp[Question=='Q2A_2' | Question=='Q2B_2', .(group_mean = mean(Value), group_sd = sd(Value)), keyby = .(
```

```
##    male_criminal group_mean  group_sd
## 1:             0      1.200 1.1510864
## 2:             1      0.675 0.9069179
```

```
dp[Question=='Q3A_2' | Question=='Q3B_2', .(group_mean = mean(Value), group_sd = sd(Value)), keyby = .(
```

```
##    male_criminal group_mean  group_sd
## 1:             0      0.525 0.9844626
## 2:             1      0.340 0.2701851
```

```
dp[Question=='Q7A_2' | Question=='Q7B_2', .(group_mean = mean(Value), group_sd = sd(Value)), keyby = .(
```

```
##    male_criminal group_mean  group_sd
## 1:             0      1.300 0.7582875
## 2:             1      1.175 1.0242884
```

```
dp[Question=='Q8A_2' | Question=='Q8B_2', .(group_mean = mean(Value), group_sd = sd(Value)), keyby = .(
```

```
##    male_criminal group_mean group_sd
## 1:             0       1.82 1.279453
## 2:             1       1.85 1.731088
```

```
dp[Question=='Q10A_2' | Question=='Q10B_2', .(group_mean = mean(Value), group_sd = sd(Value)), keyby =
```

```
##    male_criminal group_mean  group_sd
## 1:             0      0.525 0.7847505
## 2:             1      0.200 0.1732051
```

```
dp[Question=='Q12A_2' | Question=='Q12B_2', .(group_mean = mean(Value), group_sd = sd(Value)), keyby =
```

```
##    male_criminal group_mean  group_sd
## 1:             0      0.625 0.8770215
## 2:             1      0.400 0.6244998
```

For the misdemeanors, we observe longer average sentences for females in 5 out of the 6 cases.

Pilot Study Findings

Firstly from the pilot study, we realized that we were only asking one version of the survey at first without randomly assigning respondents to the second half of the study. We also realized the questions on whether individuals are law students or not was not asked. Additionally, we noticed that the surveys did not have an equal balance of genders among the misdemanors and felonies (even though there were an equal number of male and female respondents overall). Moreover, we received feedback on the survey from our respondents that the misdemeanors should be asked together and the felonies should should be asked together. This is because the accepted ranges differs from 0 to 5 to 0 to 20 between the two groups, and constantly flunctuating scaled confused our participants.

We also added language clarifying that the study results were anonymous and refined the language used to invite our respondents to our study to make it easier to understand.

In terms of analysis, we see that for the small sample size of this pilot test, the differences in sentence lengths (‘Value’) between male and female criminals are not significant. We observe that although not statistically significant, the mean sentence length is actually longer for female criminals in 4 of the 6 felonies and 5 of the 6 misdemeanors. The felony crime having the largest apparent difference relative to the variation is from Question 9, which shows a difference (ETA) of ~ -2.35 years, but the HC standard error is 2.06 years, so the effect is not statistically significant. For this crime, we would need to reduce the standard error roughly in half to render the effect statistically significant; this implies that we would need our main study sample size to be at least 4x as large as the pilot sample.