Demystify fairness and discrimination in insurance, and avoid some pitfalls

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

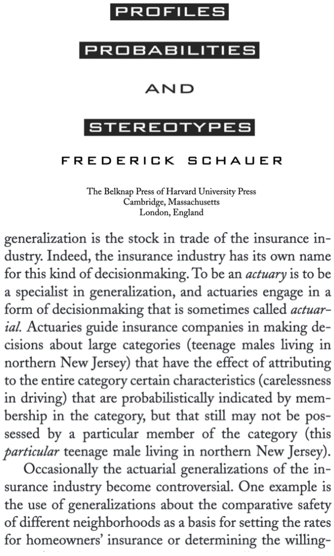
Financial Conduct Authority (FCA), London, November 2024

FOUNDATION FOR SCIENCE

What is an “actuary”?

- “actuarial” ?

“*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).



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

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

- Poisson regression (frequency) on categories, or not

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

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

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

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).

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
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮

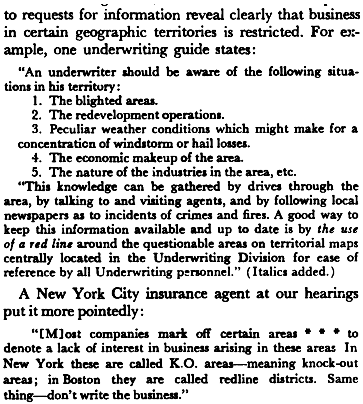
It is possible to predict the "race" based on non-protected variables, e.g. names and geolocation, see "Bayesian Improved Surname Geocoding (BISG)", Elliott et al. (2009), Imai and Khanna (2016)

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.”

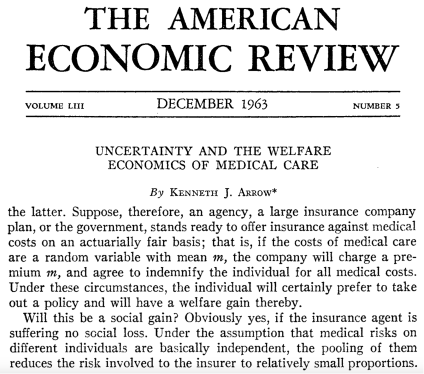
What is a “actuarial fairness”?

- “**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

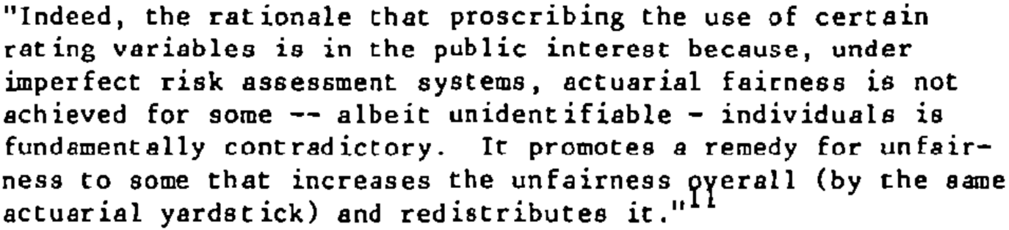
By KENNETH J. ARROW*

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)

What is a "actuarial fairness"?



"Indeed, the rationale that proscribing the use of certain rating variables is in the public interest because, under imperfect risk assessment systems, actuarial fairness is not achieved for some -- albeit unidentifiable - individuals is fundamentally contradictory. It promotes a remedy for unfairness to some that increases the unfairness overall (by the same actuarial yardstick) and redistributes it."[11]

"*Indeed, the rationale that proscribing the use of certain rating variables is in the public interest because, under imperfect risk assessment systems, actuarial fairness is not achieved for some – albeit unidentifiable - individuals is fundamentally contradictory. It promotes a remedy for unfairness to some that increases the unfairness overall (by the same actuarial yardstick) and redistributes it*," Casey et al. (1976), cited in Walters (1981)

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

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

- ▶ but still, there is no “**law of one price**” in insurance, Froot et al. (1995)

