Statistics 141 Project Phase 3

Group 1 — Eleazar Birke, Sophia Miller, Sophie Miller, John Stoehr, Shardul Vijay

Wednesday, December 8th, 2021

~~~
---
title: "Final Project Phase"
author: "Group 1"
date: "12/6/2021"
header-includes: |
  \usepackage{fancyhdr}
  \pagestyle{fancy}
  \fancyhead[CO,C]{Homework 9 - MATH 141}
  \fancyfoot[CO,C]{}
  \fancyfoot[C]{\thepage}
  \usepackage{float}
output:
  bookdown::pdf_document2:
    fig_caption: yes
    toc: no
    number_section: no
urlcolor: red
---
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```


```{r echo=FALSE, message=FALSE}
# load libraries
library(tidyverse)
library(openintro)
library(dplyr)
library(infer)
library(kableExtra)
library(janitor)
library(gghighlight)
```
~~~

Introduction:

Our data comes from Rossi et al.'s Criminal Recidivism Data (cited below), which can be downloaded in the form of a csv file for use in R. This data was collected in 1980, and summarizes information tracking 432 individuals released from Maryland state prisons in the

1970s for one year after release. This study was run as an experiment, where half of the individuals were provided with financial aid, and the other half did not receive any aid.

The study explicitly lays out each of the 62 variables' contexts, but neglects terms like categorical & numerical, and their respective subcategories. As we continued to develop and investigate our dataset, we converted the below variables with an X into discrete numeric variables, assigning 0's and 1's where appropriate.

Classification of each variable:

week: numerical, discrete

arrest: numerical, discrete

fin: categorical, nominal (X - binary)

age: numerical, discrete

race: categorical, nominal (X - binary)

wexp: categorical, nominal (X - binary)

mar: categorical, nominal (X - binary)

paro: categorical, nominal (X - binary)

prio: numerical, discrete

educ: numerical, discrete (discrete values represent ranges in education, i.e 2 = 6th grade or less, maybe we could tinker with this)

emp1-emp52: categorical, nominal(X - binary)

Cited Data Set:

Rossi, P.H., R.A. Berk, and K.J. Lenihan (1980). Money, Work, and Crime: Some Experimental Results. New York: Academic Press.

Research Question and Hypotheses

When examining this data set, we looked at some sources on recidivism to further explain what we were observing in the dataset, in order to perform data exploration more accurately. In order to examine the relationships between potential reasons for recidivism, we further looked at reports analyzing the data set we chose, alongside coming up with a main research question to dictate our formal line of research, written below.

Is there a relationship between employment and recidivism?

In looking at the relationship between employment and recidivism, we began to also look at the variable of financial aid, and examine the variable of employment through two lenses — through length of employment and the speed at which employment was obtained. In looking at these variables, we created a series of hypotheses that further guided our data exploration. These are listed below:

H0: Receipt of financial aid has no effect on the likelihood of recidivism.

$H0 = p_{arrest w/financial aid} - p_{arrest w/o finaid} = 0$

HA: Receipt of financial aid has an effect on the likelihood of recidivism.

$HA = p_{arrest w/financial aid} - p_{arrest w/o finaid)} \neq 0

H0: Individuals that were employed for more than half the duration of the study (26 weeks) were no more or less likely to be re-arrested than those who were not.

$H0 = p_{proportion arrested w/ half employment} - p{arrest w/o half employment} = 0$

HA: Individuals that were employed for more than half the duration of the study (26 weeks) were less likely to be re-arrested than those who were not.

$HA = p_{proportion arrested w/ half employment} - p{arrest w/o half employment} \neq 0$

In our hypotheses, we examined the length of employment, but shifted the precedence on the speed at which employment was garnered to the financial aid variable, as it might explain why a person may actively seek employment more readily, or the opposite. One thing that is objectively absent from our analysis is a discussion of the variable involving race, for the following reasons.

