Fairness and discrimination in actuarial pricing

Arthur Charpentier\textsuperscript{1} \& Laurence Barry\textsuperscript{2}

\textsuperscript{1} Univeristé du Québec à Montréal \hspace{1cm} \textsuperscript{2} Chaire Pari

Convention A - ASTIN-IAA - September 2022
Agenda

- Charpentier (2022) Insurance: Discrimination, Biases and Fairness, Institut Louis Bachelier
Notations

\[
\begin{align*}
    y \in \{0, 1\} & \quad \text{variable of interest (classically binary)} \\
    p \in \{0, 1\} & \quad \text{protected variable (sensitive)} \\
    x \in \mathbb{R}^d & \quad \text{‘explanatory’ variables} \\
    s \in [0, 1] & \quad \text{score, classically } s = s(x, p) \\
    \hat{y} \in \{0, 1\} & \quad \text{classifier, classically } \hat{y} = 1(s > t)
\end{align*}
\]

**Fairness Through Unawareness**, Kusner et al. (2017)
Protected attribute \( p \) is not explicitly used in decision function \( \hat{y} \).
Ethics, Fairness and Discrimination

- **Accuracy:** \( \pi(x) = \mathbb{E}_{P}[Y|X = x] \) (historical probability) (is)
- **Fairness:** \( \pi^*(x) = \mathbb{E}_{P^*}[Y|X = x] \) (targeted probability) (ought, Hume (1739))

- “Technology is neither good nor bad; nor is it neutral”, Kranzberg (1986)
- “Machine learning won’t give you anything like gender neutrality ‘for free’ that you didn’t explicitly ask for”, Kearns and Roth (2019)
- “at the core of insurance business lies discrimination between risky and non-risky insureds”, Avraham (2017)

Discrimination could be described as “actuarially fair”, therefore characterizing “fair discrimination" when the difference in premium reflects a difference in risk, Meyer and Rothstein (2004),

- “Insurance rates are color-blind and solely based on risk”, (Property Casualty Insurers Association of America, 2015 CFA study) cited in https://evolutionofraceandinsurance.org/2000s
"actuarial fair discrimination"?

Bohren et al. (2019) on statistical discrimination, or efficient discrimination, as in Becker (1957) (inspired by Edgeworth (1922) up to Phelps (1972))

Becker (2005) says “if young Moslem Middle Eastern males were in fact much more likely to commit terrorism against U.S. than were other groups, putting them through tighter security clearance would reduce current airport terrorism”,

“racial profiling” is “effective”, even though “such profiling is ‘unfair’ to the many young male Moslems who are not terrorists, and to the many minority shoppers who are honest ...”
## Protected Attributes?

<table>
<thead>
<tr>
<th></th>
<th>CA</th>
<th>HI</th>
<th>GA</th>
<th>NC</th>
<th>NY</th>
<th>MA</th>
<th>PA</th>
<th>FL</th>
<th>TX</th>
<th>AL</th>
<th>ON</th>
<th>NB</th>
<th>NL</th>
<th>QC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>X</td>
<td>X</td>
<td>✗</td>
<td>X</td>
<td>✗</td>
<td>X</td>
<td>✗</td>
<td>X</td>
<td>X</td>
<td>✗</td>
<td>X</td>
<td>✗</td>
<td>X</td>
<td>✗</td>
</tr>
<tr>
<td>Age</td>
<td>X</td>
<td>X</td>
<td>✗</td>
<td>X</td>
<td>✗</td>
<td>X</td>
<td>✗</td>
<td>X</td>
<td>X</td>
<td>✗</td>
<td>X</td>
<td>✗</td>
<td>X</td>
<td>✗</td>
</tr>
<tr>
<td>Driving experience</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Credit history</td>
<td>X</td>
<td>X</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Education</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>✗</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>✗</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Occupation</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>✗</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>✗</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Employment status</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>✗</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>✗</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Marital status</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Housing situation</td>
<td>X</td>
<td>X</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Address/ZIP code</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Insurance history</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>

