Math 150 - Methods in Biostatistics - Project Submission 1

Victor Machado

April 19, 2019

```
knitr::opts_chunk$set(message=FALSE, warning=FALSE, fig.height=3.5, fig.width=8,
                      fig.align = "center")
library(tidyverse)
library(broom)
library(tidylog)
library(survival)
library(survminer)
library(readr)
```

The CD4 count is a "snapshot of how well your immune system is functioning". CD4 cells are white blood cells that fight infection. The more a person has the btter, these are the cells that the HIV virus kills. When CD4 count drops below 200, a person is diagnosed with AIDS. A normal range for CD4 cells is about 500-1,500 cells/mm3. The following histogram shows that more than half of the patients in the study had this count at lower than 150 cells/mm3.

(https://www.hiv.va.gov/patient/diagnosis/labs-CD4-count.asp)

```
AIDSdata <- read_csv("AIDSdata.csv")
cd4 <- AIDSdata$cd4
hist(cd4)
```
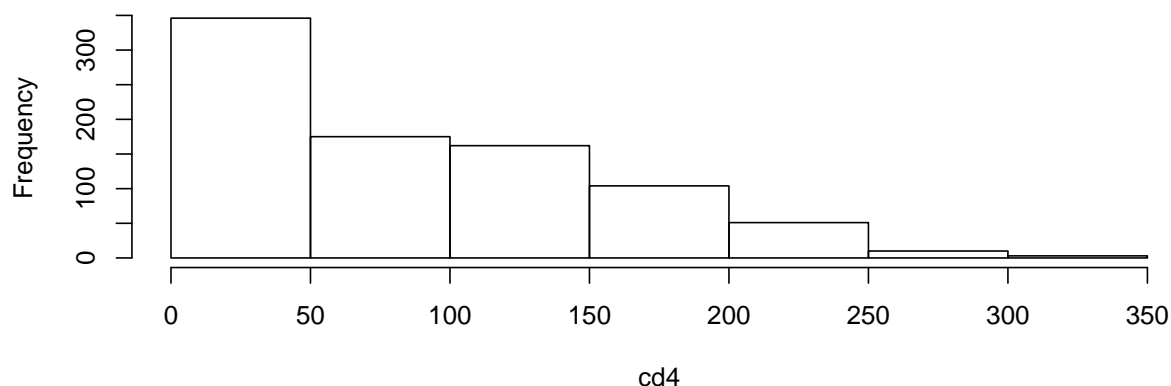

The following barplot shows that most of the patients in the study were males 84%. According to the CDC, from 1996-2000, out of all the people with AIDS 77.4% were male.
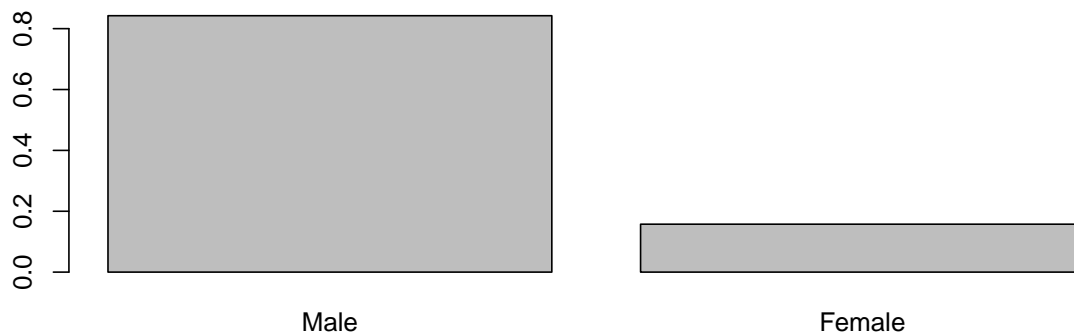
(https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5021a2.htm#tab1)

```
sex <- factor(AIDSdata$sex, labels=c("Male", "Female"))
summary(sex)
```

```
##   Male Female
##    717    134
```

```r
prop.table(table(sex))
```

```
## sex
##      Male    Female
## 0.8425382 0.1574618
```

```r
barplot(prop.table(table(sex)))
```


In the study, 53% of the patients who participated were White, 28% of the patients were Black, 18% of the patients were Hispanic, 1% of the patients were Asian or Pacific Islander and 1% of the patients were American Indian or Alaskan Native.