The breakdown of the races in the dataset delineate variables of “white” and “nonwhite” options which prohibit nuanced examination of the individualized races and their respective results. Therefore, we did not use the variable of race in our study, as racial prejudices already skew the population of incarcerated people. Each state’s racial distribution of incarceration population differs, and none of them correctly reflect the overall population of each state. Due to systematic discrimination and the resulting socio-economic conditions people of color are subject to, the distribution of race within incarcerated populations are greatly biased. The study published in 1980 explains “there may have been many reasons for the overrepresentation of blacks and Hispanics relative to whites in the TARP groups. The main reason, however, as we will show, lay in the socioeconomic position occupied by these groups within their states.”

(Rossi, 124) In order to consider race as a valuable variable in determining employment rate post release, we would first have to isolate and investigate US race relations and the impact of systematic racism within both the nation and individual states.

According to table 7.3, the distribution of the white and Black population within federal prisons in 1973 was equal, 48% and 48% respectively. In contrast, the distribution of race within Georgia's prison population is skewed, with 61% of incarcerations categorized as Black. The distribution of race within Texas prisons differs from both Georgia's and the Country. The population within Texas prisons are divided into white, Black and Hispanic groups, with the percentages respectively following as: 38%, 45%, and 15%.

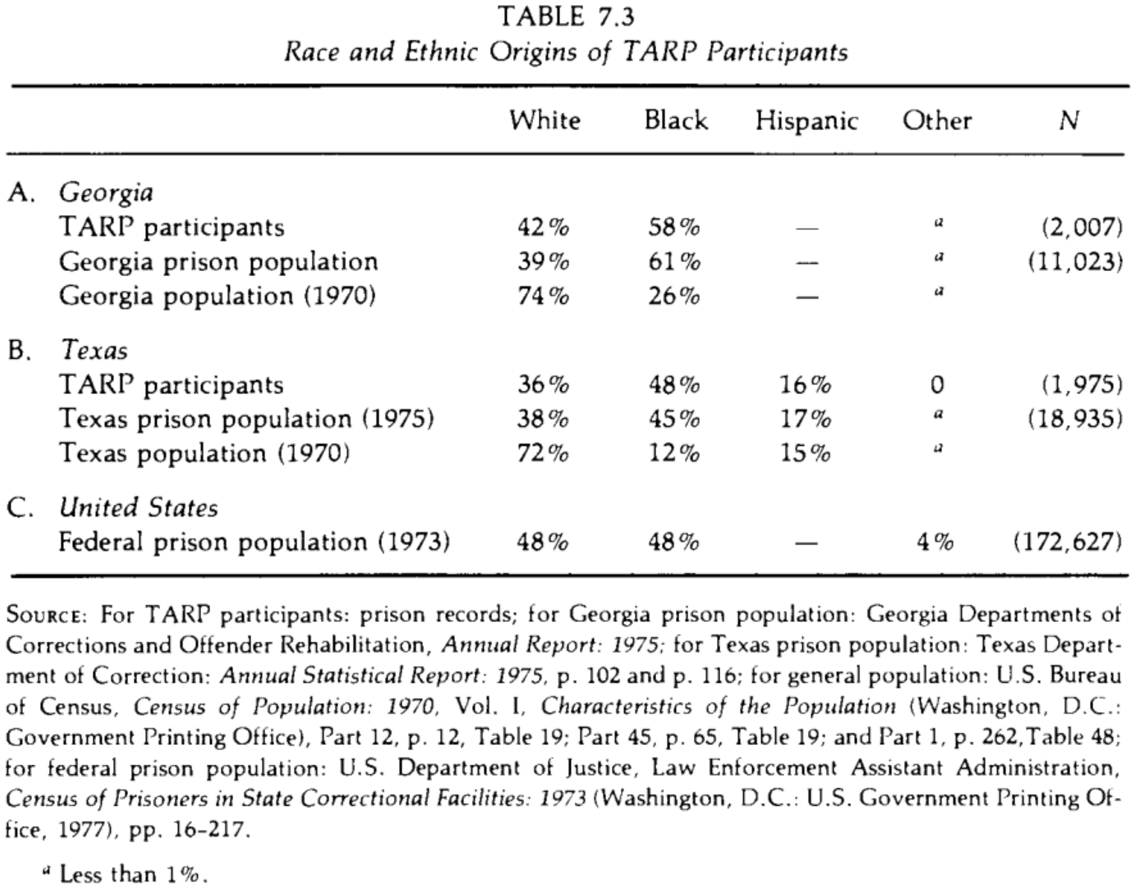


TABLE 7.3
Race and Ethnic Origins of TARP Participants

	White	Black	Hispanic	Other	*N*
A. *Georgia*					
TARP participants	42%	58%	—	[a]	(2,007)
Georgia prison population	39%	61%	—	[a]	(11,023)
Georgia population (1970)	74%	26%	—	[a]	
B. *Texas*					
TARP participants	36%	48%	16%	0	(1,975)
Texas prison population (1975)	38%	45%	17%	[a]	(18,935)
Texas population (1970)	72%	12%	15%	[a]	
C. *United States*					
Federal prison population (1973)	48%	48%	—	4%	(172,627)

SOURCE: For TARP participants: prison records; for Georgia prison population: Georgia Departments of Corrections and Offender Rehabilitation, *Annual Report: 1975;* for Texas prison population: Texas Department of Correction: *Annual Statistical Report: 1975,* p. 102 and p. 116; for general population: U.S. Bureau of Census, *Census of Population: 1970,* Vol. I, *Characteristics of the Population* (Washington, D.C.: Government Printing Office), Part 12, p. 12, Table 19; Part 45, p. 65, Table 19; and Part 1, p. 262,Table 48; for federal prison population: U.S. Department of Justice, Law Enforcement Assistant Administration, *Census of Prisoners in State Correctional Facilities: 1973* (Washington, D.C.: U.S. Government Printing Office, 1977), pp. 16–217.

[a] Less than 1%.

