PHIL 408Q/PHPE 308D
Fairness

Eric Pacuit, University of Maryland

May 2, 2024

Research on algorithmic fairness studies the prospects of unbiased assessment. Bias in error rates is one form of bias, but not the only form and often considered not the most important form. Can bias in error rates and other important forms of bias be simultaneously eliminated?

One lesson that emerges from some of these studies is that eliminating one form of bias can mean that it is impossible to eliminate another. Sometimes, then, we face a conflict between eliminating different forms of bias.

Research on algorithmic fairness studies the prospects of unbiased assessment. Bias in error rates is one form of bias, but not the only form and often considered not the most important form. Can bias in error rates and other important forms of bias be simultaneously eliminated?

One lesson that emerges from some of these studies is that eliminating one form of bias can mean that it is impossible to eliminate another. Sometimes, then, we face a conflict between eliminating different forms of bias.

Here, I argue that, not only do we face a conflict in eliminating different forms of bias, we also face a conflict in eliminating one form of bias across different groupings. Eliminating a certain form of bias across groups for one way of categorizing people in a population can mean that it is impossible to eliminate that form of bias across groups for another way of classifying them.

Rush T. Stewart (2022). *Identity and the limits of fair assessment*. Journal of Theoretical Politics 2022, Vol. 34(3), pp. 415 - 442.

Setup

A single property y of interest.

Individuals in N either have property y or lack it: $Y : N \to \{0, 1\}$

Call a function $h : N \to [0, 1]$ an assessor.

For concreteness, interpret $h(i)$ as the assessor's probability that i has property y.

Setup

The quantity $P(Y = 1) = \mu$, for example, is the proportion of people in N that have property y, the prevalence of y in the population.

Call μ the base rate for y in N.

Given a **partition** $\pi = \{G_1, \ldots, G_m\}$ of N, let $P_k = P(\cdot \mid G_k)$. So, $P_1(Y = 1) = \mu_1$ is the base rate for y in group 1 is μ_1 and $P_2(h = 0.5)$ is the proportion of people to which h assigns 0.5 in G_2, and so on.

Strong Calibration

An assessor is **(strongly) calibrated** if

$$P_k(Y = 1 \mid h = p) = p \text{ for all } p \in [0, 1] \text{ and } k = 1, 2, \ldots, m \text{ such that } P_k(h = p) > 0.$$

Strong Calibration

An assessor is **(strongly) calibrated** if

$$P_k(Y = 1 \mid h = p) = p \text{ for all } p \in [0, 1] \text{ and } k = 1, 2, \ldots, m \text{ such that } P_k(h = p) > 0.$$

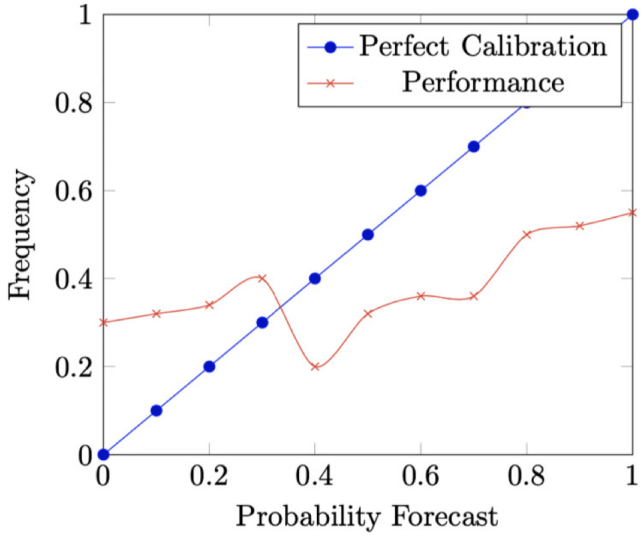
E.g., consider weather forecasting. Suppose that each day, a forecaster announces a probability of rain for that day. The forecaster is calibrated if it rains on 10% of the days she announces that it will rain with probability 0.1, and it rains on 85% of the days she predicts rain with probability 0.85, etc.

Strong Calibration

If the assessor is calibrated, not only would it not be **overconfident** in one group and **underconfident** in another, it would not be over- or underconfident in any of its assessments.

When the assessor is calibrated, Kleinberg et al. write "we are justified in treating people with the same score comparably with respect to the outcome, rather than treating people with the same score differently based on the group they belong to."

Calibration



Predictive Equity

An assessor h satisfies **predictive equity** (also called **weak calibration for groups**) for a partition π if

$$P_k(Y = 1 \mid h = p) = P_j(Y = 1 \mid h = p) \text{ for all } G_k, G_j \in \pi$$

Put another way, among people assigned the same assessment score, the proportion of people who have property y is the same across all groups in the partition.

Limitation Result

An assessor is **perfect** if $h(i) = Y(i)$ for all $i \in N$.

Limitation Result

An assessor is **perfect** if $h(i) = Y(i)$ for all $i \in N$.

Observation 1. Let h be an assessor for N. The following are equivalent:

1. h is calibrated for all binary partitions.
2. h is calibrated for all partitions.
3. h is perfect.