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

- ▶ estimating “**expected losses**” means maximizing “**accuracy**”

average losses / empirical losses

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

least squares

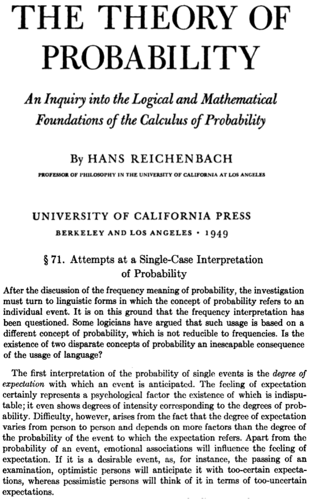
i.e. we want to minimize the error between observed loses y and predictions $\widehat{y}$.

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

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

“*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*,"

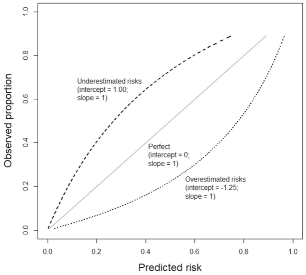
- If 40 out of 100 in this group are found to have the disease, the risk is underestimated
- If we observe that in this group, 10 out of 100 have the disease, we have overestimated the risk.

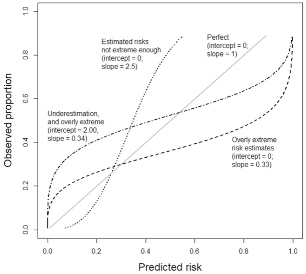
The prediction $\widehat{m}(\boldsymbol{X})$ of Y is a well-calibrated prediction if

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

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

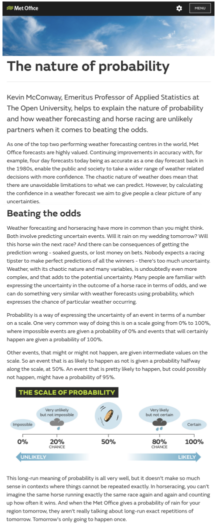
estimate risk $\widehat{y}=20\%$





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)



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

Discrimination? Individual vs. Group Treatment

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

- "**Ten Oever**" judgement (*Gerardus Cornelis Ten Oever v Stichting Bedrijfspensioenfonds voor het Glazenwassers – en Schoonmaakbedrijf*, in April 1993), the Advocate General Van Gerven (1993) argued that "*the fact that women generally live longer than men has no significance at all for the life expectancy of a specific individual and it is not acceptable for an individual to be penalized on account of assumptions which are not certain to be true in his specific case*," as mentioned in De Baere and Goessens (2011).
- Schanze (2013) used the term "**injustice by generalization**," from Britz (2008) ("**Generalisierungsunrecht**")

→ Actuarial pricing is essentially discriminatory... and unfair.

"At the core of insurance business lies discrimination".

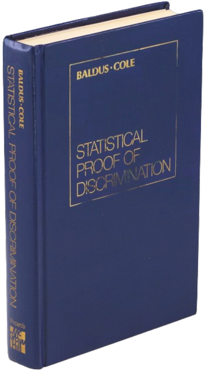
- "*What is unique about insurance is that* **even statistical discrimination which by definition is absent of any malicious intentions, poses significant moral and legal challenges**. *Why? Because on the one hand, policy makers would like insurers to treat their insureds equally, without discriminating 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)

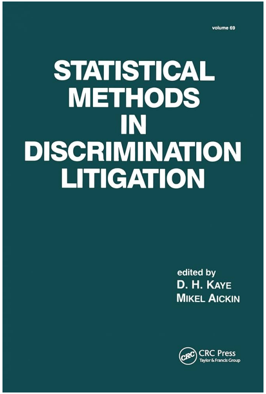
- "*Technology is neither good nor bad; nor is it neutral*," Kranzberg (1986)

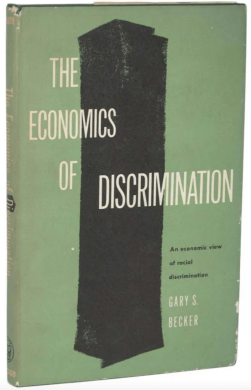
- "*Machine learning won't give you anything like gender neutrality 'for free' that you didn't explicitly ask for*," Kearns and Roth (2019)

