Multivariate Optimal Transport

We have mentioned that, in dimension one

Theorem 4.6: Optimal map for continuous univariate distributions

The optimal Monge map $T^\star$ such that $T^\star_\# \mathbb{P}_{\mathtt{A}} = \mathbb{P}_{\mathtt{B}}$ is $T^\star = F_{\mathtt{B}}^{-1} \circ F_{\mathtt{A}}$.

- $T^\star$ is an increasing mapping.

Example Univariate Gaussian

$$x_{\mathtt{B}} = T^\star(x_{\mathtt{A}}) = \mu_{\mathtt{B}} + \sigma_{\mathtt{B}}\sigma_{\mathtt{A}}^{-1}(x_{\mathtt{A}} - \mu_{\mathtt{A}}).$$

Multivariate Optimal Transport

Theorem 4.7: Optimal map for continuous multivariate distributions, Brenier (1991)

With a quadratic cost, the optimal Monge map $T^\star$ is unique, and it is the gradient of a convex function, $T^\star = \nabla\varphi$.

Example Multidimensional Gaussian

$$\boldsymbol{x}_{\mathrm{B}} = \mathcal{T}^\star(\boldsymbol{x}_{\mathrm{A}}) = \boldsymbol{\mu}_{\mathrm{B}} + \boldsymbol{A}(\boldsymbol{x}_{\mathrm{A}} - \boldsymbol{\mu}_{\mathrm{A}}),$$

where $\boldsymbol{A}$ is a symmetric positive matrix that satisfies $\boldsymbol{A}\boldsymbol{\Sigma}_{\mathrm{A}}\boldsymbol{A} = \boldsymbol{\Sigma}_{\mathrm{B}}$, which has a unique solution given by $\boldsymbol{A} = \boldsymbol{\Sigma}_{\mathrm{A}}^{-1/2}\big(\boldsymbol{\Sigma}_{\mathrm{A}}^{1/2}\boldsymbol{\Sigma}_{\mathrm{B}}\boldsymbol{\Sigma}_{\mathrm{A}}^{1/2}\big)^{1/2}\boldsymbol{\Sigma}_{\mathrm{A}}^{-1/2}$, where $\boldsymbol{M}^{1/2}$ is the square root of the square (symmetric) positive matrix $\boldsymbol{M}$ based on the Schur decomposition ($\boldsymbol{M}^{1/2}$ is a positive symmetric matrix), as described in Higham (2008).

Generalized Linear Model

Definition 4.39: Exponential family, McCullagh and Nelder (1989)

The distribution of Y is in the exponential family if its density (with respect to some appropriate measure) is

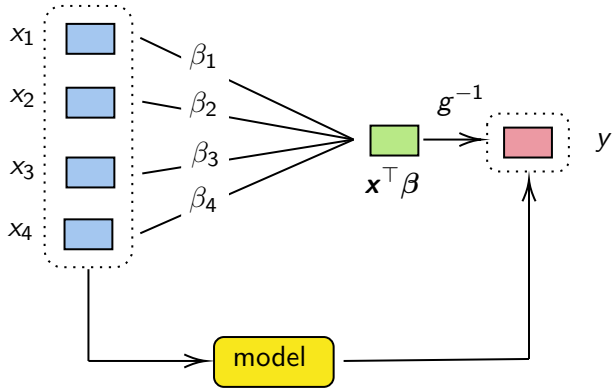
$$f_{\theta,\varphi}(y) = \exp\left(\frac{y\theta - b(\theta)}{\varphi} + c(y,\varphi)\right),$$

where θ is the canonical parameter, φ is a nuisance parameter, and $b : \mathbb{R} \to \mathbb{R}$ is some $\mathbb{R} \to \mathbb{R}$ function.

- Such as the binomial, Poisson, Gaussian, gamma distributions, etc.
- Also compound Poisson / Tweedie (from Tweedie (1984)).

Generalized Linear Model

- Given some dataset $(y_i, \boldsymbol{x}_i)$, suppose that $\mu(\boldsymbol{x}) = g^{-1}(\boldsymbol{x}^\top\beta)$



- OLS, $\mu(\boldsymbol{x}) = \boldsymbol{x}^\top\beta$ and $\widehat{\beta}^{\text{ols}} = \text{argmin}\left\{\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top\beta)^2\right\} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$.

