Discrimination and Interpretability of Predictive Models

Arthur Charpentier
with Marie-Pier Côté, Olivier Côté Agathe Fernandes-Machado
Ewen Gallic, François Hu, Marouane Il-Idrissi & Philipp Ratz

Institut des Actuaires, Paris, May 2025

"**Discrimination** *is the act, practice, or an instance of separating or distinguishing categorically rather than individually*," Merriam-Webster (2022).

FOUNDATION FOR SCIENCE

What is an "actuary"?

"*To be an* **actuary** *is to be a specialist in generalization, and actuaries engage in a form of decision making that is sometimes called actuarial. Actuaries guide insurance companies in making decisions about* **large categories that have the effect of attributing to the entire category certain characteristics that are probabilistically indicated by membership in the category, but that still may not be possessed by a particular member of the category**," Schauer (2006).

[Most] "*actuaries cannot think of individuals except as members of groups*" claimed Brilmayer et al. (1979). Each individual is assigned the same value as all other members of the group to which it is assigned.

See also Mowbray (1921) or Bailey and Simon (1960), or more recently Board (2005) and Finger (2006)

PROFILES

PROBABILITIES

AND

STEREOTYPES

FREDERICK SCHAUER

The Belknap Press of Harvard University Press
Cambridge, Massachusetts
London, England

generalization is the stock in trade of the insurance industry. Indeed, the insurance industry has its own name for this kind of decisionmaking. To be an *actuary* is to be a specialist in generalization, and actuaries engage in a form of decisionmaking that is sometimes called *actuarial.* Actuaries guide insurance companies in making decisions about large categories (teenage males living in northern New Jersey) that have the effect of attributing to the entire category certain characteristics (carelessness in driving) that are probabilistically indicated by membership in the category, but that still may not be possessed by a particular member of the category (this *particular* teenage male living in northern New Jersey).

Occasionally the actuarial generalizations of the insurance industry become controversial. One example is the use of generalizations about the comparative safety of different neighborhoods as a basis for setting the rates for homeowners' insurance or determining the willing-

What is an "actuarial model" (as in most actuarial textbooks)?

- linear regression on categories - "**segmentation**"

$+\beta_3$ ceteris paribus

$$\widehat{y}(\text{man}) = \beta_0 + \beta_1 \mathbf{1}_{\text{urban}} + \beta_2 \mathbf{1}_{\text{young}} + \boxed{\beta_3}\, \mathbf{1}_{\text{man}} = \widehat{y}(\text{woman}) + \boxed{\beta_3}$$

- Poisson regression (frequency) on categories, or not

$$\widehat{y}(\text{man}) = \exp\left[\beta_0 + \beta_1 \mathbf{1}_{\text{urban}} + \beta_2\, \boxed{\mathbf{1}_{\text{young}}} + \boxed{\beta_3}\, \mathbf{1}_{\text{man}}\right] = \widehat{y}(\text{woman}) \cdot \boxed{\exp[\beta_3]}$$

$\times e^{\beta_3}$ ceteris paribus

$$\widehat{y}(\text{man}) = \exp\left[\beta_0 + \beta_1 \mathbf{1}_{\text{urban}} + \beta_2\, \boxed{\text{age}} + \boxed{\beta_3}\, \mathbf{1}_{\text{man}}\right] = \widehat{y}(\text{woman}) \cdot \boxed{\exp[\beta_3]}$$

If β_3 small, $e^{\beta_3} \approx 1 + \beta_3$, i.e. "$\beta_3 = 0.2$" $\longleftrightarrow$ "+20% for men"

Thus "**interpretation**" is simple (if we do not discuss what "ceteris paribus" means).

“The myth of the actuary” (objectivity vs. subjectivity)

- The rhetoric of insurance exclusion – numbers, objectivity and statistics – forms what Brian Glenn calls “*the myth of the actuary*,” “*a powerful rhetorical situation in which decisions appear to be based on objectively determined criteria when they are also largely based on subjective ones*” or “**the subjective nature of a seemingly objective process**.” “*Virtually every aspect of the insurance industry is predicated on stories first and then numbers*,” Glenn (2000, 2003)
- Importance of **interpretation** and **explainability** of models
- Some models have a high accuracy... for wrong reasons...

“The myth of the actuary” (objectivity vs. subjectivity)

- E.g., classifiers, $y \in \{0,1\}$
- why a prediction of $\widehat{y} = 1$?

“*On a collection of additional 60 images, the classifier predicts “Wolf” if there is snow (or light background at the bottom), and “Husky” otherwise, regardless of animal color, position, pose, etc.*,” Ribeiro et al. (2016)

- Also, was $\widehat{m}(\boldsymbol{x}_i) = 1.32\%$ a good prediction if $y_i = 1$?

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the “Husky vs Wolf” task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: “Husky vs Wolf” experiment results.

From Econometrics to Machine Learning. Why could there be a problem?

- **Econometrics** is dead, long live "**artificial intelligence**"
- "**Machine learning**" context, i.e. black boxes, with less intuitive interpretation
- "**Big data**" context, i.e. easy to get proxies for protected/sensitive variables

y	urban	age	race
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

y	urban	age	zip	lastname	model	credit
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮

It is possible to predict the "race" based on non-protected variables, e.g. names and geolocation, see "Bayesian Improved Surname Geocoding (BISG)", Elliott et al. (2009), Imai and Khanna (2016)

- Problem of "**Indirect discrimination**", or "**statistical** / **proxy discrimination**"

Where could there be a problem?

Ratemaking is an issue, but also **underwriting**,

"**Redlining**", for loans, but also insurance, Kerner (1968)

"*use of a* **red line around the questionable areas on territorial maps** *centrally located in the Underwriting Division for ease of reference by all Underwriting personnel* [...] *mark off certain areas * * * to denote a lack of interest in business arising in these areas In New York these are called K.O. areas meaning* **knock-out areas***; in Boston they are called* **redline districts. Same thing – don't write the businesss.**"

to requests for information reveal clearly that business in certain geographic territories is restricted. For example, one underwriting guide states:

"An underwriter should be aware of the following situations in his territory:

1. The blighted areas.
2. The redevelopment operations.
3. Peculiar weather conditions which might make for a concentration of windstorm or hail losses.
4. The economic makeup of the area.
5. The nature of the industries in the area, etc.

"This knowledge can be gathered by drives through the area, by talking to and visiting agents, and by following local newspapers as to incidents of crimes and fires. A good way to keep this information available and up to date is by *the use of a red line* around the questionable areas on territorial maps centrally located in the Underwriting Division for ease of reference by all Underwriting personnel." (Italics added.)

A New York City insurance agent at our hearings put it more pointedly:

"[M]ost companies mark off certain areas * * * to denote a lack of interest in business arising in these areas In New York these are called K.O. areas—meaning knock-out areas; in Boston they are called redline districts. Same thing—don't write the business."

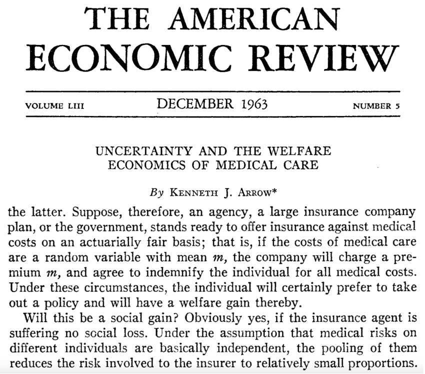
What is a "actuarial fairness"?

- "**Actuarial fairness**" ?

... "*on an* **actuarially fair** *basis; that is, if the costs of medical care are a random variable with mean m, the company will charge a premium m, and agree to indemnify the individual for all medical costs*," Arrow (1963).

"**actuarially fair premiums**" = "**expected losses**"

of the insured risk, see also Frezal and Barry (2020).



