Auditing Robust Fairness Metrics

1 Statistical Parity Robustness

1.1 Preliminaries

We will first define some preliminaries.

x : vector of protected attributes. In most basic example, x is 1-dimensional and $x \in [0, 1]$

x' : vector of protected attributes. In most basic example, x' is 1-dimensional and $x' \in [0, 1]$

y : predicted output label, $y \in [0, 1]$ for all dimensions of x and x′

row_i : individual represented as (x_i, x'_i, y_i)

n : size of dataset (i.e., number of rows)

D : dataset D is a set of rows $\{(x_i, x'_i, y_i)\}$ for $i \in [1, n]$

From these preliminaries, we define the following sets, representing subgroups of the global population.

d_{11} : $\{i \in d_{11} | y_i = 1 \wedge x_i = 1\}$

d_{10} : $\{i \in d_{10} | y_i = 1 \wedge x_i = 0\}$

d_1 : $\{i \in d_1 | x_i = 1\}$

We take the definition of γ statistical parity to be

$$|P[y = 1|x = 1] - P[y = 1|x = 0]| < \gamma,$$

and derive the equivalent representation

$$\frac{|d_{11}|}{|d_1|} - \frac{d_{10}}{n - |d_1|} < \gamma$$