Generalized Linear Model

- We modeling, we want to solve problems such as

$$\begin{cases} \min\limits_{f \in \mathcal{F}} \{f(\boldsymbol{x})\} \\ \text{s.t. } f \in \mathcal{F}_0, \text{ where } \boldsymbol{x} \text{ is given.} \end{cases} \quad \text{or} \quad \begin{cases} \min\limits_{\boldsymbol{\theta} \in \mathbb{R}^k} \{f_{\boldsymbol{\theta}}(\boldsymbol{x})\} \\ \text{s.t. } \boldsymbol{\theta} \in \Theta, \text{ where } \boldsymbol{x} \text{ is given.} \end{cases}$$

- In optimisation, it is usually written the other way round,

$$\begin{cases} \min\limits_{\boldsymbol{x} \in \mathbb{R}^k} \{f(\boldsymbol{x})\} \\ \text{s.t. } \boldsymbol{x} \in \mathcal{E}, \text{ where } f \text{ is given.} \end{cases}$$

Generalized Linear Model

$$\begin{cases} \min_{\boldsymbol{x}\in\mathbb{R}^k} \{f(\boldsymbol{x})\} \\ \text{s.t. } \boldsymbol{x} \in \mathcal{E}, \text{ where } f \text{ is given.} \end{cases}$$

can be written

$$\min_{\boldsymbol{x}\in\mathbb{R}^k} \{f(\boldsymbol{x}) + \lambda p(\boldsymbol{x})\}$$

where $\lambda > 0$ is some penalty factor, and $p(\cdot)$ is some function. With

$$p(\boldsymbol{x}) = \begin{cases} 0 \text{ si } \boldsymbol{x} \in \mathcal{E} \\ +\infty \text{ si } \boldsymbol{x} \notin \mathcal{E} \end{cases}$$

the two problems are equivalent. Other p functions can be considered.

- p is said to be an exact penalty if the two problems are equivalent.

Generalized Linear Model

- E.g., if $\mathcal{E} = \mathbb{R}_+^k = \{\boldsymbol{x} : \boldsymbol{x} \geq \boldsymbol{0}\}$, we can consider

$$p(\boldsymbol{x}) = -\sum_{i=1}^{k} \log(x_i) \quad \text{(as suggested by Frisch (1955))}$$

Generalized Linear Model

- Consider problems

$$\min_{\boldsymbol{x}\in\mathbb{R}^k}\{f(\boldsymbol{x})\} \text{ under constraint } g(\boldsymbol{x})=\boldsymbol{0} \quad \text{or} \quad \min_{\boldsymbol{x}\in\mathbb{R}^k}\{f(\boldsymbol{x})\} \text{ under constraint } g(\boldsymbol{x})\leq\boldsymbol{0}$$

- Karush-Kuhn-Tucker condition is

$$\begin{cases}\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^\star,\boldsymbol{z}^\star)=\boldsymbol{0}\\ \nabla_{\boldsymbol{z}}\mathcal{L}(\boldsymbol{x}^\star,\boldsymbol{z}^\star)=\boldsymbol{0}\end{cases}$$

where

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{z})=f(\boldsymbol{x})+\boldsymbol{z}^\top g(\boldsymbol{x})$$

is the Lagrangian problem (parameter $\boldsymbol{z}$ are multipliers)

Generalized Linear Model

Definition 4.40: Ridge Estimator (OLS), Hoerl and Kennard (1970)

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^k}{\text{argmin}} \left\{ \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{k}\beta_j^2 \right\}.$$

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\mathbb{I})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

Definition 4.41: Ridge Estimator (GLM)

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^k}{\text{argmin}} \left\{ -\sum_{i=1}^{n}\log f(y_i|\mu_i = g^{-1}(\boldsymbol{x}_i^\top\boldsymbol{\beta})) + \lambda\sum_{j=1}^{k}\beta_j^2 \right\}.$$

Generalized Linear Model

Definition 4.42: LASSO Estimator (OLS), Tibshirani (1996)

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{lasso}} = \text{argmin} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{k} |\beta_j| \right\}.$$

Definition 4.43: LASSO Estimator (GLM)

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{lasso}} = \text{argmin} \left\{ - \sum_{i=1}^{n} \log f(y_i | \mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})) + \lambda \sum_{j=1}^{k} |\beta_j| \right\}.$$

Generalized Linear Model

```
> library(glmnet)
> fit_ridge = glmnet(x, y, alpha = 0)
> fit_lasso = glmnet(x, y, alpha = 1)
```

- Elastic net

$$\min\left\{\frac{1}{2}\sum_{i=1}^{n}(y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta})^2+\lambda_1\sum_{j=1}^{k}|\beta_j|+\frac{\lambda_2}{2}\sum_{j=1}^{k}\beta_j^2\right\},$$

e.g. $\lambda_1=\alpha\lambda$ and $\lambda_2=(1-\alpha)\lambda$ (two parameters — one for the global regularization, one for the trade-off between Ridge (Tikhonov) vs. Lasso)

to go further (for more details on penalization issues)