Data Exploration: ***This section should include your exploration of your data that leads to your question and hypotheses. Include details on how you wrangle your data or if made additional computations.***

In order to showcase our hypothesis & results, we wanted to explore the effect of employment on recidivism within the Rossi 1970 dataset. Sofia generated a plot showing how employment appears to affect recidivism. By looking at cumulative weeks of employment, the plot below depicts the relationship between time spent in formal employment, formally unemployed, and these two groups' composite individual propensity for rearrest. This visual, Figure 1, as well as Table 1 accompanying it, provided us with a clear visual to present as we navigate the Rossi dataset.

Given our observations regarding steady employment's efficacy for rearrest prevention, we wanted to explore when these individuals found a job after release. We also considered that these employment dates may coincide with financial aid presence or lack thereof. We found individuals' employment date follow a fairly normal distribution immediately upon release, regardless of financial aid.

But there was a spike in employment, as the study was coming to a close, skewing the histogram data to the right, perhaps indicative of individuals who received financial aid were planning to gain employment upon their last week of receipt. We observed a mild effect regarding the proportion of those in receipt of aid and those who were not in receipt of financial aid as the study and aid came to a close.

The last exploration we ventured into stemmed from our growing assumption that financial variables were the ideal predictor variables in the dataset for whether or not individuals return to jail. We used a permutations test for financial aid recipients and their counterparts (no financial aid recipients) to explore the overall effect the financial aid had on recidivism.

The exploration we conducted align with the findings in the Rossi dataset's publication, "TARP demonstrated that the provision of limited amounts of financial aid to released prisoners in the form of minimum unemployment benefit payments for periods of between 3 and 6 months can decrease the arrests experienced by the ex-felons in the year following release by 25% to 50%." (Rossi, 7). Such that financial variables, though they come with nuance & are not the sole predictive data metric, are of the few available to ascertain predictions regarding recidivism.

Code for initial

~~~
```{r SophiaCode, echo = TRUE}
### pivot to create columns indicating week of employment and whether or not the individual
was employed
recidivism_col <- rownames_to_column(recidivism,var = "individual")
###sum weeks employed for each individual and make that a new column
emp_long <- pivot_longer(recidivism_col,emp1:emp52,names_to = "emp_week",values_to =
"emp_stat")
emp_long$emp_num <- ifelse(emp_long$emp_stat=="yes",1,0)

emp_c <- emp_long %>%
  add_count(individual,wt = emp_num) %>%
  rename(total_weeks_emp = n)
emp_c2 <- emp_c %>%
  select(-emp_num)
~~~