THE AMERICAN ECONOMIC REVIEW

VOLUME LIII — DECEMBER 1963 — NUMBER 5

UNCERTAINTY AND THE WELFARE ECONOMICS OF MEDICAL CARE

By KENNETH J. ARROW*

the latter. Suppose, therefore, an agency, a large insurance company plan, or the government, stands ready to offer insurance against medical costs on an actuarially fair basis; that is, if the costs of medical care are a random variable with mean m, the company will charge a premium m, and agree to indemnify the individual for all medical costs. Under these circumstances, the individual will certainly prefer to take out a policy and will have a welfare gain thereby.

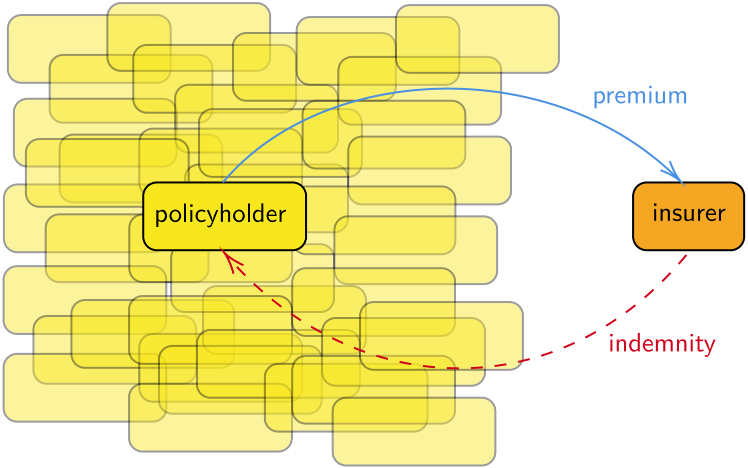
Will this be a social gain? Obviously yes, if the insurance agent is suffering no social loss. Under the assumption that medical risks on different individuals are basically independent, the pooling of them reduces the risk involved to the insurer to relatively small proportions.

"*governments must recognise that there is a difference between unfair discrimination and insurers differentiating prices according to risk*," Swiss Re (2015), cited in Meyers and Van Hoyweghen (2018)

So "actuarial fairness" has to do with "accuracy"?

Following Arrow (1963), "**actuarially fair premiums**" = "**expected losses**"

"*Insurance is the contribution of the many to the misfortune of the few*"



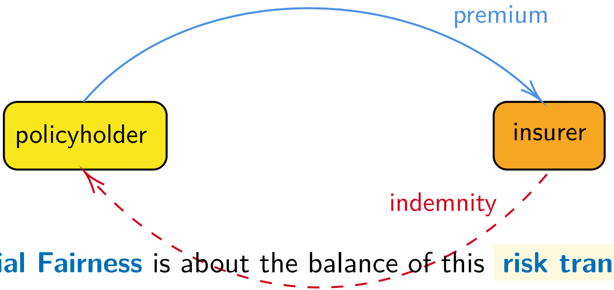
So "actuarial fairness" has to do with "accuracy"?

- There is no "**law of one price**" in insurance, see Froot et al. (1995)

→ with different models and different portfolio, we can have two different premiums

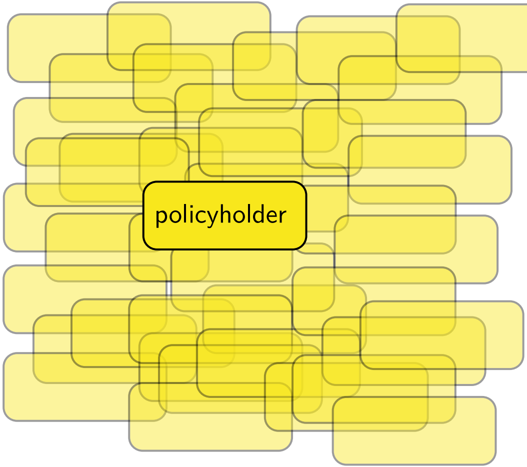
"*Insurance is the contribution of the many to the misfortune of the few*"

- Insurance is a **risk transfer** (from a policyholder to an insurance company)



- **Actuarial Fairness** is about the balance of this **risk transfer**

So "actuarial fairness" has to do with "accuracy"?



"Insurance is the contribution of the many to the misfortune of the few"

- As discussed in Charpentier (2025), insurance is also a **risk sharing** (among policyholders)
- **Fairness** (and **equity**) have to do with risk sharing and cross-subsidies within risk classes
- what is "**expected losses**"?

  $\mathbb{E}(Y)$ or $\mathbb{E}(Y \mid \boldsymbol{X})$?

  and what should be $\boldsymbol{X}$?

So “actuarial fairness” has to do with “accuracy”?

When y is binary, $y \in \{0,1\}$, hard to assess if $\widehat{y} = 12.2486\%$ is accurate or not...

“*If we are asked to find the probability holding for an individual future event, we must first incorporate the case in a suitable reference class*,” Reichenbach (1971)

“*When we speak of the ‘probability of death’, the exact meaning of this expression can be defined in the following way only.* **We must not think of an individual, but of a certain class as a whole**, *e.g., ‘all insured men forty-one years old living in a given country and not engaged in certain dangerous occupations’. A probability of death is attached to the class of men or to another class that can be defined in a similar way.* **The phrase ‘probability of death’, when it refers to a single person, has no meaning for us at all**,” von Mises (1928, 1939)

THE THEORY OF PROBABILITY

An Inquiry into the Logical and Mathematical Foundations of the Calculus of Probability

By HANS REICHENBACH

PROFESSOR OF PHILOSOPHY IN THE UNIVERSITY OF CALIFORNIA AT LOS ANGELES

UNIVERSITY OF CALIFORNIA PRESS

BERKELEY AND LOS ANGELES · 1949

§ 71. Attempts at a Single-Case Interpretation of Probability

After the discussion of the frequency meaning of probability, the investigation must turn to linguistic forms in which the concept of probability refers to an individual event. It is on this ground that the frequency interpretation has been questioned. Some logicians have argued that such usage is based on a different concept of probability, which is not reducible to frequencies. Is the existence of two disparate concepts of probability an inescapable consequence of the usage of language?

The first interpretation of the probability of single events is the *degree of expectation* with which an event is anticipated. The feeling of expectation certainly represents a psychological factor the existence of which is indisputable; it even shows degrees of intensity corresponding to the degrees of probability. Difficulty, however, arises from the fact that the degree of expectation varies from person to person and depends on more factors than the degree of the probability of the event to which the expectation refers. Apart from the probability of an event, emotional associations will influence the feeling of expectation. If it is a desirable event, as, for instance, the passing of an examination, optimistic persons will anticipate it with too-certain expectations, whereas pessimistic persons will think of it in terms of too-uncertain expectations.

So “actuarial fairness” has to do with “accuracy”?

As explained in Van Calster et al. (2019), ”*among patients with an estimated risk of 20%, we expect 20 in 100 to have or to develop the event*,”

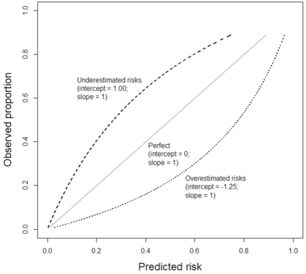
- If 40 out of 100 in this group are found to have the disease, the risk is underestimated
- If we observe that in this group, 10 out of 100 have the disease, we have overestimated the risk.

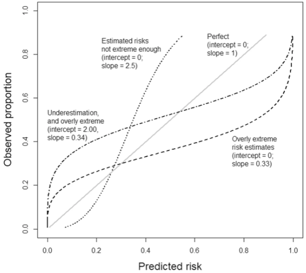
The prediction $\widehat{m}(\boldsymbol{X})$ of Y is a well-calibrated prediction if

$$\mathbb{E}\big[\underbrace{Y}\,\big|\,\widehat{Y}=\widehat{y}\big]=\widehat{y},\ \forall\widehat{y}$$

20 out of 100 (proportion $y=1$)

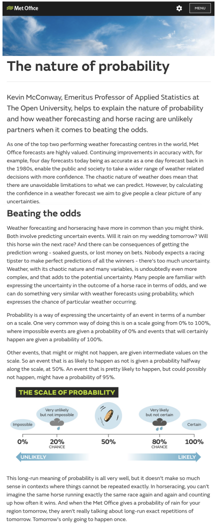
estimate risk $\widehat{y}=20\%$





So “actuarial fairness” has to do with “accuracy”?

