Fair Prediction with Disparate Impact

- **Task**: Binary classification → but results extend to multiclass classification
- **Important assumptions**: The observed outcome Y is a suitable outcome measure for the purpose of assessing fairness properties
- **What the paper shows**
  - Different fairness criteria cannot be all simultaneously satisfied when prevalence differs across groups
  - Disparate impact can arise when a (recidivism) prediction instrument fails to satisfy the criterion of error rate balance

- **Paper outline**
  - # fairness criteria
  - showing that instrument that satisfies predictive parity cannot have equal FPR and FNR across groups when prevalence differs across those groups
  - Differences in FPR & FNR can result in disparate impact

- **Notation**
  - S: the risk score → higher values → higher risks
  - X = Covariates
  - $R \in \{w, b\}$: The group to which individuals belong - It might be an element of X
  - $Y \in \{0, 1\}$: the outcome - 1 if the indiv recidivates
  - s_{HR}: high-risk score threshold

- Definitions
  - Calibration: A risk score S is said to be **well-calibrated** if it reflect the same likelihood of recidivism irrespective of the individual's group

$$P(Y=1 \mid S=s, R=b) = P(Y=1 \mid S=s, R=w), \quad \forall s$$

  - Predictive parity: A score S satisfies **predictive parity** at a threshold S_{HR} if if the probability of recidivism among high risk offenders is the same regardless of the group membership

$$P(Y=1 \mid S \geq S_{HR}, R=b) = P(Y=1 \mid S \geq S_{HR}, R=w)$$

⚠ A model that is well-calibrated is not guaranteed to satisfy predictive parity

  - Error-rate balance: A score S satisfies **error-rate balance** at a threshold S_{HR} If the FPR & FNR are equal across groups:

$$\begin{cases} P(S > S_{HR} \mid Y=0, R=b) = P(S > S_{HR} \mid Y=0, R=w) \\ P(S \leq S_{HR} \mid Y=1, R=b) = P(S \leq S_{HR} \mid Y=1, R=w) \end{cases}$$

  - Statistical parity: A score S satisfies **statistical parity** at a threshold S_{HR} if

$$P(S > S_{HR} \mid R=b) = P(S > S_{HR} \mid R=w)$$

- **Results:**
  - If a score satisfies predictive parity while the groups have ≠ prevalence, the score cannot achieve equal FPR & FNR across groups

$$\mathrm{FPR} = \frac{p}{1-p}\frac{1-\mathrm{PPV}}{\mathrm{PPV}}(1-\mathrm{FNR}).$$

- One has no control over the event prevalence in each group: p. Yet, one can choose to:

(i) Allow unequal false negative rates to retain equal PPV's and achieve equal false positive rates
(ii) Allow unequal false positive rates to retain equal PPV's and achieve equal false negative rates
(iii) Allow unequal PPV's to achieve equal false positive and false negative rates

↳ In case (iii): Having $FPR_w = FPR_b$ & $FNR_w = FNR_b$ One can allow unequal PPV values & set different s_{HR} for each group