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 $\widehat{y}_i$

  → 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

  → classical in computer science

Several definitions of “fairness” or “non-discriminatory”

sensitive sensitive

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

score $\widehat{y}$

outcome y

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

score $\widehat{y}$

outcome y

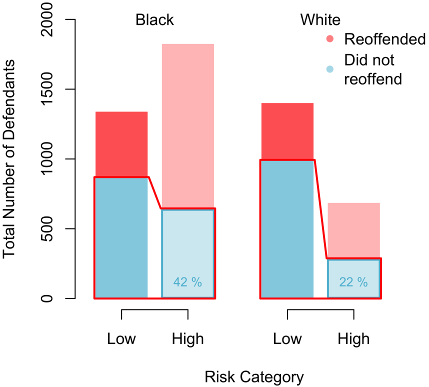
calibration → $\mathbb{E}[\,Y \mid \widehat{Y} = u, S = \mathrm{A}\,] \overset{?}{=} \mathbb{E}[\,Y \mid \widehat{Y} = u, S = \mathrm{B}\,], \forall u$

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%

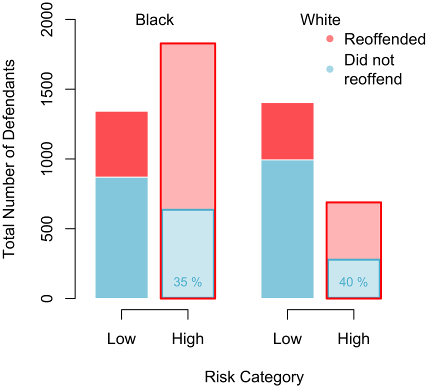


$$\mathbb{P}[\widehat{Y} = \text{high} \mid Y = \text{no}, S = \text{black}] = 42\% \overset{?}{=} \mathbb{P}[\widehat{Y} = \text{high} \mid Y = \text{no}, S = \text{white}] = 22\%,$$

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 $35 \approx 40\%$



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

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"),

decision $\widehat{y}$

$$\text{demographic parity} \rightarrow \mathbb{P}[\,\widehat{Y}=1 \,|\, S=\mathsf{A}\,] \overset{?}{=} \mathbb{P}[\,\widehat{Y}=1 \,|\, S=\mathsf{B}\,]$$

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

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

score $\widehat{m}$

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

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

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

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

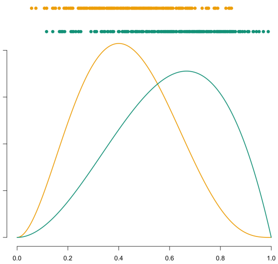
For **decisions** ($\widehat{y} \in \{0,1\}$, e.g., "obtain a loan"), we considered

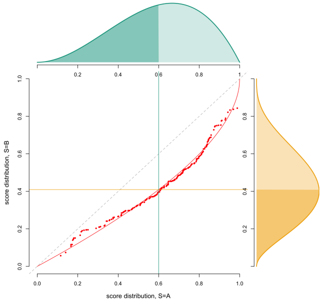
$$\text{demographic parity} \;\rightarrow\; \mathbb{E}[\,\widehat{Y}\mid S=\text{A}\,] \overset{?}{=} \mathbb{E}[\,\widehat{Y}\mid S=\text{B}\,]$$

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

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

score $\widehat{y}$





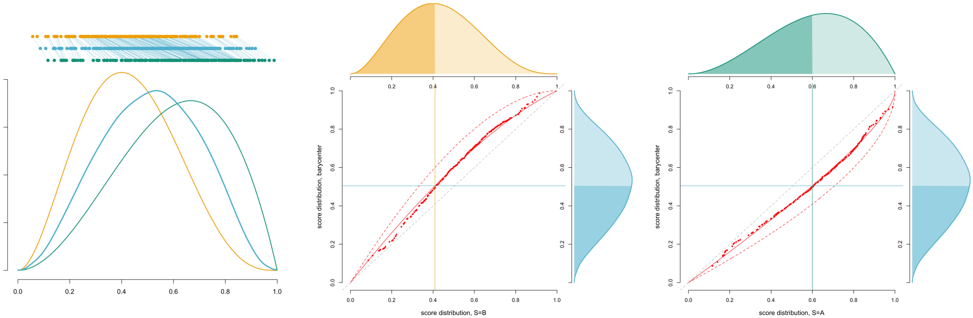
- individual in group A with a score $\widehat{y}(\text{A}) = 60\%$ corresponding to quantile α (here 0.5)
- in group B, the same quantile α corresponds to $\widehat{y}(\text{B}) = 40\%$