Accuracy

Definition 4.44: ROC curve

The ROC curve is the parametric curve

$$\{\mathbb{P}[m(\boldsymbol{X}) > t | Y = 0], \mathbb{P}[m(\boldsymbol{X}) > t | Y = 1]\} \text{ for } t \in [0, 1],$$

when the score $m(\boldsymbol{X})$ and Y evolve in the same direction (a high score indicates a high risk).

$$C(t) = \text{TPR} \circ \text{FPR}^{-1}(t),$$

where

$$\begin{cases} \text{FRP}(t) = \mathbb{P}[m(\boldsymbol{X}) > t | Y = 0] = \mathbb{P}[m_0(\boldsymbol{X}) > t] \\ \text{TPR}(t) = \mathbb{P}[m(\boldsymbol{X}) > t | Y = 1] = \mathbb{P}[m_1(\boldsymbol{X}) > t]. \end{cases}$$

Accuracy

```
> library(ROCR)
> pred = prediction(df$yhat, df$y)
> roc = performance(pred,"tpr","fpr")
> plot(roc)
> auc = performance( pred,"auc")
```

Definition 4.45: AUC, area under the ROC curve

The area under the curve is defined as the area below the ROC curve,

$$\text{AUC} = \int_0^1 C(t)\mathrm{d}t = \int_0^1 \text{TPR} \circ \text{FPR}^{-1}(t)\mathrm{d}t.$$

Calibration

- Well-calibration was initially discussed in forecasting

Definition 4.46: Well-calibrated (1), Van Calster et al. (2019), Krüger and Ziegel (2021)

The forecast X of Y is a well-calibrated forecast of Y if $\mathbb{E}(Y|X) = X$ almost surely, or $\mathbb{E}[Y|X = x] = x$, for all x.

- one can define "well-calibration" in prediction

Definition 4.47: Well-calibrated (2), Zadrozny and Elkan (2002); Cohen and Goldszmidt (2004)

The prediction $m(\boldsymbol{X})$ of Y is a well-calibrated prediction if $\mathbb{E}[Y|m(\boldsymbol{X}) = \widehat{y}] = \widehat{y}$, for all $\widehat{y}$.

Calibration

"**Well calibrated** *classifiers are probabilistic classifiers for which the output can be directly interpreted as a confidence level. For instance, a well calibrated (binary) classifier should classify the samples such that among the samples to which it gave a*[predicted probability] *value close to 0.8, approximately 80% actually belong to the positive class*," scikit learn: Probability calibration

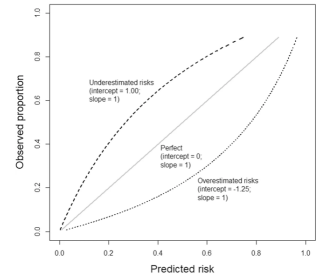
- "*Suppose that a forecaster sequentially assigns probabilities to events. He is* **well calibrated** *if, for example, of those events to which he assigns a probability 30 percent, the long-run proportion that actually occurs turns out to be 30 percent*," Dawid (1982).
- "*Out of all the times you said there was a 40 percent chance of rain, how often did rain actually occur? If, over the long run, it really did rain about 40 percent of the time, that means your forecasts were* **well calibrated**," Silver (2012),
- "*we desire that the estimated class probabilities are reflective of the true underlying probability of the sample*," Kuhn and Johnson (2013)

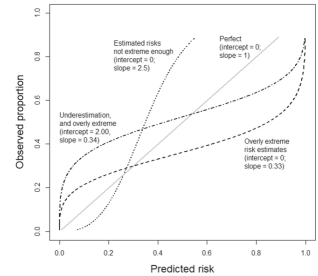
Calibration

- See Murphy and Epstein (1967), Roberts (1968), Gneiting and Raftery (2005) on ensemble methods for weather forecasting, or more generally Lichtenstein et al. (1977), Oakes (1985), Gneiting et al. (2007).

