# New Definition of Fairness

Suppose, the protected attribute is G , predicted value is $\hat{y}$ and the true value is $y$. Now if the value $\frac{P(G|\hat{y},y)}{P(G|y)}$ is below some particular threshold, we can suspect unfairness.

## Motivation Towards the New Definition:

We know **Equality of Odds** is satisfied if the prediction $\hat{y}$ is conditionally independent to the protected attribute $G$, given the true value $y$:

$$P(\hat{y}|y, G) = P(\hat{y}|y)$$

.

Also, we say that a classifier f has disparate impact if

$$\frac{P(f(y) = 1|G = 0)}{P(f(y) = 1|G = 1)} \leq \tau$$

where $\tau$ is the threshold.

Being motivated by this, we may say that if the value $\frac{P(\hat{y}|y,G)}{P(\hat{y}|y)}$ is below some particular threshold, there may be unfairness. Now

$$\begin{aligned}\frac{P(\hat{y}|y, G)}{P(\hat{y}|y)} &= \frac{P(\hat{y}, y, G)P(y)}{P(\hat{y}, y)P(y, G)} \\ &= \frac{P(G|\hat{y}, y)}{P(G|y)}\end{aligned}$$

So if the value $\frac{P(G|\hat{y},y)}{P(G|y)}$ is below some particular threshold, we suspect unfairness.