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

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



"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)

- 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)

See philosophical discussions about **affirmative action**, e.g., Rubenfeld (1997); Pojman (1998); Anderson (2004)

“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,

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

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” ?

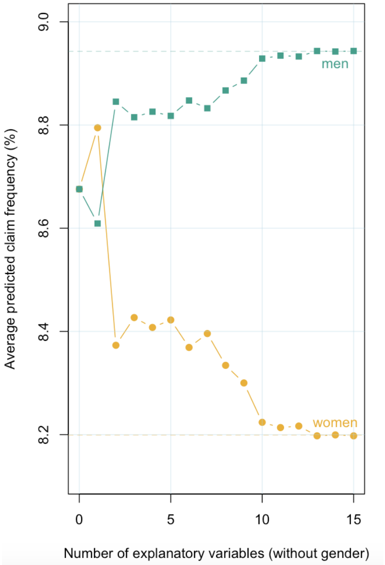
Discrimination in the data, or in the model?

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 ($s = \text{A}$) with height $x = 6'3$ (or 190 cm). If that person had been a women ($s = \text{B}$) 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

Determining whether discrimination because of sex or ethnic identity is being practiced against persons seeking passage from one social status or locus to another is an important problem in our society today. It is legally important and morally important. It is also often quite difficult. This article is an exploration of some of the issues of measurement and assessment involved in one example of the general problem, by means of which we hope to shed some light on the difficulties. We will proceed in a straightforward and indeed naive way, even though we know how misleading an unsophisticated approach to the problem is. We do this because we think it quite likely that other persons interested in questions of bias might proceed in just the same way, and careful exposure of the mistakes in our discovery procedure may be instructive.

Data and Assumptions

The particular body of data chosen for examination here consists of applications for admission to graduate study at the University of California, Berkeley, for the fall 1973 quarter. In the admissions cycle for that quarter, the Graduate Division at Berkeley received approximately 15,000 applications, some of which were later withdrawn or transferred to a different proposed entry quarter by the applicants. Of the applications finally remaining for the fall 1973 cycle 12,763 were sufficiently complete to permit a deceision to admit or to deny admission. The question we wish to pursue is whether the decision to admit or to deny was influenced by the sex of the applicant. We cannot know with any certainty the influences on the evaluators in the Graduate Admissions Office, or on the faculty reviewing committees, or on any other administrative personnel participating in the chain of actions that led to a decision on an individual application. We can, however, say that if the admissions decision and the sex of the applicant are statistically associated in the results of a series of applications, we may judge that *bias* existed, and we may then seek to find whether *discrimination* existed. By "bias" we mean here a pattern of association between a particular decision and a particular sex of applicant, of sufficient strength to make us confident that it is unlikely to be the result of chance alone. By "discrimination" we mean the exercise of decision influenced by the sex of the applicant when that is immaterial to the qualifications for entry.

The simplest approach (which we shall call approach A) is to examine the aggregate data for the campus. This approach would surely be taken by many persons interested in whether bias in admissions exists on any campus. Table 1 gives the data for all 12,763 applications to the 101 graduate departments and interdepartmental graduate majors to which application was made for fall 1973 (we shall refer to them all as departments). There were 8442 male applicants and 4321 female applicants. About 44 percent of the males and about 35 percent of the females were admitted. Just this kind of simple calculation of proportions impels us to examine the data further. We will pursue the question by using a familiar statistic, chi-square. As already noted, we are aware of the pitfalls ahead in this naive approach, but we intend to stumble into every one of them for didactic reasons.