“*Suppose the Met Office says that the probability of rain tomorrow in your region is 80%. They aren’t saying that it will rain in 80% of the land area of your region, and not rain in the other 20%. Nor are they saying it will rain for 80% of the time. What they are saying is there is an 80% chance of rain occurring at any one place in the region, such as in your garden.* [...] *A forecast of 80% chance of rain in your region should broadly mean that, on about 80% of days when the weather conditions are like tomorrow’s, you will experience rain where you are. [...] If it doesn’t rain in your garden tomorrow, then the 80% forecast wasn’t wrong, because it didn’t say rain was certain. But* **if you look at a long run of days, on which the Met Office said the probability of rain was 80%, you’d expect it to have rained on about 80% of them.**” McConway (2021)


Met Office

MENU

The nature of probability

Kevin McConway, Emeritus Professor of Applied Statistics at The Open University, helps to explain the nature of probability and how weather forecasting and horse racing are unlikely partners when it comes to beating the odds.

As one of the top two performing weather forecasting centres in the world, Met Office forecasts are highly valued. Continuing improvements in accuracy with, for example, four day forecasts today being as accurate as a one day forecast back in the 1980s, enable the public and society to take a wider range of weather related decisions with more confidence. The chaotic nature of weather does mean that there are unavoidable limitations to what we can predict. However, by calculating the confidence in a weather forecast we aim to give people a clear picture of any uncertainties.

Beating the odds

Weather forecasting and horseracing have more in common than you might think. Both involve predicting uncertain events. Will it rain on my wedding tomorrow? Will this horse win the next race? And there can be consequences of getting the prediction wrong - soaked guests, or lost money on bets. Nobody expects a racing tipster to make perfect predictions of all the winners - there's too much uncertainty. Weather, with its chaotic nature and many variables, is undoubtedly even more complex, and that adds to the potential uncertainty. Many people are familiar with expressing the uncertainty in the outcome of a horse race in terms of odds, and we can do something very similar with weather forecasts using probability, which expresses the chance of particular weather occurring.

Probability is a way of expressing the uncertainty of an event in terms of a number on a scale. One very common way of doing this is on a scale going from 0% to 100%, where impossible events are given a probability of 0% and events that will certainly happen are given a probability of 100%.

Other events, that might or might not happen, are given intermediate values on the scale. So an event that is as likely to happen as not is given a probability halfway along the scale, at 50%. An event that is pretty likely to happen, but could possibly not happen, might have a probability of 95%.

This long-run meaning of probability is all very well, but it doesn't make so much sense in contexts where things cannot be repeated exactly. In horseracing, you can't imagine the same horse running exactly the same race again and again and counting up how often it wins. And when the Met Office gives a probability of rain for your region tomorrow, they aren't really talking about long-run exact repetitions of tomorrow. Tomorrow's only going to happen once.

So “actuarial fairness” has to do with “accuracy”?

This concept goes beyond the simple issue of personalization (discussed in Barry and Charpentier (2020))

There are usually classical assumptions for “model” $\widehat{y}$,

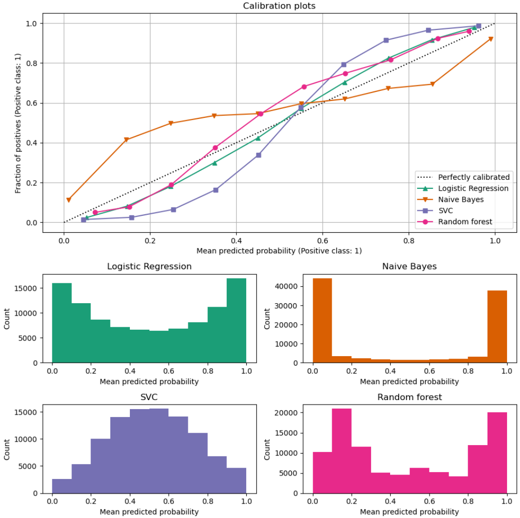
- (globally) well balanced, $\mathbb{E}[\underbrace{\widehat{Y}}_{\text{premium collected}}] = \mathbb{E}[\underbrace{Y}_{\text{losses paid}}]$
- (locally) well balanced, $\mathbb{E}[\widehat{Y} \mid \widehat{Y} = \widehat{y}] = \mathbb{E}[Y \mid \widehat{Y} = \widehat{y}] = \widehat{y}$, $\forall \widehat{y}$ (“calibration”)

From "accuracy" to "calibration"

Following Wilks (1990), define the calibration curve as

$$g : \begin{cases} [0,1] \to [0,1] \\ p \mapsto g(p) := \mathbb{E}[Y \mid \widehat{m}(\boldsymbol{X}) = p] \end{cases}$$

g is estimated using local regression of $\{(y_i, \widehat{m}(\boldsymbol{x}_i))\}$



“At the core of insurance business lies discrimination”.

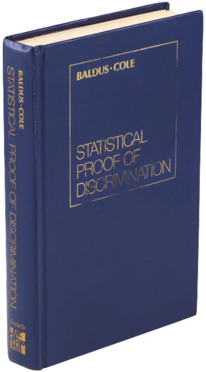
- ”*What is unique about insurance is that* **even statistical discrimination which by definition is absent of any malicious intentions, poses significant moral and legal challenges**. *Why? Because on the one hand, policy makers would like insurers to treat their insureds equally, without discriminating based on race, gender, age, or other characteristics, even if it makes statistical sense to discriminate* (...) *On the other hand,* **at the core of insurance business lies discrimination** *between risky and non-risky insureds. But riskiness often statistically correlates with the same characteristics policy makers would like to prohibit insurers from taking into account.* ” Avraham (2017)

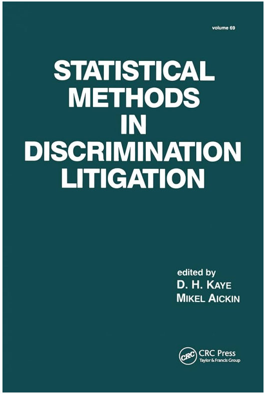
- “*Technology is neither good nor bad; nor is it neutral*,” Kranzberg (1986)

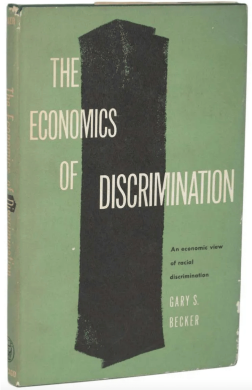
- “*Machine learning won’t give you anything like gender neutrality ‘for free’ that you didn’t explicitly ask for*,” Kearns and Roth (2019)

Quantifying discrimination, isn't it an old problem?

See Becker (1957) or Baldus and Cole (1980), among (many) others.







Several papers over the past 15 years revisited several notions and concepts.

Is there a (simple) way to quantify unfairness ?

