

# 1 The COMPAS data

Table 1 shows the [confusion matrices](#) for Black and white defendants, respectively, based on the actual COMPAS data.

Table 1: Confusion matrices for COMPAS defendants.

| (a) Black defendants |      |      |       | (b) white defendants |      |      |       |
|----------------------|------|------|-------|----------------------|------|------|-------|
|                      | Low  | High | Total |                      | Low  | High | Total |
| Surv                 | 990  | 805  | 1795  | Surv                 | 1139 | 349  | 1488  |
| Recid                | 532  | 1369 | 1901  | Recid                | 461  | 505  | 966   |
| Total                | 1522 | 2174 | 3696  | Total                | 1600 | 854  | 2454  |

## 1.1 Disparate impact

Consider the following mathematical definition of **disparate impact**:

$$\frac{\Pr(High|White)}{\Pr(High|Black)} < 1 - \epsilon,$$

where  $\epsilon$  is a small positive constant. An algorithm exhibits  $\epsilon$ -disparate impact if and only if this inequality is true.

- (1) Explain in words what it would mean for the LHS to equal 1?
- (2) How far from 1 would you tolerate this score deviating before alleging bias? Determine a value for  $\epsilon$  in your group.
- (3) Compute the disparate impact score using the definition and confusion matrices in Table 1.
- (4) How does this score relate to ProPublica's allegations in Angwin et al. (2016)?

## 1.2 Equalized Odds

An algorithm exhibits  $\epsilon$ -**equalized odds** if and only if, for some  $\epsilon > 0$ :

$$|\Pr(High|White \cap Survived) - \Pr(High|Black \cap Survived)| < \epsilon$$

- (5) Does the COMPAS algorithm exhibit equalized odds (using your value of  $\epsilon$ )?
- (6) Is COMPAS biased according to equalized odds? What values of  $\epsilon$  would you be willing to tolerate?

### 1.3 Well-calibration

An algorithm is said to be  $\epsilon$ -**well-calibrated** if and only if:

$$|\Pr(High|Black) - \Pr(Recidivated|Black)| < \epsilon,$$

for some  $\epsilon > 0$  and *for all groups* (e.g., the property holds for white defendants as well).

- (7) Is COMPAS well-calibrated? What, if any, rebuttal would you make against ProPublica's claims of bias?

## 2 Competing notions of fairness

Barocas et al. (2023) show that while there are dozens of competing notions of statistical fairness, all are mathematical equivalent to just three (see [their dictionary](#)), which they call **independence**, **separation**, and **calibration** (see Section 1.3). Independence is equivalent to the disparate impact that we calculated in Section 1.1, and separation is equivalent to the equalized odds we calculated in Section 1.2. Thus, the three measures we calculated form a complete set.

- (8) Does COMPAS meet all three statistical criteria for fairness? Does it meet *any*?

## 3 The impossibility of fairness

Kleinberg et al. (2016) show that under all but the most trivial of conditions, satisfying all three notions of fairness is impossible.

- (9) What are the implications of this for our discussions of fairness?

## References

- Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. *Machine Bias*. ProPublica; ProPublica. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press. <https://mitpress.mit.edu/9780262048613/fairness-and-machine-learning/>.
- Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *arXiv Preprint arXiv:1609.05807*. <https://arxiv.org/abs/1609.05807/>.