We must first make clear two assumptions that underlie consideration of the data in this contingency table approach. Assumption 1 is that in any given discipline male and female applicants do not differ in respect of their intelligence, skill, qualifications, promise, or other attribute deemed legitimately pertinent to their acceptance as students. It is precisely this assumption that makes the study of "sex bias" meaningful, for if we did not hold it any differences in acceptance of applicants by sex could be attributed to differences in their qualifications, promise as scholars, and so on. Theoretically one could test the assumption, for example, by examining presumably unbiased estimators of academic qualification such as Graduate Record Examination scores, undergraduate grade point averages, and so on. There are, however, enormous practical difficulties in this. We therefore predicate our discussion on the validity of assumption 1.

Assumption 2 is that the sex ratios of applicants to the various fields of graduate study are not importantly associated with any other factors in admission. We shall have reason to challenge this assumption later, but it is crucial in the first step of our exploration, which is the investigation of bias in the aggregate data.

Tests of Aggregate Data

We pursue this investigation by computing the expected frequencies of male and female applicants admitted and denied, from the marginal totals of Table 1, on the assumption that men and women applicants have equal chances of admission to the university (that is, on the basis of assumptions 1 and 2). This computation, also given in Table 1, shows that 277 fewer women and 277 more men were admitted than we would have expected under the assumptions noted. That is a large number, and it is unlikely that so large a bias to the disadvantage of women would occur by chance alone. The chi-square value for this table is 110.8, and the probability of a chi-square that large (or larger) under the assumptions noted is vanishingly small.

We should on this evidence judge that bias existed in the fall 1973 admissions. On that account, we should look for the responsible parties to see whether they give evidence of discrimination. Now, the outcome of an application for admission to graduate study is determined mainly by the faculty of the department to which the prospective student applies. Let us then examine each of the departments for indications of bias. Among the 101 departments we find 16 that either had no women applicants or denied admission to no applicants of either sex. Our computations, therefore, except where otherwise noted, will be based on the remaining 85. For a start let us identify those of the 85 with bias sufficiently large to occur by chance less than five times in a hundred. There prove to be four such departments. The deficit in the number of women admitted to these four (under the assumptions for calculating expected frequencies as given above) is 26. Looking further, we find six departments biased in the opposite direction, at the same probability levels; these account for a deficit of 64 men.

These results are confusing. After all, if the campus had a shortfall of 277 women in graduate admissions, and we look to see who is responsible, we ought to find somebody. So large a deficit ought not simply to disappear. There is even a suggestion of a surplus of women. Our method of examination must be faulty.

Some Underlying Dependencies

We have stumbled onto a paradox, sometimes referred to as Simpson's in this context (*1*) or "spurious correlation" in others (*2*). It is rooted in the falsity of assumption 2 above. We have assumed that if there is bias in the proportion of women applicants admitted it will be because of a link between sex of applicant and decision to admit. We have given much less attention to a prior linkage, that between sex of applicant and department to which admission is sought. The tendency of men and women to seek entry to different departments is marked. For example, in our data almost two-thirds of the applicants to English but only 2 percent of the applicants to mechanical engineering are women. If we cast the application data into a 2 × 101 contingency table, distinguishing department and sex of applicants, we find this table has a chi-square of 3091 and that the probability of obtaining a chi-square value that large or larger by chance is about zero. For the 2 × 85 table on the departments used in most of the analysis, chi-square is 3027 and the probability about zero. Thus the sex distribution of applicants is anything but random among the departments. In examining the data in the aggregate as we did in our initial approach, we pooled data from these very different, independent decision-making units. Of course, such pooling would not nullify assumption 2 if the different departments were equally difficult to enter. We will address ourselves to that question in a moment.