- classical fairness concept are related to so called "**group fairness**", where we have a statistical (overall perspective),
- in some problems, we focus on discrimination in "continuous outcomes",
  - $\widehat{m}(\boldsymbol{x}_i, s_i) \in [0, 1]$ (score) that could also be denoted $\widehat{y}_i$
  - $\widehat{m}(\boldsymbol{x}_i, s_i) \in \mathbb{R}_+$ (premium) that could also be denoted $\widehat{y}_i$

  $\rightarrow$ classical in insurance modeling
- in some problems, we focus on discrimination in binary decisions $\widehat{y}_i \in \{0, 1\}$, usually obtained as
  - $\widehat{y}_i = \mathbf{1}(\widehat{m}(\boldsymbol{x}_i, s_i) > \text{threshold}) \in \{0, 1\}$ (class) that could also be denoted

  $\rightarrow$ classical in computer science

Several definitions of "fairness" or "non-discriminatory"

sensitive sensitive

demographic parity $\rightarrow$ $\mathbb{E}[\,\widehat{Y} \mid S = \mathrm{A}\,] \overset{?}{=} \mathbb{E}[\,\widehat{Y} \mid S = \mathrm{B}\,]$

score $\widehat{y}$

outcome y

equalized odds $\rightarrow$ $\mathbb{E}[\,\widehat{Y} \mid Y = y, S = \mathrm{A}\,] \overset{?}{=} \mathbb{E}[\,\widehat{Y} \mid Y = y, S = \mathrm{B}\,], \forall y$

score $\widehat{y}$

outcome y

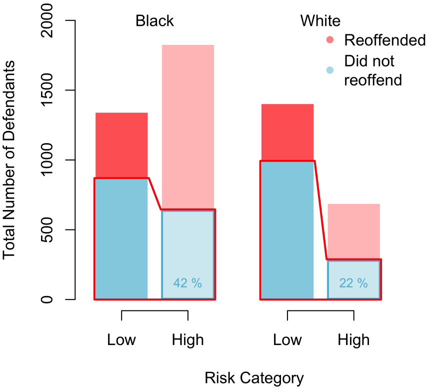
calibration $\rightarrow$ $\mathbb{E}[\,Y \mid \widehat{Y} = u, S = \mathrm{A}\,] \overset{?}{=} \mathbb{E}[\,Y \mid \widehat{Y} = u, S = \mathrm{B}\,], \forall u$

score $\widehat{y}$

Isn't it a problem to have several definitions?

From Feller et al. (2016),

- for White people, among those who did not re-offend (y), 22% were wrongly classified ($\widehat{y}$),
- for Black people, among those who did not re-offend, 42% were wrongly classified,
- **Problem**, since 42% $\gg$ 22%

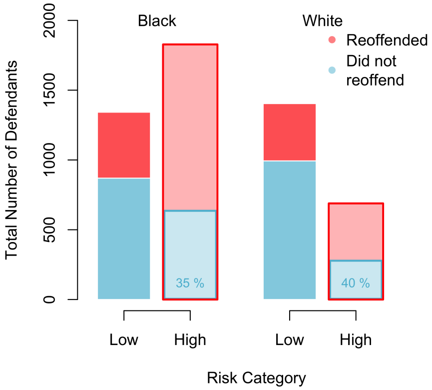


$$\mathbb{P}[\widehat{Y} = \text{high} \mid Y = \text{no}, S = \text{black}] = 42\% \stackrel{?}{=} \mathbb{P}[\widehat{Y} = \text{high} \mid Y = \text{no}, S = \text{white}] = 22\%,$$

Isn't it a problem to have several definitions?

From Dieterich et al. (2016),

- for White people, among those who were classified as high risk ($\widehat{y}$), 40% did not re-offend (y),
- for Black people, among those who were classified as high risk ($\widehat{y}$), 35% did not re-offend (y),
- **No problem**, since $35 \approx 40\%$



$$\mathbb{P}[Y = \text{no} \,|\, \widehat{Y} = \text{high}, S = \text{black}] = 35\% \overset{?}{=} \mathbb{P}[Y = \text{no} \,|\, \widehat{Y} = \text{high}, S = \text{white}] = 40\%.$$

Is it always possible to have a sensitive-free model (with respect to ...)?

For **decisions** ($\widehat{y} \in \{0,1\}$, e.g., "obtain a loan"),

$$\text{demographic parity} \rightarrow \mathbb{P}[\underbrace{\widehat{Y}=1}_{\text{decision } \widehat{y}} \mid S=A] \overset{?}{=} \mathbb{P}[\underbrace{\widehat{Y}=1}_{\text{decision } \widehat{y}} \mid S=B]$$

those decisions are usually based on **scores**, and **thresholds**

$$\text{demographic parity} \rightarrow \mathbb{E}[\underbrace{\widehat{m}(\boldsymbol{X},S)}_{\text{score } \widehat{m}} > t \mid S=A] \overset{?}{=} \mathbb{E}[\underbrace{\widehat{m}(\boldsymbol{X},S)}_{\text{score } \widehat{m}} > t \mid S=B]$$

One can achieve demographic parity, simply selecting **different thresholds**

$$\text{demographic parity} \rightarrow \mathbb{E}[\widehat{m}(\boldsymbol{X},S) > t_A \mid S=A] \overset{?}{=} \mathbb{E}[\widehat{m}(\boldsymbol{X},S) > t_B \mid S=B]$$

(with that strategy, usually impossible to achieve equalized odds)

Is it always possible to have a sensitive-free model (with respect to ...)?

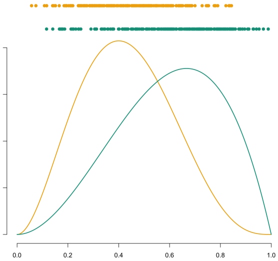
For **decisions** ($\widehat{y} \in \{0, 1\}$, e.g., "obtain a loan"), we considered

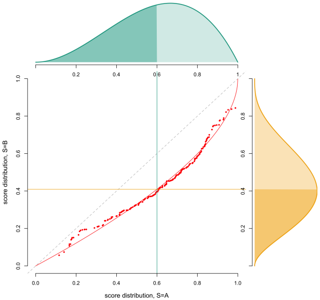
$$\text{demographic parity} \ \rightarrow \ \mathbb{E}[\,\widehat{Y}\,|\,S = \mathsf{A}\,] \overset{?}{=} \mathbb{E}[\,\widehat{Y}\,|\,S = \mathsf{B}\,]$$

and we can consider the analogous for **scores** (possibly used to assess premiums),

$$\text{demographic parity} \ \rightarrow \ \mathbb{E}[\,\widehat{m}(\boldsymbol{X}, S)\,|\,S = \mathsf{A}\,] \overset{?}{=} \mathbb{E}[\,\widehat{m}(\boldsymbol{X}, S)\,|\,S = \mathsf{B}\,]$$

score $\widehat{y}$





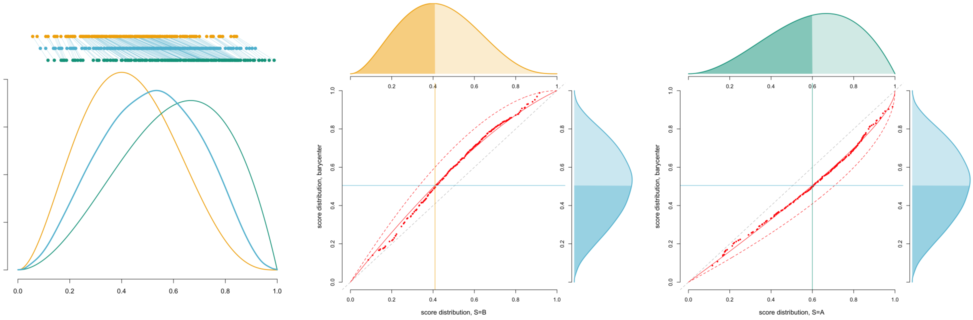
- individual in group A with a score $\widehat{y}(\mathsf{A}) = 60\%$ corresponding to quantile α (here 0.5)
- in group B, the same quantile α corresponds to $\widehat{y}(\mathsf{B}) = 40\%$

Is it always possible to have a sensitive-free model (with respect to ...)?