~~~
emp_fin <- pivot_wider(emp_c2,names_from ="emp_week",values_from ="emp_stat") %>%
  mutate(emp_half = case_when((total_weeks_emp <= 26) ~ "No",(total_weeks_emp > 26) ~
"Yes"))
emp_fin$arrested <- ifelse(emp_fin$arrest == 1,"arrested","not arrested")
view(emp_fin)

###histogram of cumulative weeks employed for all individuals which separates out groups who
were arrested during the period of the study
ggplot(data = emp_fin, aes(x = total_weeks_emp)) +  geom_histogram(bins=40) + labs(x =
"weeks of employment maintained by an individual", y = "frequency", title = "Cumulative
Weeks Employed") + facet_wrap(~arrested)

###For people that were arrested, what proportion were employed for at least half the duration of
the study (26 weeks)?
emp_calc <- emp_fin %>%
  count(arrest,emp_half) %>%
  group_by(emp_half) %>%
  mutate(prop = n/sum(n))
view(emp_calc)
```
~~~

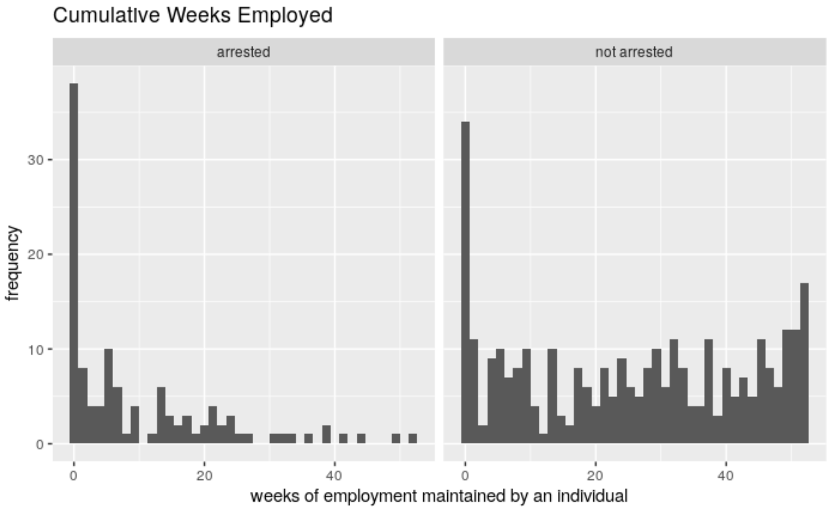
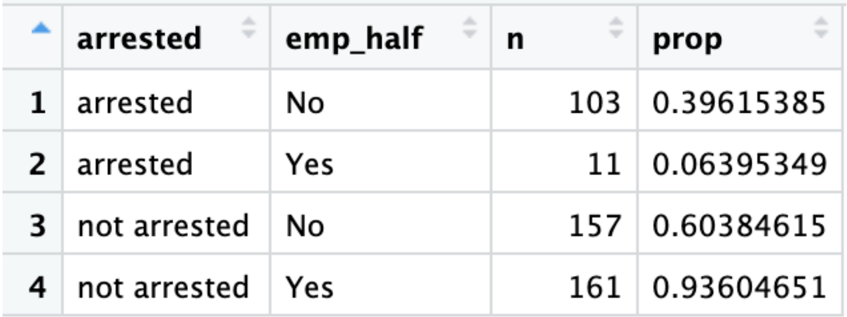


Figure 1. Histogram showing the frequency of individuals who maintained employment for a certain number of weeks during the duration of the study, separated by individuals who were or not arrested at any point.



	arrested	emp_half	n	prop
1	arrested	No	103	0.39615385
2	arrested	Yes	11	0.06395349
3	not arrested	No	157	0.60384615
4	not arrested	Yes	161	0.93604651

Table 1. Proportion populations of arrested group & unarrested group with respect to their employment duration exceeding half the study duration (26 weeks out of 52 weeks total)

The proportion of individuals employed for 26 weeks steadily (emp_half : Yes : total = 171 individuals) & remained unarrested was 94%, the remaining 6%, though they managed to maintain employment, they were rearrested. Those who could not maintain employment for at least half the study period (emp_half : No : total = 260 individuals) yielded a small difference in proportion, with 60% remaining unarrested, and 40% experiencing a rearrest.