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

Let us first examine an alternative to aggregating the data across the 85 departments and then computing a statistic—namely, computing a statistic on each department first and aggregating those. Fisher gives a method for aggregating the results of such independent experiments (*3*). If we apply his method to the chi-square statistics of the 85 individual contingency tables, we obtain a value that has a probability of occurrence by chance alone, that is, if sex and admission are unlinked for any major, of about 29 times in 1000 (*4*). Another common aggregation procedure, proposed to us in this context by E. Scott, yields a result having a probability of 6 times in 10,000 (*5*). This is consistent with the evidence of bias in some direction purportedly shown by Table 1. However, when we examine the *direction* of bias, the picture changes. For instance, if we apply Fisher's method to the *one-sided* statistics, testing the hypothesis of no bias or of bias in favor of women, we find that we could have obtained a value as large as or larger than the one observed, by chance alone, about 85 times in 100 (*6*).

Our first, naive approach of examining the aggregate data, computing expected frequencies under certain assumptions, computing a statistic, and deciding therefrom that bias existed in favor of men has now been cast into doubt on at least two grounds. First, we could not find many biased decision-making units by examining them individually. Second, when we take account of the differences among departments in the proportions of men and women applying to them and avoid this problem by computing a statistic on each department separately, and aggregating those statistics, the evidence for campus-wide bias in favor of men is extremely weak; on the contrary, there is evidence of bias in favor of women.

The missing piece of the puzzle is yet another fact: not all departments are equally easy to enter. If we cast the data into a 2 × 101 table, distinguishing department and decision to admit or deny, we find that this table has a chi-square value of 2195, with an associated probability of occurrence by chance (under assumptions 1 and 2) of about zero, showing that the odds of gaining admission to different departments are widely divergent. (For the 2 × 85 table chi-square is 2121 and the probability about zero.) Now, these odds of getting into a graduate program are in fact strongly associated with the tendency of men and women to apply to different departments in different degree. *The proportion of women applicants tends to be high in departments that are hard to get into and low in those that are easy to get into.* Moreover this phenomenon is more pronounced in departments with large numbers of applicants. Figure 1 is a scattergram of proportion of applicants that are women plotted against proportion of applicants that are admitted. The association is obvious on inspection although the relationship is certainly not linear (*7*). If we use a weighted correlation (*8*) as a measure of the relationship for all 85 departments in the plot we obtain $\hat{\rho} = .56$. If we apply the same measure to the 17 departments with the largest numbers of applicants (accounting for two-thirds of the total population of applicants) we obtain $\hat{\rho} = .65$, while the remaining 68 departments have a corresponding $\hat{\rho} = .39$. The significance of $\hat{\rho}$ under the hypothesis of no association can be calculated. All three values obtained are highly significant.

The effect may be clarified by means of an analogy. Picture a fishnet with two different mesh sizes. A school of fish, all of identical size (assumption 1), swim toward the net and seek to pass. The female fish all try to get through the small mesh, while the male fish all try to get through the large mesh. On the other side of the net all the fish are male. Assumption 2 said that the sex of the fish had no relation to the size of the mesh they tried to get through. It is false. To take another example that illustrates the danger of incautious pooling of data, consider two departments of a hypothetical university—machismatics and social warfare. To machismatics there apply 400 men and 200 women; these are admitted in exactly equal proportions, 200 men and 100 women. To social warfare there apply 150 men and 450 women; these are admitted in exactly equal proportions, 50 men and 150 women. Machismatics admitted half the applicants of each sex, social warfare admitted a third of the applicants of each sex. But about 73 percent of the men applied to machismatics and 27 percent to social warfare, while about 69 percent of the women applied to social warfare and 31 percent to machismatics. When these two departments are pooled and expected frequencies are computed in the usual way (with assumption 2), there is a deficit of about 21 women (Table 2). A discrepancy in that direction that large or larger would be expectable less than 2 percent of the time by chance; yet both departments were seen to have been absolutely fair in dealing with their applicants.

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	Outcome				Difference	
	Observed		Expected			
	Admit	Deny	Admit	Deny	Admit	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.

The creation of bias in our original situation is, of course, much more complex, since we are aggregating many tables. It results from an interaction of the three factors, choice of department, sex, and admission status, whose broad outlines are suggested by our plot but which cannot be described in any simple way.

In any case, aggregation in a simple and straightforward way (approach A) is misleading. More sophisticated methods of aggregation that do not rely on assumption 2 are legitimate but have their difficulties. We shall have more to say on this later.