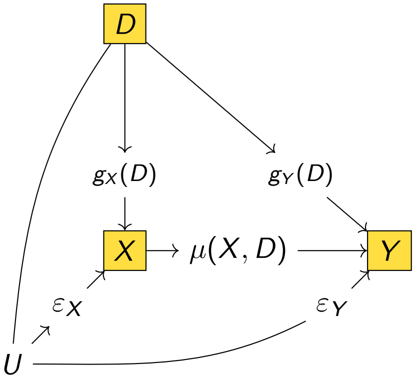
- To get a fair model (**neutral with respect to s**), consider an average between the two models,

score in group A with quantile α score in group B with quantile α

$$\widehat{y}^\star = \mathbb{P}[S=\mathrm{A}]\cdot\widehat{y}(\mathrm{A}) + \mathbb{P}[S=\mathrm{B}]\cdot\widehat{y}(\mathrm{B})$$



A spectrum of fair premiums with Causal Graphs



As in Côté et al. (2024), consider a

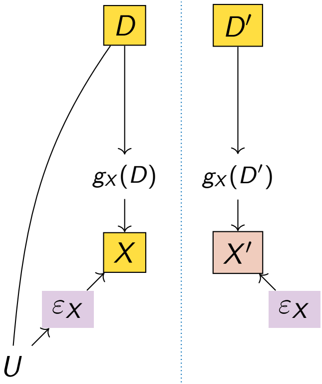
structural causal model | Markov property

$$\begin{cases} X = \psi_X(D, \varepsilon_X) = g_X(D) + \varepsilon_X \\ Y = \psi_Y(D, X, \varepsilon_Y) = g_Y(D) + \mu(X) + \varepsilon_Y \end{cases}$$

with $\varepsilon_X \perp\!\!\!\perp D$ and $\varepsilon_Y \perp\!\!\!\perp D$.

- abduction use the evidence (X, D) to determine the value of the noise ε_X
- prediction use the estimated noise ε_X to compute the counterfactual of X as $\psi_X(D', \varepsilon_X)$

Optimal Transport for Counterfactual, Gaussian Additive Case



Pearl (2009) suggested a

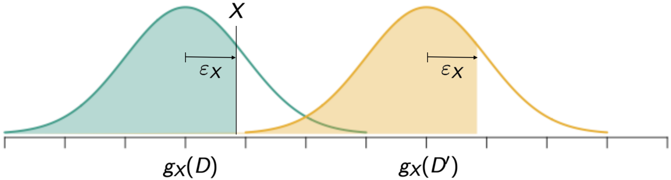
twin network representation of the counterfactual

- **abduction** use the evidence (X, D):

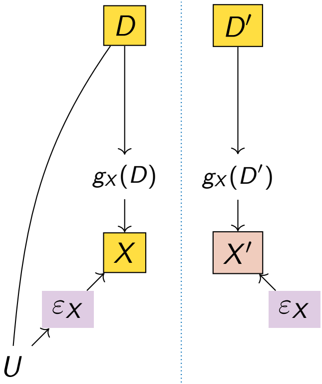
  $$\varepsilon_X = X - g_X(D)$$

- **prediction** use the same estimated noise ε_X to compute the counterfactual of X

  $$X' = g_X(D') + \varepsilon_X$$



Optimal Transport for Counterfactual, General Case



Charpentier et al. (2023a) and then Fernandes Machado et al. (2025a,b) extended this using

optimal transport on causal graphs

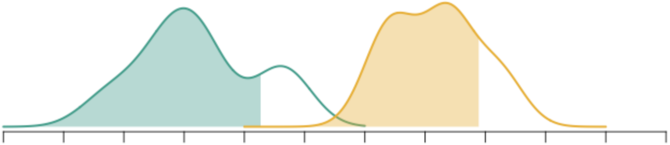
- $F_A(\cdot)$ cdf of $X \mid D = A$, $F_A(x) = \mathbb{P}(X \leq x \mid D = A)$

  abduction $u = F_A(X)$ probability level in group A

- $F_B(\cdot)$ cdf of $X \mid D = B$,
  $F_B(x) = \mathbb{P}(X \leq x \mid D = B)$

  prediction $F_B^{-1}(u)$ quantile of level u in group B

counterfactual is $X' = F_B^{-1}(F_A(X))$



A spectrum of fair premiums, Côté et al. (2024, 2025)

$$\boxed{\mu^B(x,d)=\mathbb{E}(Y\mid X=x,D=d)\;\leftarrow\;\text{best estimate}}$$

$$\boxed{\mu^U(x)=\mathbb{E}(Y\mid X=x)\;\leftarrow\;\text{unaware}}$$

$$\boxed{\mu^A(x)=\mathbb{E}_D(Y\mid X=x,D)=\mathbb{E}_D(\mu^B(x,D))\;\leftarrow\;\text{aware}}$$

$$\mu^A(x)=\mathbb{P}(D=A)\mu^B(x,d=A)+\mathbb{P}(D=B)\mu^B(x,d=B)\text{ if }D\in\{A,B\}$$

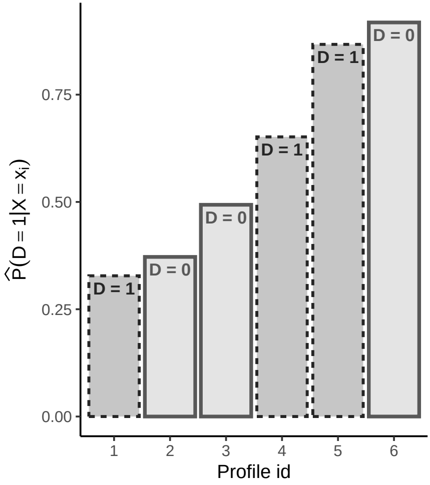
$$\boxed{\mu^C(x,d)=\mathbb{E}(Y\mid\varepsilon_X=x-\Pi_d(x))\;\leftarrow\;\text{corrective}}$$

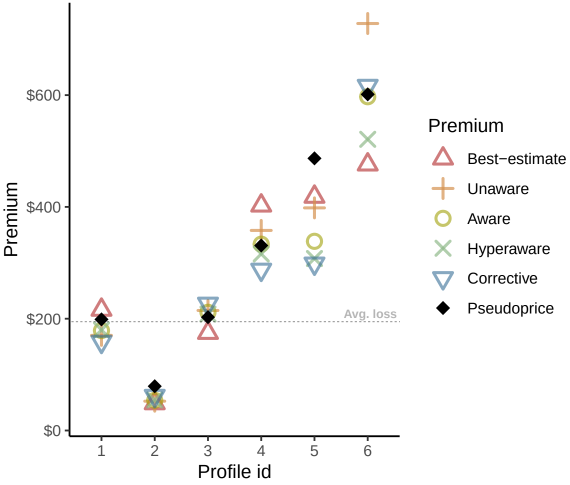
$$\mu^C(x,d=A)=\mathbb{P}(D=A)\mu^B(x,d=A)+\mathbb{P}(D=B)\cdot F_B\circ F_A^{-1}(\mu^B(x,d=A))$$

$$\boxed{\mu^H(x)=\mathbb{E}_D(Y\mid X=x,D)=\mathbb{E}_D(\mu^C(x,D))\;\leftarrow\;\text{hyperaware}}$$

A spectrum of fair premiums, Côté et al. (2024, 2025)

(on a real insurance portfolio, we compared the five premiums, and the “real one”)





"In order to treat some persons equally, we must treat them differently"

- Supreme Court Justice Harry Blackmun stated, in 1978,

  "*In order to get beyond racism, we must first take account of race. There is no other way. And* **in order to treat some persons equally, we must treat them differently**," Knowlton (1978), cited in Lippert-Rasmussen (2020)

- To counteract **disparate impact**, intentional **disparate treatment** is necessary
- See philosophical discussions about **affirmative action**, e.g., Rubenfeld (1997); Pojman (1998); Anderson (2004)
- In 2007, John G. Roberts of the U.S. Supreme Court submits

  "*The way to stop discrimination on the basis of race is to* **stop discriminating on the basis of race**," Sabbagh (2007) and Turner (2015)

- corresponds to the "colorblind" approach
- Rejects any form of **disparate treatment**, even for corrective purposes, and reproduction of historical inequalities will lead to **disparate impact**

"Neutral with respect to some sensitive attribute?"

What does "**neutral with respect to** s" really means ?