However, according to the CDC, from 1996-2000, the highest rate of people with AIDS by race was Black with 44.9%, then White with 34%, Hispanic with 19.7%, Asian/Pacific Islander with 0.8% and American Indian/Alaska Native with 0.4%.

(https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5021a2.htm#tab1)

```r
ethnicity <- factor(AIDSdata$raceth, labels=c("White", "Black", "Hispanic", "Asian/Pacific", "Native"))
summary(ethnicity)
```

```
##         White         Black      Hispanic Asian/Pacific        Native
##           447           235           150            10             9
```

```r
prop.table(table(ethnicity))
```

```
## ethnicity
##         White         Black      Hispanic Asian/Pacific        Native
##    0.52526439    0.27614571    0.17626322    0.01175088    0.01057579
```

```r
barplot(prop.table(table(ethnicity)))
```
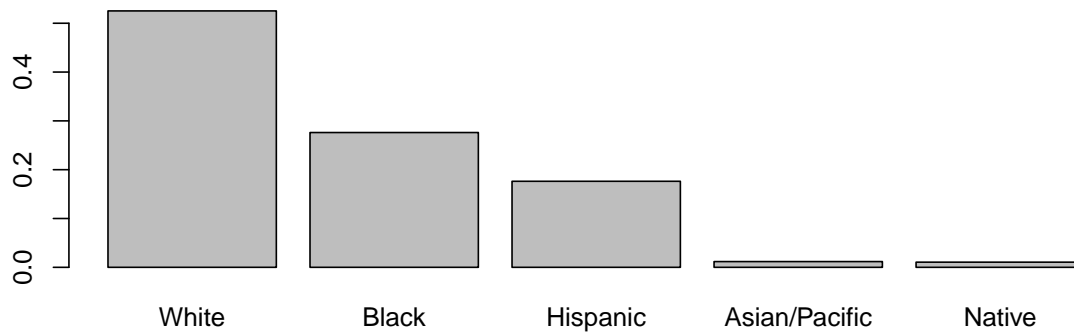
A total of 1156 patients not previously treated with lamivudine (3TC) or protease inhibitors were stratified according to CD4 cell count (50 or fewer vs 51 to 200 cells per cubic millimeter) and then were randomly assigned to one of two daily regiments:
600mg of zidovudine (ZDV) and 300 mg of lamivudine (3TC) or
The same regiment along with 2400 mg of indivanir.
Stavudine could be substituted for zidovudine.
The primary end point was the time to the development of acquired immunodeficiency syndrome (AIDS) or death.

```r
stratified_data <- factor(AIDSdata$strat2, label=c("CD4<=50","CD4>50"))
barplot(prop.table(table(stratified_data)))
```
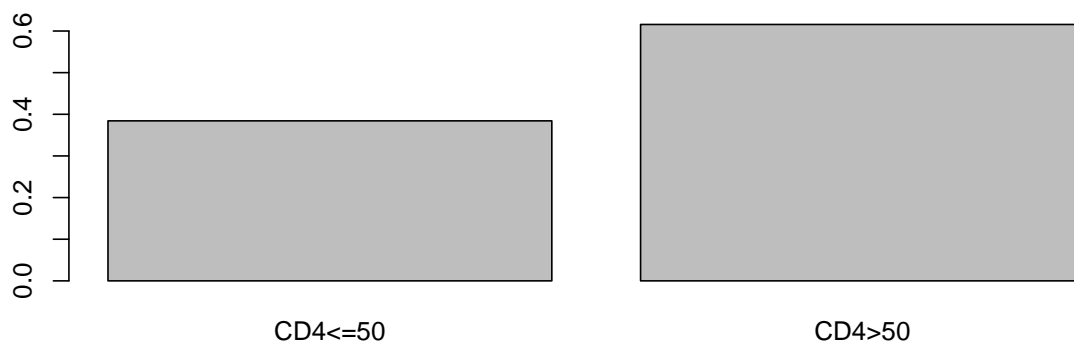

The patients, had to be more than 16 years old and had to have laboratory documentation of HIV-1 infection, a CD4 cell count of 200 per cubic millimeter or less within the 60 days before entry into the study, at least 3 months of zidovudine treatment, no more than 1 week of prior lamivudine treatment, no prior treatment with protease inhibitors and a Karnofsky score of at least 70.

