Algorithmic Fairness in Practice

Now that we've discussed algorithmic fairness and bias, let's work through a real example and see how an algorithm can be biased.

In May 2016, Jeff Larson and others from ProPublica published a story about algorithmic bias in criminal justice risk assessment scores. These scores are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts to even more fundamental decisions about defendants' freedom. In Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin, the results of such assessments are given to judges during criminal sentencing.

In 2014, then U.S. Attorney General Eric Holder warned that the risk scores might be injecting bias into the courts. He called for the U.S. Sentencing Commission to study their use. "Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice," he said, adding, "they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society." The sentencing commission did not, however, launch a study of risk scores. So, ProPublica did, as part of a larger examination of the powerful, largely hidden effect of algorithms in American life.

ProPublica obtained the risk scores assigned to more than 7,000 people arrested in Broward County, Florida, in 2013 and 2014 and checked to see how many were charged with new crimes over the next two years. The score proved remarkably unreliable in forecasting violent crime. In addition, ProPublica was able to show the algorithm was racially biased.

ProPublica completed a thorough analysis involving logistic regression, survival analysis and other statistical methods (check out more details here if interested), but here we will be exploring how the algorithm is biased and communicating this bias.

The data for ProPublica's analysis is contained in the file `compas-scores-two-years.csv`. Below are the variables we will be using:

- `race`: Race of the individual (we will only focus on `African-American` and `Caucasian` race categories).
- `two_year_recid`: Indicator if the individual reoffended (commited another crime) within 2 years.
- `decile_score`: Risk score, 1-10, 1 being low and 10 being high.
- `score_text`: score group, "Low": `decile_score` = 1-3, "Medium": `decile_score` = 4-7, "High": `decile_score` = 8-10.

Question 1

While there are several race/ethnicity categories represented in this dataset, we will limit our analyses to those who self-identified as Caucasian or African-American. Read in the data and filter the data frame to only include Caucasian and African-American individuals. How many African-American individuals are represented in this dataset and how many Caucasian individuals are represented?

```r
library(tidyverse)
library(ggplot2)
library(readr)
```

```r
# Your code here
```

Question 2

Make 2 bar charts of `decile_score`, one for each race group. What do you notice about the distributions of scores for the two groups?

```
# Your code here
```

Question 3

Is the risk score a good predictor of two-year recidivism (i.e., committing another crime within 2 years)? Create a new variable called `binary_score` that is equal to 0 if `score_text` is equal to "Low" (this will be the "low-risk" group) and 1 otherwise (this will be the "high-risk" group). Create a 2x2 table of `binary_score` and `two_year_recid` using the `table` function. Calculate accuracy, sensitivity, specificity, false positive rate and false negative rate by hand. What is the accuracy? Are the sensitivity and specificity balanced? Are the false positive rate and false negative rate balanced?

- Here, false positive rate is the number of false positives over the total number of true negatives, and false negative rate is the number of false negatives over the total number of true positives.

```
# Your code here
```

Question 4

Now calculate the accuracy, sensitivity, specificity, false positive rate and false negative rate for each race group. Does the algorithm perform better for one group over the other? Describe how the model is biased.

- Hint: think about what false positives, false negatives, false positive rate and false negative rate mean in this context.

```
# Your code here
```