Data & Models I

\[
\begin{align*}
  y \in \{0, 1\} & \quad \text{variable of interest (classically binary)} \\
  p \in \{0, 1\} & \quad \text{protected variable (sensitive)} \\
  x \in \mathbb{R}^d & \quad \text{‘explanatory’ variables} \\
  s \in [0, 1] & \quad \text{score, classically } s = s(x, p) \\
  \hat{y} \in \{0, 1\} & \quad \text{classifier, classically } \hat{y} = 1(s > t)
\end{align*}
\]

**Fairness Through Unawareness**, Kusner et al. (2017)

Protected attribute $p$ is not explicitly used in decision function $\hat{y}$. 
policyholder
$(x_i, p_i)$
Data & Models III

policyholder

\((x_i, p_i)\)

inferred data

\(m_d(x, p)\)

possibles scores
Data & Models IV

- commercial discount
- multiple decisions from policyholder (related to various personal features)
- multiple decisions from agent / broker

\[
\begin{align*}
\text{policyholder} & : (x_i, p_i) \\
\text{underwriting} & : m_u(x, p) \\
\text{infered data} & : m_d(x, p)
\end{align*}
\]
Data & Models V

-.policyholder\((x_i, p_i)\)

- underwriting\(m_u(x, p)\)

- claim a loss\(\ell(x, p)\)

(infered data
\[m_d(x, p)\]

(true value of the loss)
Data & Models VI

- Policyholder: \( (x_i, p_i) \)
- Underwriting: \( m_u(x, p) \)
- Claim a loss: \( \ell(x, p) \)
- Claims manager: \( m_c(x, p) \)
- Inferred data: \( m_d(x, p) \)

- Go to court?
- Possible discrimination
- Internal process
Data & Models VII

- policyholder $(x_i, p_i)$
- underwriting $m_u(x, p)$
- claim a loss $\ell(x, p)$
- claims manager $m_c(x, p)$
- fraud detection $m_f(x, p, x'(p))$
- fraud scores

inferred data $m_d(x, p)$
policyholder \((x_i, p_i)\)

underwriting \(m_u(x, p)\)

claim a loss \(\ell(x, p)\)

indemnity \(y_i\)

inferred data \(m_d(x, p)\)

claims manager \(m_c(x, p)\)

fraud detection \(m_f(x, p, x'(p))\)
policyholder
$(x_i, p_i)$

underwriting
$m_u(x, p)$

claim a loss
$\ell(x, p)$

indemnity
$y_i$

inferred data
$m_d(x, p)$

claims manager
$m_c(x, p)$

fraud detection
$m_f(x, p, x'(p))$

dataset
$(y_i, x_i, p_i)$

(training data)
Defining Group Fairness

**Demographic Parity**, (Corbett-Davies et al. (2017), Agarwal (2021))

Decision function $\hat{y}$ satisfies demographic parity if $Y \perp P$, i.e.

$$P[\hat{Y} = y | P = 0] = P[\hat{Y} = y | P = 1], \forall y \text{ or } E[\hat{Y} | P = 0] = E[\hat{Y} | P = 1]$$

We can compare $s(X)$ conditional on $Y$, but also on $P$.
Defining Group Fairness II

**Equal Opportunity**, Hardt et al. (2016)

True positive parity $\Pr[\hat{Y} = 1|P = 0, Y = 1] = \Pr[\hat{Y} = 1|P = 1, Y = 1]$ or false positive parity $\Pr[\hat{Y} = 1|P = 0, Y = 0] = \Pr[\hat{Y} = 1|P = 1, Y = 0]$. 