Disaggregation

The most radical alternative to approach A is to consider the individual graduate departments, one by one. However, this approach (which we may call approach B) also poses difficulties. Either we must sample randomly from the different departments, or we must take account of the probability of obtaining unusual sex ratios of admittees by chance in a number of simultaneously conducted independent experiments. That is, in examining 85 separate departments at the same time for evidence of bias we are conducting 85 simultaneous experiments,

---

Dr. Bickel is professor of statistics, Dr. Hammel is professor of anthropology and associate dean of the Graduate Division, and Mr. O'Connell is a member of the data processing staff of the Graduate Division, at the University of California, Berkeley 94720.

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

sensitive sensitive

$$\mathbb{P}[\widehat{Y}=\text{yes} \mid S=\text{men}] \quad \geq \quad \mathbb{P}[\widehat{Y}=\text{yes} \mid S=\text{women}]$$

overall admission

$$\mathbb{P}[\widehat{Y}=\text{yes} \mid X=x, S=\text{men}] \quad \leq \quad \mathbb{P}[\widehat{Y}=\text{yes} \mid X=x, S=\text{women}], \ \forall x.$$

conditional on program

"*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) ?

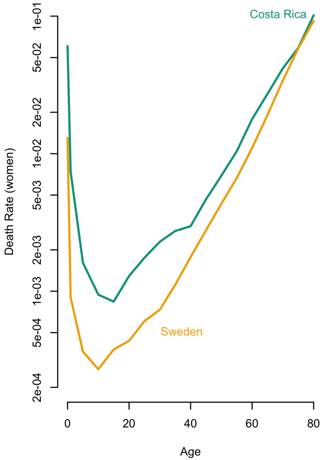
Simpson's paradox can also be seen a an **omitted variable bias** problem,

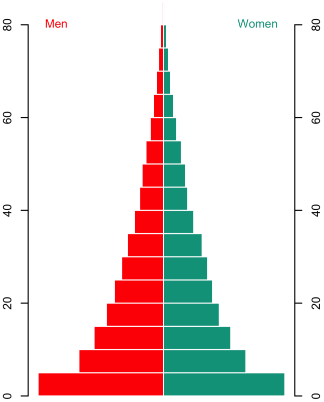
$$\begin{cases} y_i = \beta_0 + \boldsymbol{x}_1^\top \boldsymbol{\beta}_1 + \boldsymbol{x}_2^\top \boldsymbol{\beta}_2 + \varepsilon_i \text{ true model} \\ y_i = b_0 + \boldsymbol{x}_1^\top \boldsymbol{b}_1 + \eta_i \text{ estimated models} \end{cases}$$

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

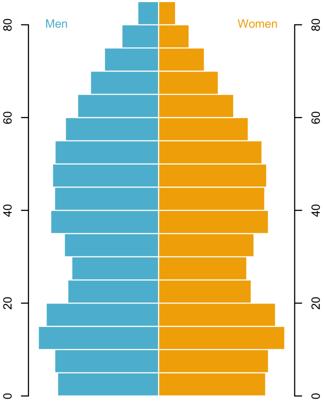
so that $\mathbb{E}[\widehat{\boldsymbol{b}}_1] = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_{12} \neq \boldsymbol{\beta}_1$.

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





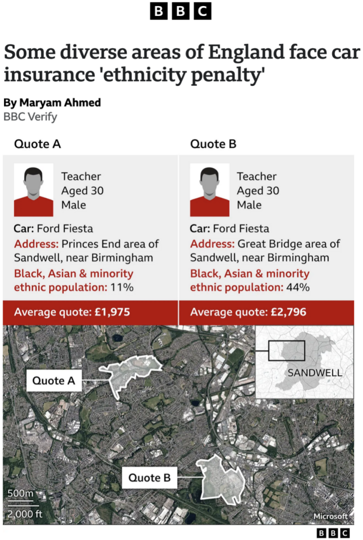
Costa Rica (1960)



Sweden (1958-1962)