Code for original employment date for prisoners released dataframe.

~~~
```{r time_until_employment, echo = TRUE}
data_dim <- dim(recidivism) # the dimension of the dataframe (rows,columns)
emp_indices <- seq(12,63,1) # sequence of numbers from 12 to 63 - indices of the columns of interest (e.x. emp4) in the data frame
week_emp <- c()
for(i in 1:data_dim[1]){
counter <- 0
for(j in emp_indices){
if(is.na(recidivism[[i,j]])){
counter <- NA
break
}
if(recidivism[[i,j]] == "yes"){
break
}
counter <- counter + 1
}
week_emp <- c(week_emp,counter)
}
we <- data.frame(week_emp)
we
week_emp_h <- ggplot(data=we, aes(x=week_emp)) + geom_histogram(bins=30) + labs(x = "first week of employment for an individual", y = "frequency", title = "First Week Employed")
week_emp_h
~~~

~~~
```
Code taking above build dataframe and generating a new dataframe to depict concurrent effects of financial aid & initial employment week proportional to each other.
```{r time_until_employment, echo = TRUE}
data_dim <- dim(recidivism) # the dimension of the dataframe (rows,columns)
fa_indices <- seq(4,1) # sequence of numbers from 12 to 63 - indices of the columns of interest (e.x. emp4) in the data frame
fa <- c()
for(i in 1:data_dim[1]){
counter <- 0
for(j in fa_indices){
if(is.na(recidivism[[i,j]])){
counter <- NA
break
}
if(recidivism[[i,j]] == "yes"){
break
}
counter <- counter + 1
}
fa <- c(fa,counter)
}
we <- data.frame(fa,werd)
we <- we %>% mutate(fa_char = as.character(fa)) # mutate so we may plot the fa variable as a categorical variable for our aes(fill)
we
fa_plot <- ggplot(data=we, aes(x=werd)) + geom_histogram(bins=30, color="orange", aes(fill=fa_char)) + labs(x = "first week of employment for an individual", y = "frequency", title = "First Week Employed")
###
fa_plot


```
~~~