Below we can see the proportion of patients according to their karnofsky score, where
100 means normal no complaint/no evidence of the disease 90 means normal activity possible; minor signs/symptoms of the disease 80 means normal activity with effort;some signs/symptoms of the disease 70

means care for self; normal activity/active work not possible

```
karnof <- factor(AIDSdata$karnof, label=c("100", "90", "80", "70"))
barplot(prop.table(table(karnof)))
```
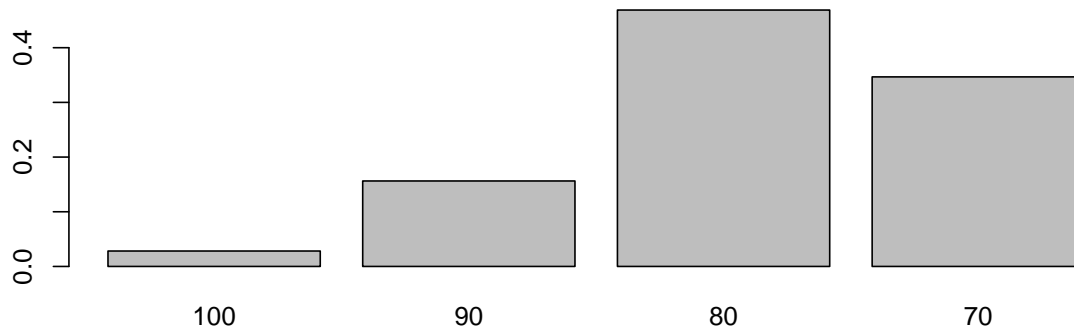

IV drug use history was also kept track of, 84% of the patients in the study had never used IV drugs and about 16% had previously used IV drugs or never used it. Meaning many of these patients were using IV drugs for the first time.
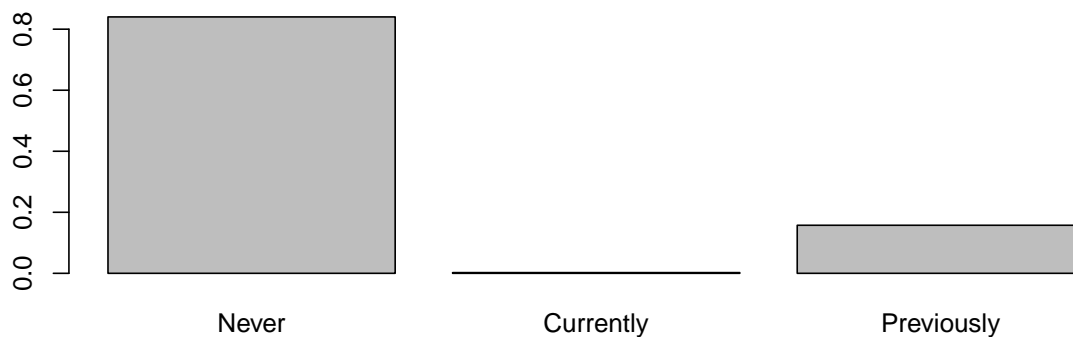
```
ivhistory <- factor(AIDSdata$ivdrug, label=c("Never", "Currently", "Previously"))
summary(ivhistory)
```

```
##      Never  Currently Previously
##        715          2        134
```

```
prop.table(table(ivhistory))
```

```
## ivhistory
##       Never   Currently  Previously
## 0.840188014 0.002350176 0.157461810
```

```
barplot(prop.table(table(ivhistory)))
```

For something new, I would like to include bootstrapping the survival model. I am planning on using the book Bootstrap Methods and their Application by A.C. Davison and D.V. Hinkley, specifically section 3.5 and the initial sections for preliminary information.

In the section it mentions a few different ways of resampling based on different assumptions. One is called Conditional bootstrapping for censored data, where censoring times are resampled from the conditional censoring distribution. Another is the "weird"" bootstrap, which is used when interest is focused on the survival or hazard functions. Another is called the "cases" bootstrap which is, "ordinary bootstrap sampling under the random censorship model".

I would like to learn more about these different techniques and implement these simulations for bootstrapping as well as the assumptions, basic ideas and the purpose of bootstrapping.