![Graphs showing true and false positive rates](image-url)
<table>
<thead>
<tr>
<th>Definition</th>
<th>Authors</th>
<th>Formula</th>
<th>Independence/Conditional Independence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical Parity</td>
<td>Dwork et al. (2012)</td>
<td>$P[\hat{Y} = 1</td>
<td>P = p] = \text{cst, } \forall p$</td>
</tr>
<tr>
<td>Conditional Statistical Parity</td>
<td>Corbett-Davies et al. (2017)</td>
<td>$P[\hat{Y} = 1</td>
<td>P = p, X = x] = \text{cst}_x, \forall p, y \quad \hat{Y} \perp \perp P$</td>
</tr>
<tr>
<td>Equalized Odds</td>
<td>Hardt et al. (2016)</td>
<td>$P[\hat{Y} = 1</td>
<td>P = p, Y = y] = \text{cst}_y, \forall p, y$</td>
</tr>
<tr>
<td>Equalized Opportunity</td>
<td>Hardt et al. (2016)</td>
<td>$P[\hat{Y} = 1</td>
<td>P = p, Y = 1] = \text{cst, } \forall p$</td>
</tr>
<tr>
<td>Predictive Equality</td>
<td>Corbett-Davies et al. (2017)</td>
<td>$P[\hat{Y} = 1</td>
<td>P = p, Y = 0] = \text{cst, } \forall p$</td>
</tr>
<tr>
<td>Balance (Positive)</td>
<td>Kleinberg et al. (2017)</td>
<td>$\mathbb{E}[S</td>
<td>P = p, Y = 1] = \text{cst, } \forall p$</td>
</tr>
<tr>
<td>Balance (Negative)</td>
<td>Kleinberg et al. (2017)</td>
<td>$\mathbb{E}[S</td>
<td>P = p, Y = 0] = \text{cst, } \forall p$</td>
</tr>
<tr>
<td>Conditional Accuracy Equality</td>
<td>Berk et al. (2017)</td>
<td>$P[Y = y</td>
<td>P = p, \hat{Y} = y] = \text{cst}_y, \forall p, y$</td>
</tr>
<tr>
<td>Predictive Parity</td>
<td>Chouldechova (2017)</td>
<td>$P[Y = 1</td>
<td>P = p, \hat{Y} = 1] = \text{cst, } \forall p$</td>
</tr>
<tr>
<td>Calibration</td>
<td>Chouldechova (2017)</td>
<td>$P[Y = 1</td>
<td>P = p, S = s] = \text{cst}_s, \forall p, s$</td>
</tr>
<tr>
<td>Well-Calibration</td>
<td>Chouldechova (2017)</td>
<td>$P[Y = 1</td>
<td>P = p, S = s] = s, \forall p, s$</td>
</tr>
<tr>
<td>Accuracy Equality</td>
<td>Berk et al. (2017)</td>
<td>$P[\hat{Y} = Y</td>
<td>P = p] = \text{cst, } \forall p$</td>
</tr>
<tr>
<td>Treatment Equality</td>
<td>Berk et al. (2017)</td>
<td>$\frac{\text{FN}_p}{\text{FP}_p} = \text{cst}_p, \forall p$</td>
<td></td>
</tr>
</tbody>
</table>
Individual Fairness

Consider some distances $D$ on $\{0, 1\} \times \{0, 1\}$ or $[0, 1] \times [0, 1]$, and $d$ on $\mathbb{R}^p \times \mathbb{R}^p$.

**Lipschitz property**, Duivesteijn and Feelders (2008)

$$D(\hat{y}_i, \hat{y}_j) \text{ or } D(s_i, s_j) \leq d(x_i, x_j), \ \forall i, j = 1, \ldots, n.$$ 

**Counterfactual fairness**, Kusner et al. (2017) If the prediction in the real world is the same as the prediction in the counterfactual world where the individual would have belonged to a different demographic group, we have counterfactual equity, i.e.

$$\mathbb{P}[Y_{P^{-p}}^* = y|X = x] = \mathbb{P}[Y_{P^{-p'}}^* = y|X = x], \ \forall p', x, y.$$
Penalizing to get a more fair pricing model

Inspired by Goodfellow et al. (2018), to avoid un-fairness: penalize to a dependence measure between \( \hat{y} \) and \( p \), e.g. for demographic parity

\[
\arg\min_{\theta} \left\{ \mathcal{L}(h_\theta(x), y) + \lambda \text{corr}(\hat{y}, p) \right\}
\]

or for equalized odds, if \( y \in \{0, 1\} \)

\[
\arg\min_{\theta} \left\{ \mathcal{L}(h_\theta(x), y) + \lambda_0 \text{corr}(\hat{y}, p|y = 0) + \lambda_1 \text{corr}(\hat{y}, p|y = 1) \right\}
\]

(that could be extented to more general \( y \)’s), see Grari et al. (2022).