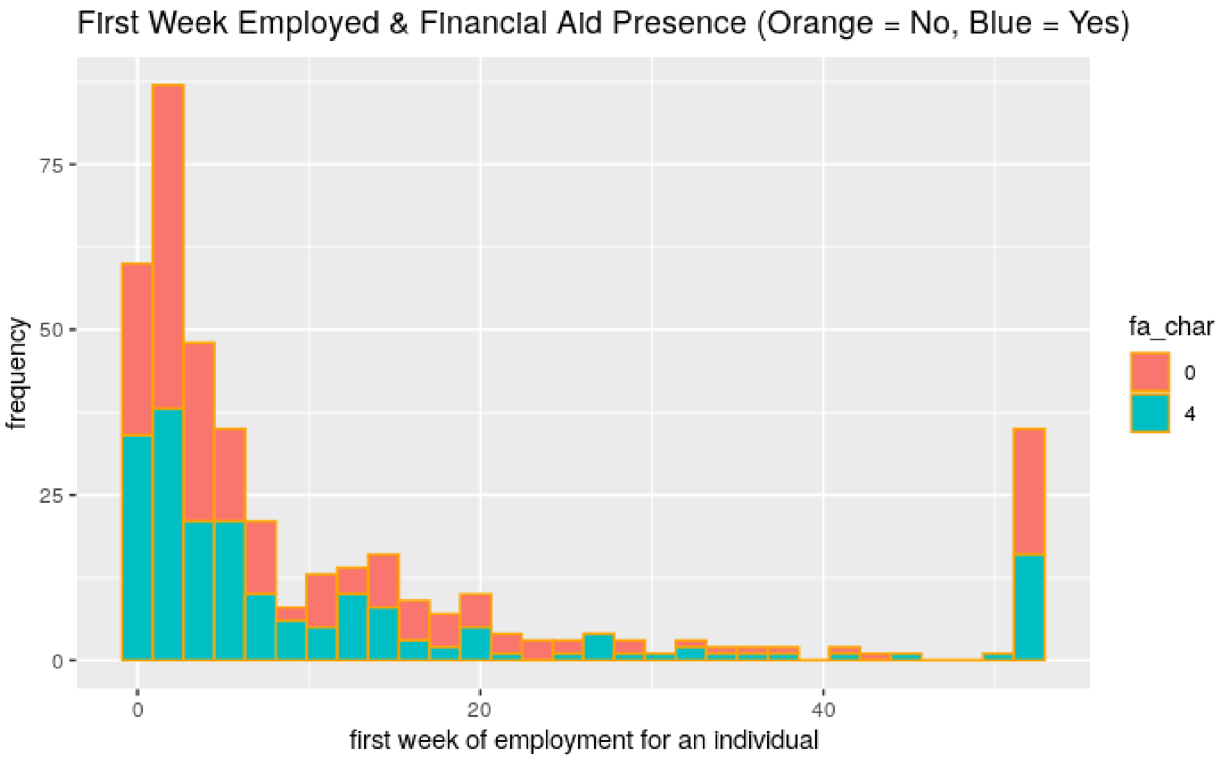


Figure 2. Graphical representation of individuals specific employment week post confinement release. The overlay of colors shown where 4 represents the proportion of individuals finding employment at that given week having received financial aid, and 0 represents the proportion of those who found employment at that week without financial aid. The sudden peak in employment at or near 52 weeks as compared to most individuals finding employment relatively soon after their release date depicts a large skew to the right, and illustrates that as financial aid comes to an end we observe a quantifiable effect on individual employment status.

Methods: ***This section should include the processes you took on performing your statistical method.***

~~~
```{r StatAnalysis echo=TRUE}
###basic bootstrapping analysis
yesno <- recidivism %>% select(arrest, fin)
set.seed(334422)
k <- 1000 # number of trials
n <- 432 # sample size
bs_yesno <- rep_sample_n(yesno,n,replace=TRUE,reps=k)
bs_yesno_prop <- bs_yesno %>%
group_by(arrest)
print(bs_yesno_prop)
~~~

```
bs_yesno_prop %>% count(fin == "yes")

###permutation test

#determining diff in prop for our sample
recid_prop <- emp_fin %>%
  count(arrested,fin) %>%
  group_by(fin) %>%
  mutate(prop = n / sum(n))
recid_prop
p_yes <- recid_prop %>%
  filter(fin == "yes", arrested == "not arrested") %>%
  pull(prop)
p_no <- recid_prop %>%
  filter(fin == "no", arrested == "not arrested") %>%
  pull(prop)
p_diff <- p_yes-p_no
p_diff

emp_fin %>%
  specify(arrested ~ fin, success = "not arrested") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 5, type = "permute")

##1000 shuffles
set.seed(334422) # set seed for replicability
shuffles <- emp_fin %>%
  specify(arrested ~ fin, success = "not arrested") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in props", order = c("yes", "no"))