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

Disentangling correlations



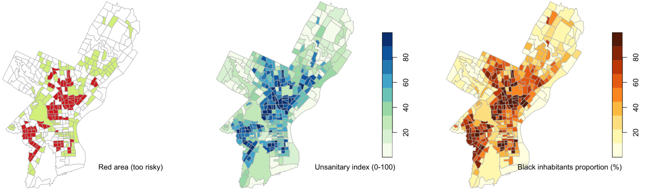
BBC

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

By Maryam Ahmed
BBC Verify

Quote A	Quote B
Teacher Aged 30 Male	Teacher Aged 30 Male
Car: Ford Fiesta	**Car:** Ford Fiesta
Address: Princes End area of Sandwell, near Birmingham	**Address:** Great Bridge area of Sandwell, near Birmingham
Black, Asian & minority ethnic population: 11%	**Black, Asian & minority ethnic population:** 44%
Average quote: £1,975	**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”?

sensitive attribute 1

$$\mathbb{E}[\,\widehat{m}(\boldsymbol{X},S_1,S_2)\,|\,S_1=\mathrm{A}\,]\neq\mathbb{E}[\,\widehat{m}(\boldsymbol{X},S_1,S_2)\,|\,S_1=\mathrm{B}\,]$$

$$\mathbb{E}[\,\widehat{m}(\boldsymbol{X},S_1,S_2)\,|\,S_2=\mathrm{C}\,]\approx\mathbb{E}[\,\widehat{m}(\boldsymbol{X},S_1,S_2)\,|\,S_2=\mathrm{D}\,]$$

sensitive attribute 2

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

sensitive attribute 1

$$\mathbb{E}[\,\widetilde{m}(\boldsymbol{X},S_1,S_2)\,|\,S_1=\mathrm{A}\,]=\mathbb{E}[\,\widetilde{m}(\boldsymbol{X},S_1,S_2)\,|\,S_1=\mathrm{B}\,]$$

$$\mathbb{E}[\,\widetilde{m}(\boldsymbol{X},S_1,S_2)\,|\,S_2=\mathrm{C}\,]\neq\mathbb{E}[\,\widetilde{m}(\boldsymbol{X},S_1,S_2)\,|\,S_2=\mathrm{D}\,]$$

sensitive attribute 2

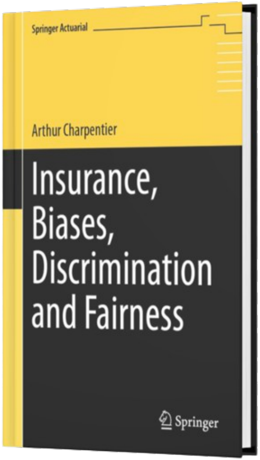
“The myth of the actuary” (objectivity vs. subjectivity)

- 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

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.

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.

Britz, G. (2008). *Einzelfallgerechtigkeit versus Generalisierung: verfassungsrechtliche Grenzen statistischer Diskriminierung*. Mohr Siebeck.

Casey, B., Pezier, J., and Spetzler, C. (1976). *The Role of Risk Classification in Property and Casualty Insurance: A Study of the Risk Assessment Process : Final Report*. Stanford Research Institute.

References

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

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, 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. *forthcoming*.

De Baere, G. and Goessens, E. (2011). Gender differentiation in insurance contracts after the judgment in case c-236/09, Association Belge des Consommateurs Test-Achats asbl v. conseil des ministres. *Colum. J. Eur. L.*, 18:339.

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

References

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. *Thirty-Eighth Annual Conference on Neural Information Processing Systems*.

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*.

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

References

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.

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.

References

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.

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.

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.

References

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

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

Schanze, E. (2013). Injustice by generalization: notes on the Test-Achats decision of the european court of justice. *German Law Journal*, 14(2):423–433.

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.

Van Gerven, G. (1993). Case c-109/91, Gerardus Cornelis Ten Oever v. Stichting bedrijfspensioenfonds voor het glazenwassers-en schoonmaakbedrijf. *EUR-Lex*, 61991CC0109.

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

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

Walters, M. A. (1981). Risk classification standards. In *Proceedings of the Casualty Actuarial Society*, volume 68, pages 1–18.