We have seen that accuracy was assessed with respect to data in the portfolio,

$$\boxed{\overline{y}} = \underset{\gamma\in\mathbb{R}}{\text{argmin}}\left\{\boxed{\sum_{i=1}^{n}(y_i-\gamma)^2}\right\} \text{ or } \boxed{\mathbb{E}[Y]} = \underset{\gamma\in\mathbb{R}}{\text{argmin}}\left\{\boxed{\sum_{y}(y-\gamma)^2\mathbb{P}[Y=y]}\right\}$$

based on observations from the insurer's portfolio. Technically, should we consider

- expected values / probabilities / independence properties based on $\mathbb{P}$ (portfolio)
- expected values / probabilities / independence properties based on $\mathbb{Q}$ (market)

(ongoing work *Why portfolio-specific fairness should fail to extend market-wide: Selection bias in insurance* with M.P. Côté & O. Côté)

Should we ask for neutrality "in the portfolio" or for some "targeted population" ?

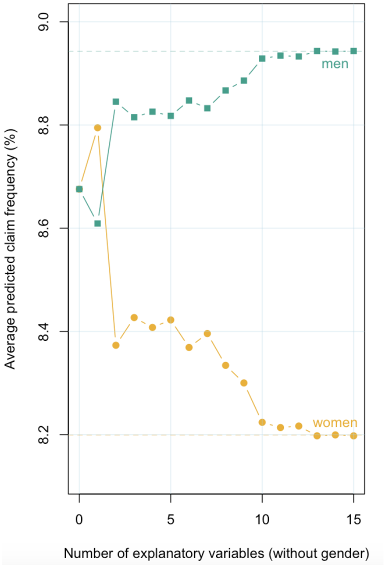
Discrimination in the data, or in the model?

On a French motor dataset, average claim frequencies are 8.94% (men) and 8.20% (women).

Consider some logistic regression to estimate annual claim frequency, on k explanatory variables excluding gender.

	men	women
$k = 0$	8.68%	8.68%
$k = 2$	8.85%	8.37%
$k = 8$	8.87%	8.33%
$k = 15$	8.94%	8.20%
empirical	8.94%	8.20%

Models simply tend to reproduce what was observed in the data (see "**is-ought**" problem, in Hume (1739)).



Discrimination in the data, or in the model?

David Hume's "**is-ought**" problem, in Hume (1739)

what **is** observed, what is **statistically normal**

$$\pi(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}}[Y|\boldsymbol{X}=\boldsymbol{x}] \text{ where } \mathbb{P} \text{ is the historical probability}$$

$\neq$ what **should be**, what we expect from an **ethical norm**

$$\pi(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}^\star}[Y|\boldsymbol{X}=\boldsymbol{x}] \text{ where } \mathbb{P}^\star \text{ is some "fair" probability}$$

"*keep in mind that machine learning can only be used to memorize patterns that are present in your training data. You can only recognize what you've seen before. Using machine learning trained on past data to predict the future is making the assumption that the future will behave like the past*," Chollet (2021)

Classical **clausula rebus sic stantibus** ("with things thus standing") in predictive modeling (statistics and machine learning)

Discrimination in the data, or in the model?

- change the training data to de-bias (through weights) : **pre-processing**

if we can draw i.i.d. copies of a random variable X_i's, under probability $\mathbb{P}$, then

$$\frac{1}{n}\sum_{i=1}^{n} h(x_i) \to \mathbb{E}_{\mathbb{P}}[h(X)], \text{ as } n \to \infty \text{ "\textbf{law of large numbers}"}$$

but if we want to reach $\mathbb{E}_{\mathbb{Q}}[h(X)]$, consider

$$\frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{\mathrm{d}\mathbb{Q}(x_i)}{\mathrm{d}\mathbb{P}(x_i)}}_{\text{weight } \omega_i} h(x_i) \to \mathbb{E}_{\mathbb{Q}}[h(X)], \text{ as } n \to \infty.$$

- keep the biases data, but distort the outcome : **post-processing**
- add a fairness constraint (penalty) in the optimization problem : **in-processing**
  as classical adversarial techniques, Grari et al. (2021)

Discrimination, with different perspectives

- Regulatory perspective, "**group fairness**" (discussed previously)
- Policyholders perspective, "**individual fairness**"

A decision satisfies individual fairness if "*had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same*."

- also named "**counterfactual fairness**" in Kusner et al. (2017), and should be related to classical causal inference problem, (conditional) average treatment effect (the "treatement" being the sensitive attribute),

  "*other things being equal*" ?**ceteris paribus** ? See "revolving variable" in Kilbertus et al. (2017). Consider a men ($s = \mathrm{A}$) with height $x = 6'3$ (or 190 cm). If that person had been a women ($s = \mathrm{B}$) would she have height $x = 6'3$?

  (hint: no, consider similar quantiles, as discussed previously, see Charpentier et al. (2023a))

What if we neither observe nor collect sensitive personal information (s) ?

September 27, 2023, the Colorado Division of Insurance exposed a new proposed regulation entitled Concerning Quantitative Testing of External Consumer Data and Information Sources, Algorithms, and Predictive Models Used for Life Insurance Underwriting for Unfairly Discriminatory Outcomes. Use of **BIFSG** (Bayesian Improved First Name Surname and Geocoding), from Elliott et al. (2009).

Consider 12 people living near Atlanta, GA (Fulton & Gwinnett counties),

	last	first	county	city	zipcode	whi	bla	his	asi
2	RADLEY	OLIVIA	Fulton	Fairburn	30213	14	83	1	0
3	BOORSE	KEISHA	Fulton	Atlanta	30331	97	0	3	0
4	MAZ	SAVANNAH	Gwinnett	Norcross	30093	5	6	76	13
5	GAULE	NATASHIA	Gwinnett	Snellville	30078	67	19	14	0
6	MCMELLEN	ISMAEL	Gwinnett	Lilburn	30047	73	15	6	3
7	WASHINGTON	BRYN	Gwinnett	Norcross	30093	0	95	3	0

(ongoing *Predicting Unobserved Multi-Class sensitive Attributes : Enhancing Calibration with Nested Dichotomies for Fairness* with A.M. Patrón Piñerez, A. Fernandes Machado, & E. Gallic)

Can we use aggregate data related to sensitive information ($\overline{s}$) ?

Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

P. J. Bickel, E. A. Hammel, J. W. O'Connell

Data and Assumptions

Tests of Aggregate Data

Some Underlying Dependencies

Table 1. Decisions on applications to Graduate Division for fall 1973, by sex of applicant—naive aggregation. Expected frequencies are calculated from the marginal totals of the observed frequencies under the assumptions (1 and 2) given in the text. $N = 12{,}763$, $\chi^2 = 110.8$, d.f. = 1, $P = 0$ (18).

Applicants	Observed Admit	Observed Deny	Expected Admit	Expected Deny	Difference Admit	Difference Deny
Men	3738	4704	3460.7	4981.3	277.3	− 277.3
Women	1494	2827	1771.3	2549.7	− 277.3	277.3

Table 2. Admissions data by sex of applicant for two hypothetical departments. For total, $\chi^2 = 5.71$, d.f. = 1, $P = 0.19$ (one-tailed).

Applicants	Observed Admit	Observed Deny	Expected Admit	Expected Deny	Difference Admit	Difference Deny
Department of machismatics						
Men	200	200	200	200	0	0
Women	100	100	100	100	0	0
Department of social warfare						
Men	50	100	50	100	0	0
Women	150	300	150	300	0	0
Totals						
Men	250	300	229.2	320.8	20.8	− 20.8
Women	250	400	270.8	379.2	− 20.8	20.8

Fig. 1. Proportion of applicants that are women plotted against proportion of applicants admitted, in 85 departments. Size of box indicates relative number of applicants to the department.

Disaggregation

from Bickel et al. (1975), discussed as an illustration of "**Simpson's paradox**"

Can we use aggregate data related to sensitive information ($\overline{s}$) ?