Limitation Result

An assessor is **perfect** if $h(i) = Y(i)$ for all $i \in N$.

Observation 1. Let h be an assessor for N. The following are equivalent:

1. h is calibrated for all binary partitions.
2. h is calibrated for all partitions.
3. h is perfect.

In other words, outside of the unrealistic case of perfect assessment, there will be bias in confidence against some group. Observation 1 complicates any automatic inference from failure of calibration for some group to *intentional* bias on behalf of the assessor.

Limitation Result 2

An assessor h makes **perfect distinctions** if, for all $i, j \in N$, $Y(i) \neq Y(j)$ implies that $h(i) \neq h(j)$. So, for any score p, if $h(i) = p$ and $Y(i) = 1$, then for no individual j such that $Y(j) = 0$ is it the case that $h(j) = p$.

Limitation Result 2

An assessor h makes **perfect distinctions** if, for all $i, j \in N$, $Y(i) \neq Y(j)$ implies that $h(i) \neq h(j)$. So, for any score p, if $h(i) = p$ and $Y(i) = 1$, then for no individual j such that $Y(j) = 0$ is it the case that $h(j) = p$.

Observation 2. Let h be an assessor for N. The following are equivalent:

1. h satisfies predictive equity for all binary partitions.
2. h satisfies predictive equity for all partitions.
3. h makes perfect distinctions.

Limitation Result 2

An assessor h makes **perfect distinctions** if, for all $i, j \in N$, $Y(i) \neq Y(j)$ implies that $h(i) \neq h(j)$. So, for any score p, if $h(i) = p$ and $Y(i) = 1$, then for no individual j such that $Y(j) = 0$ is it the case that $h(j) = p$.

Observation 2. Let h be an assessor for N. The following are equivalent:

1. h satisfies predictive equity for all binary partitions.
2. h satisfies predictive equity for all partitions.
3. h makes perfect distinctions.

Aside from assessors that make perfect distinctions, scores will not "mean" the same thing for all groups; there will be bias against some group. In large populations, perfect distinctions is very difficult to achieve—not as difficult as perfect assessment, but difficult nonetheless.

Two Objections

1. We might consider satisfying certain fairness constraints *approximately* rather than exactly. That is, we could confine the amount of bias to which any group is subject to a certain margin of tolerance.

Two Objections

1. We might consider satisfying certain fairness constraints *approximately* rather than exactly. That is, we could confine the amount of bias to which any group is subject to a certain margin of tolerance.

2. One might be inclined to think that, while (a particular type of) unbiased assessment for multiple partitions is often desirable, we have overshot the mark by requiring it for *all* partitions.

There are simple examples of populations that allow for a imperfect assessor that is simultaneously calibrated for, say, two different non-trivial ways of partitioning the population.

Calibration Across 2 Groups

Let $N = \{1, 2, 3, 4, 5, 6\}$, and let $Y(i) = 1$ for $i = 1, 5, 6$. Consider the following two partitions

1. $\{B, W\} = \{\{1, 2, 4\}, \{3, 5, 6\}\}$ and
2. $\{M, F\} = \{\{1, 2, 3\}, \{4, 5, 6\}\}$

	B	W
M	$h(1^*) = \frac{1}{2}, h(2) = 0$	$h(3) = \frac{1}{2}$
F	$h(4) = \frac{1}{2}$	$h(5^*) = \frac{1}{2}, h(6^*) = 1$

Calibration Across 2 Groups

Let $N = \{1, 2, 3, 4, 5, 6\}$, and let $Y(i) = 1$ for $i = 1, 5, 6$. Consider the following two partitions

1. $\{B, W\} = \{\{1, 2, 4\}, \{3, 5, 6\}\}$ and
2. $\{M, F\} = \{\{1, 2, 3\}, \{4, 5, 6\}\}$

	B	W
M	$h(1^*) = \frac{1}{2}, h(2) = 0$	$h(3) = \frac{1}{2}$
F	$h(4) = \frac{1}{2}$	$h(5^*) = \frac{1}{2}, h(6^*) = 1$

h is calibrated across $\{M, F\}$ *and* across $\{B, W\}$.

No Calibration Across Groups

Suppose $N = \{1, 2, 3\}$ with $Y(1) = Y(3) = 1$ and $Y(2) = 0$:

	B	W
M	1^*	2
F	3^*	

No Calibration Across Groups

Suppose $N = \{1, 2, 3\}$ with $Y(1) = Y(3) = 1$ and $Y(2) = 0$:

	B	W
M	1^*	2
F	3^*	

- Supposing that h is imperfect and calibrated for the $\{M, F\}$ partition of N implies that individual 1 must receive a score in $(0, 1)$.
- The only such assessment consistent with calibration is $h(1) = h(2) = 1/2$.
- But then h cannot calibrated for B since, by calibration for F, $h(3) = 1$.
- Similarly, h cannot be calibrated for W since $P_W(Y = 1 \mid h = 1/2) \neq 1/2$.