#make a plot to show our shuffles, with the shaded dots representing values greater than the
difference in proportions found in our data:
shuffles %>%
  ggplot(aes(x = stat)) +
  geom_dotplot(binwidth = 0.0015) +
  gghighlight(stat >= 0.083) +
  theme(
    axis.ticks.y = element_blank(),
    axis.text.y = element_blank()
  ) +
  labs(
    x = "Differences in recidivism rates (with financial aid - no financial aid) across 1000 shuffles",
    y = NULL
  )
```

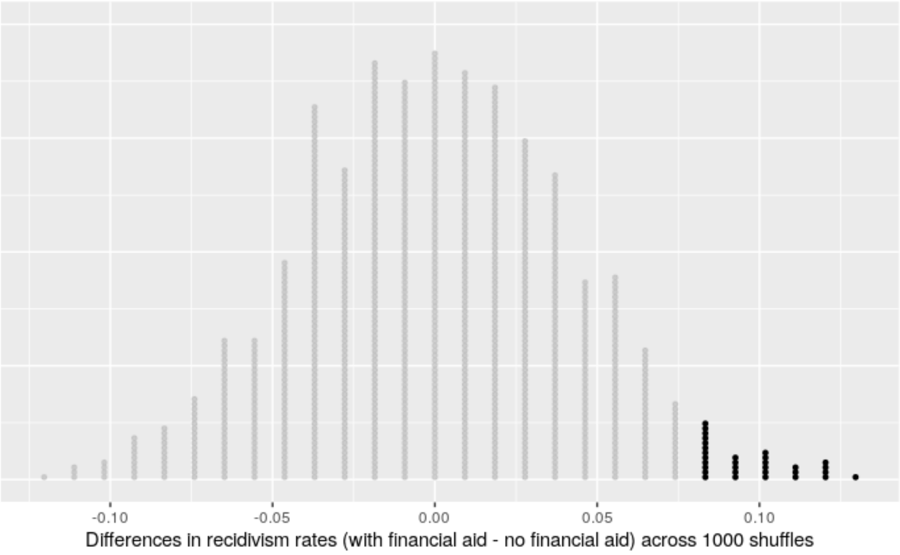


Figure 3. 1000 permutations, with black lines representing values greater than the difference in proportions found in our data (observed difference beginning around 0.08).

The variables financial aid and recidivism are *not* independent. The difference in recidivism rates of 0.083 was not due to natural variability, and financial aid recipients are less likely to return to jail compared to their non-financial aid receiving counterparts..

Alternative Hypothesis = Pfa - Pnfa > 0

In the 1000 shuffles example, we determined that there was only a ≈ 3.1% probability of obtaining a sample where ≥ 8.3% more non financial aid recipients than financial aid recipients return to jail under our null hypothesis.

After some rigorous computations and formal studies, we can conclude that the data provide strong evidence of financial aid's impact on recidivism. In this case, we **reject the null hypothesis in favor of the alternative.**

Results and Discussions: ***This section should include the results of your statistical analysis and discuss them in context of your research question. You can still include figures and tables here.***

The results of our statistical analysis present evidence for financial aid & employment being primary drivers for reducing recidivism. However, there is nuance within the construct of the prison systems, and though our results depict some promise, there are more variables to account for that determine recidivism. As observed in the Rossi publication, “The clarity of TARP results was somewhat obscured by the presence of unanticipated and undesirable side effects.” (Rossi 7) Be it disincentivizing job seeking behavior in released individuals by only providing financial aid should they remain unemployed(Rossi 7), or the postulate of the state of Georgia discriminating against skin color for a predisposition for recidivism, to that same states potential to discriminate based on skin color for employment, as well as the individual being newly released from jail (Rossi 8).

With many barriers to readjustment to society outside of prison, poor paying & labor intensive work, societal stigma around ex-inmates, “Legitimate opportunities for earnings therefore compete poorly with illegitimate sources of income.”(Rossi 8) to which they go on to address they cant quantify income from illegitimate means as well. Some societal systems provide challenges for ex convicts, and have a propensity for facilitating recidivism. We were fortunate enough to discover some hopeful sides of those societal systems that can reduce recidivism.

Discussions and Conclusions: ***This is a summary of your argument or experiment/research, and it should be related to the introduction.***

Throughout our experiment, it was important to make sure that we approached our analysis in an unbiased fashion, which was explained in our description of our variable choice. Because of that, we were able to isolate factors of employment that had a direct effect on recidivism. As shown by the table in the methods section, employment for over half the recorded experimental period has an immense effect on recidivism. In this right, our null hypothesis can be rejected in terms of the effect of employment on recidivism, as the proportion of difference is nearly more than 30% difference in rearrest rate. As for financial aid, our permutation showed that very few of the simulations actually resulted in a change in rates of recidivism. In this right, our analysis of the variables chosen was able to dodge inherent bias based on variables such as race, marriage status, or age, by isolating two variables related to financial stability. Due to this, our initial question was ratified, as there was shown to be a direct relationship between consistent employment and the rate of recidivism that an individual experienced. Our analysis showed that variables focused around financial stability were the most influential overall. This also proves that our initial hypothesis and guiding questions were correct in deciding where to put our efforts.

Works Cited

Rossi, P.H., R.A. Berk, and K.J. Lenihan (1980). *Money, Work, and Crime: Some Experimental Results.* New York: Academic Press. https://www.gwern.net/docs/sociology/1980-rossi-moneyworkandcrime.pdf

Zeisel, Hans. “Disagreement over the Evaluation of a Controlled Experiment.” American Journal of Sociology, vol. 88, no. 2, University of Chicago Press, 1982, pp. 378–89, http://www.jstor.org/stable/2779554