	Total	Men	Women	Proportions
Total	5233/12763 $\sim$ 41%	3714/8442 $\sim$ **44%**	1512/4321 $\sim$ 35%	66%-34%
Top 6	1745/4526 $\sim$ 39%	1198/2691 $\sim$ **45%**	557/1835 $\sim$ 30%	59%-41%
A	597/933 $\sim$ 64%	512/825 $\sim$ 62%	89/108 $\sim$ **82%**	88%-12%
B	369/585 $\sim$ 63%	353/560 $\sim$ 63%	17/ 25 $\sim$ **68%**	96%- 4%
C	321/918 $\sim$ 35%	120/325 $\sim$ **37%**	202/593 $\sim$ 34%	35%-65%
D	269/792 $\sim$ 34%	138/417 $\sim$ 33%	131/375 $\sim$ **35%**	53%-47%
E	146/584 $\sim$ 25%	53/191 $\sim$ **28%**	94/393 $\sim$ 24%	33%-67%
F	43/714 $\sim$ 6%	22/373 $\sim$ 6%	24/341 $\sim$ **7%**	52%-48%

Data from Bickel et al. (1975). Formalized as follows: S is the (binary) genre, $\widehat{Y}$ the admission decision, and X the program (category),

Can we use aggregate data related to sensitive information ($\overline{s}$) ?

$$\underbrace{\mathbb{P}[\widehat{Y}=\text{yes} \mid \overbrace{S=\text{men}}^{\text{sensitive}}]}_{\text{overall admission}} \geq \underbrace{\mathbb{P}[\widehat{Y}=\text{yes} \mid \overbrace{S=\text{women}}^{\text{sensitive}}]}_{\text{overall admission}}$$

$$\mathbb{P}[\widehat{Y}=\text{yes} \mid \underbrace{X=x}_{\text{conditional on program}}, S=\text{men}] \leq \mathbb{P}[\widehat{Y}=\text{yes} \mid \underbrace{X=x}_{\text{conditional on program}}, S=\text{women}], \ \forall x.$$

"*the bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seems quite fair on the whole, but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects*," Bickel et al. (1975)

What if we collect s but we miss an important predictor (x) ?

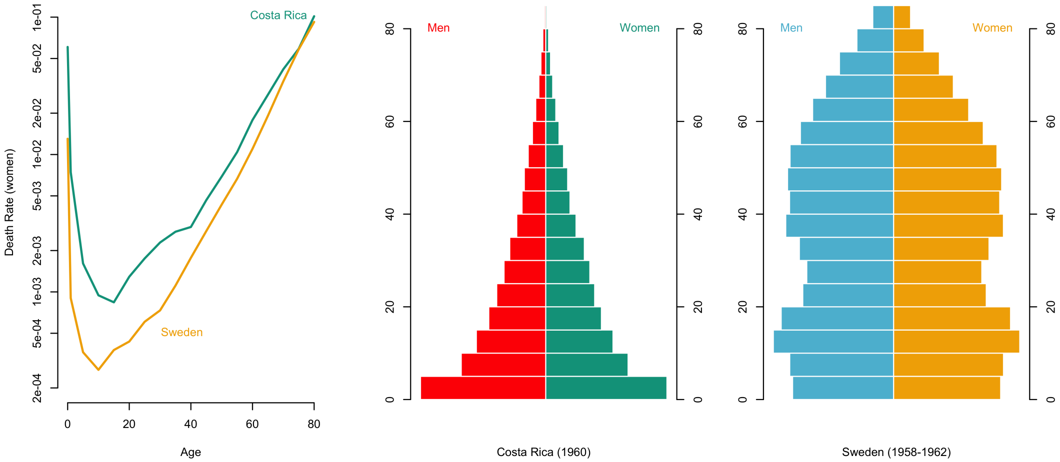
Simpson's paradox can also be seen a an **omitted variable bias** problem,

$$\begin{cases} y_i = \beta_0 + \boldsymbol{x}_1^\top \boldsymbol{\beta}_1 + \boldsymbol{x}_2^\top \boldsymbol{\beta}_2 + \varepsilon_i \text{ true model} \\ y_i = b_0 + \boldsymbol{x}_1^\top \boldsymbol{b}_1 + \eta_i \text{ estimated models} \end{cases}$$

$$\begin{aligned} \widehat{\boldsymbol{b}}_1 &= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{y} \\ &= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top [\boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}] \\ &= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_1 \boldsymbol{\beta}_1 + (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_2 \boldsymbol{\beta}_2 + (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{\varepsilon} \\ &= \boldsymbol{\beta}_1 + \underbrace{(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_2 \boldsymbol{\beta}_2}_{\boldsymbol{\beta}_{12}} + \underbrace{(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{\varepsilon}}_{\nu_i}, \end{aligned}$$

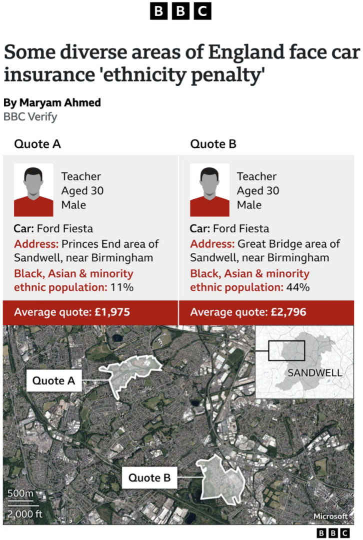
so that $\mathbb{E}[\widehat{\boldsymbol{b}}_1] = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_{12} \neq \boldsymbol{\beta}_1$.

What if we collect s but we miss an important predictor (x) ?



Overall mortality rate for women, 8.12‰ in Costa Rica, against 9.29‰ in Sweden.

Disentangling correlations


BBC

Some diverse areas of England face car insurance 'ethnicity penalty'

By Maryam Ahmed
BBC Verify

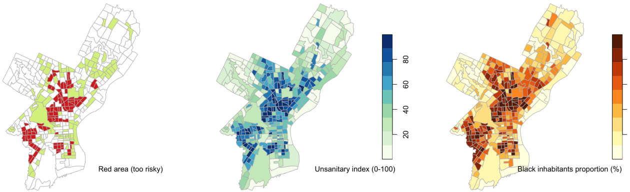
Quote A

Teacher
Aged 30
Male
Car: Ford Fiesta
Address: Princes End area of Sandwell, near Birmingham
Black, Asian & minority ethnic population: 11%
Average quote: £1,975

Quote B

Teacher
Aged 30
Male
Car: Ford Fiesta
Address: Great Bridge area of Sandwell, near Birmingham
Black, Asian & minority ethnic population: 44%
Average quote: £2,796

See some diverse areas of England face car insurance 'ethnicity penalty' (remove from the BBC website since)



y, x and s can easily be correlated variables

spurious correlations problem ?

Need to use causal models to avoid indirect discrimination

Multiple sensitive attributes, "robbing Peter to pay Paul"?

$$\mathbb{E}[\,\widehat{m}(\boldsymbol{X},S_1,S_2)\,|\,\underbrace{S_1=A\,]\;\neq\;\mathbb{E}[\,\widehat{m}(\boldsymbol{X},S_1,S_2)\,|\,S_1=B}_{\text{sensitive attribute 1}}\,]$$

$$\mathbb{E}[\,\widehat{m}(\boldsymbol{X},S_1,S_2)\,|\,\underbrace{S_2=C\,]\;\approx\;\mathbb{E}[\,\widehat{m}(\boldsymbol{X},S_1,S_2)\,|\,S_2=D}_{\text{sensitive attribute 2}}\,]$$

Distort model $\widehat{m}$ to achieve fairness with respect to S_1 $\longrightarrow$ model $\widetilde{m}$

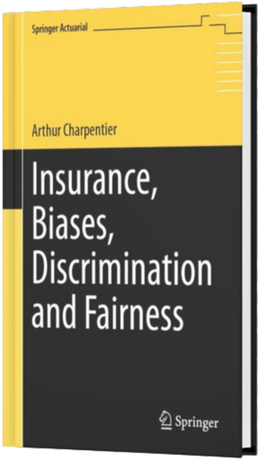
$$\mathbb{E}[\,\widetilde{m}(\boldsymbol{X},S_1,S_2)\,|\,\underbrace{S_1=A\,]\;=\;\mathbb{E}[\,\widetilde{m}(\boldsymbol{X},S_1,S_2)\,|\,S_1=B}_{\text{sensitive attribute 1}}\,]$$

$$\mathbb{E}[\,\widetilde{m}(\boldsymbol{X},S_1,S_2)\,|\,\underbrace{S_2=C\,]\;\neq\;\mathbb{E}[\,\widetilde{m}(\boldsymbol{X},S_1,S_2)\,|\,S_2=D}_{\text{sensitive attribute 2}}\,]$$

Conclusion (?)

- dealing with discrimination in insurance is tricky since actuarial pricing is deeply related to the idea of focusing on groups, and not individuals
- if we do not address properly those questions, there is no way we can get fair models
- not collecting and not using protected attributes is clearly not a good strategy
- there are still important questions that should be addressed by regulators, that should provide guidelines

To go further, Charpentier (2024) **Insurance, Biases, Discrimination and Fairness. Springer**.



References

Anderson, T. H. (2004). *The pursuit of fairness: A history of affirmative action*. Oxford University Press.

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *The American Economic Review*, 53(5):941–973.

Avraham, R. (2017). Discrimination and insurance. In Lippert-Rasmussen, K., editor, *Handbook of the Ethics of Discrimination*, pages 335–347. Routledge.

Bailey, R. A. and Simon, L. J. (1960). Two studies in automobile insurance ratemaking. *ASTIN Bulletin: The Journal of the IAA*, 1(4):192–217.

Baldus, D. C. and Cole, J. W. (1980). *Statistical proof of discrimination*. Mcgraw-Hill.

Barry, L. and Charpentier, A. (2020). Personalization as a promise: Can big data change the practice of insurance? *Big Data & Society*, 7(1):2053951720935143.

Becker, G. S. (1957). *The economics of discrimination*. University of Chicago press.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175):398–404.

Board, A. S. (2005). Risk classification (for all practice areas). *Actuarial Standard of Practice (ASOP)*, 12.

References

Brilmayer, L., Hekeler, R. W., Laycock, D., and Sullivan, T. A. (1979). Sex discrimination in employer-sponsored insurance plans: A legal and demographic analysis. *University of Chicago Law Review*, 47:505.

Charpentier, A. (2024). *Insurance: biases, discrimination and fairness*. Springer Verlag.

Charpentier, A. (2025). Les paradoxes de la segmentation et de la discrimination en assurance. *Risques*, 142.

Charpentier, A., Flachaire, E., and Gallic, E. (2023a). Causal inference with optimal transport. In Thach, N. N., Kreinovich, V., Ha, D. T., and Trung, N. D., editors, *Optimal Transport Statistics for Economics and Related Topics*. Springer Verlag.

Charpentier, A., Hu, F., and Ratz, P. (2023b). Mitigating discrimination in insurance with wasserstein barycenters. bias. In *3rd Workshop on Bias and Fairness in AI, International Workshop of ECML PKDD*.

Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.

Côté, O., Côté, M.-P., and Charpentier, A. (2024). A fair price to pay: exploiting causal graphs for fairness in insurance. *Journal of Risk and Insurance*.

Côté, O., Côté, M.-P., and Charpentier, A. (2025). Selection bias in insurance: Why portfolio-specific fairness fails to extend market-wide. *SSRN*.

References

Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(7.4):1.

Elliott, M. N., Morrison, P. A., Fremont, A., McCaffrey, D. F., Pantoja, P., and Lurie, N. (2009). Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9:69–83.

Feller, A., Pierson, E., Corbett-Davies, S., and Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *The Washington Post*, October 17.

Fernandes Machado, A., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024a). From uncertainty to precision: Enhancing binary classifier performance through calibration. *arXiv preprint arXiv:2402.07790*.

Fernandes Machado, A., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024b). Post-calibration techniques: Balancing calibration and score distribution alignment. *NeurIPS BDU Workshop*.

Fernandes Machado, A., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024c). Probabilistic scores of classifiers, calibration is not enough. *arXiv preprint arXiv:2408.03421*.

Fernandes Machado, A., Charpentier, A., and Gallic, E. (2025a). Optimal transport on categorical data for counterfactuals using compositional data and dirichlet transport. *arXiv*, 2501.15549.

References

Fernandes Machado, A., Charpentier, A., and Gallic, E. (2025b). Sequential conditional transport on probabilistic graphs for interpretable counterfactual fairness. *39th Annual AAAI Conference on Artificial Intelligence*.

Finger, R. J. (2006). Risk classification. In Bass, I., Basson, S., Bashline, D., Chanzit, L., Gillam, W., and Lotkowski, E., editors, *Foundations of Casualty Actuarial Science*, pages 287–341. Casualty Actuarial Society.

Frezal, S. and Barry, L. (2020). Fairness in uncertainty: Some limits and misinterpretations of actuarial fairness. *Journal of Business Ethics*, 167:127–136.

Froot, K. A., Kim, M., and Rogoff, K. S. (1995). The law of one price over 700 years. *National Bureau of Economic Research Cambridge*.

Glenn, B. J. (2000). The shifting rhetoric of insurance denial. *Law and Society Review*, pages 779–808.

Glenn, B. J. (2003). Postmodernism: the basis of insurance. *Risk Management and Insurance Review*, 6(2):131–143.

Grari, V., Lamprier, S., and Detyniecki, M. (2021). Fairness without the sensitive attribute via causal variational autoencoder.

References

Hu, F., Ratz, P., and Charpentier, A. (2023). Fairness in multi-task learning via wasserstein barycenters. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases – ECML PKDD*.

Hu, F., Ratz, P., and Charpentier, A. (2024). A sequentially fair mechanism for multiple sensitive attributes. *Annual AAAI Conference on Artificial Intelligence*.

Hume, D. (1739). *A Treatise of Human Nature*. Cambridge University Press Archive.

Imai, K. and Khanna, K. (2016). Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis*, 24(2):263–272.

Kearns, M. and Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Kerner, O. (1968). *Report of The National Advisory Commission on Civil Disorder*. Bantam Books.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30.

Knowlton, R. E. (1978). Regents of the university of california v. bakke. *Arkansas Law Review*, 32:499.

Kranzberg, M. (1986). Technology and history:" kranzberg's laws". *Technology and culture*, 27(3):544–560.

References

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076.

Lippert-Rasmussen, K. (2020). *Making sense of affirmative action*. Oxford University Press.

Merriam-Webster (2022). *Dictionary*.

Meyers, G. and Van Hoyweghen, I. (2018). Enacting actuarial fairness in insurance: From fair discrimination to behaviour-based fairness. *Science as Culture*, 27(4):413–438.

Mowbray, A. (1921). Classification of risks as the basis of insurance rate making with special reference to workmen's compensation. *Proceedings of the Casualty Actuarial Society*.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pojman, L. P. (1998). The case against affirmative action. *International Journal of Applied Philosophy*, 12(1):97–115.

Reichenbach, H. (1971). *The theory of probability*. University of California Press.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Rubenfeld, J. (1997). Affirmative action. *Yale Law Journal*, 107:427.

References

Sabbagh, D. (2007). *Equality and transparency: A strategic perspective on affirmative action in American law*. Springer.

Schauer, F. (2006). *Profiles, probabilities, and stereotypes*. Harvard University Press.

Swiss Re (2015). Life insurance risk selection: Required differentiation or unfair discrimination? *Sigma*.

Turner, R. (2015). The way to stop discrimination on the basis of race. *Stanford Journal of Civil Rights & Civil Liberties*, 11:45.

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., and Steyerberg, E. W. (2019). Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):1–7.

von Mises, R. (1928). *Wahrscheinlichkeit Statistik und Wahrheit*. Springer.

von Mises, R. (1939). *Probability, statistics and truth*. Macmillan.

Wilks, D. S. (1990). On the combination of forecast probabilities for consecutive precipitation periods. *Weather and Forecasting*, 5(4):640–650.