Calibration

- 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.
- Hosmer-Lemeshow test, from Hosmer Jr et al. (2013) (logistic regression), and Bier score, from Brier (1950) and Murphy (1973)
- Function plotted in psychological papers Keren (1991)

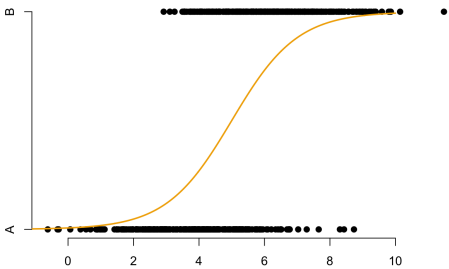


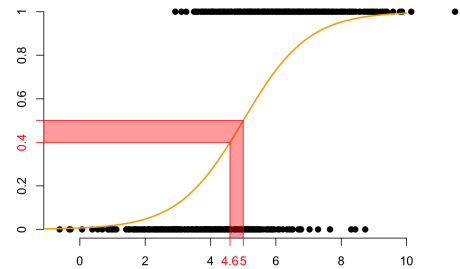


Calibration

- Consider a dataset (x_i, y_i), $y_i \in \{A, B\}$, and consider model

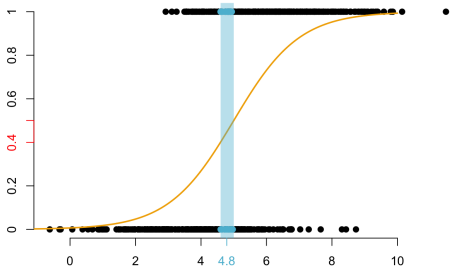
$\widehat{m}(x) = \dfrac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}$ to estimate $\mathbb{P}[Y = \mathtt{B} | X = x]$ (logistic regression). Given $[p_-, p_+] \subset [0, 1]$ (here $[0.4, 0.5]$), set $\mathcal{I} = \{i : m(\boldsymbol{x}_i) \in [p_-, p_+]\}$.

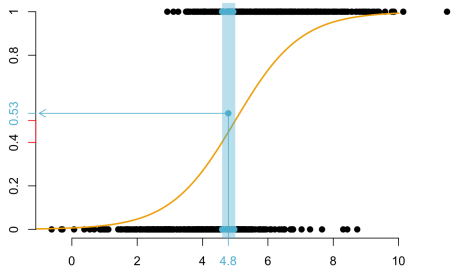




Calibration

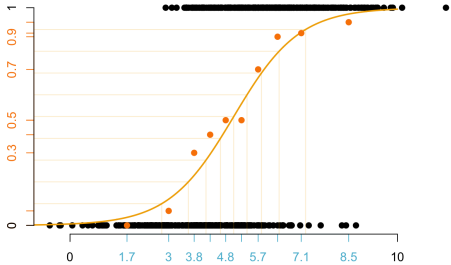
- Given $\mathcal{I} = \{i : m(\boldsymbol{x}_i) \in [p_-, p_+]\}$, set $\overline{y}_{\mathcal{I}} = \dfrac{1}{n_{\mathcal{I}}} \displaystyle\sum_{i \in \mathcal{I}} y_i$

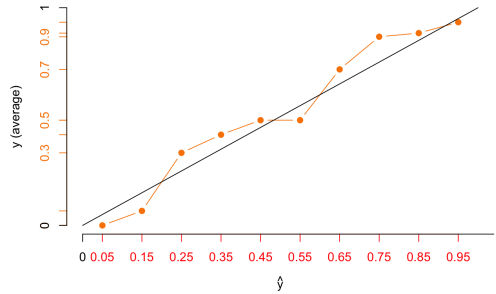




Calibration

› Compute deciles $\overline{y}_1, \cdots, \overline{y}_{10}$ associated with $[p_-, p_+]$ equal $[0, 0.1]$, $[0.1, 0.2]$, $\cdots$, $[0.9, 1]$, with midpoints p_k. Then visualize $\{(p_k, \overline{y}_k)\}$, as in `scikit-learn`.

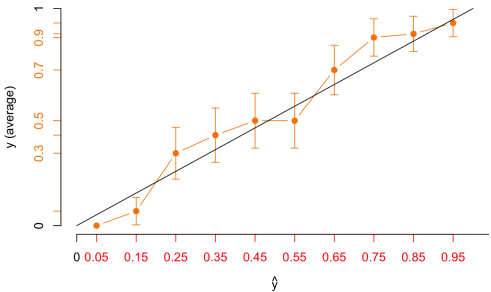


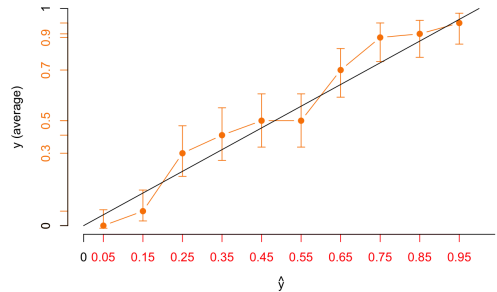


Calibration

- Asymptotic and Agresti and Coull (1998), $\tilde{n} = n + u_{1-\alpha/2}^2$ et $\tilde{p} = \frac{1}{\tilde{n}}\left(n\overline{x} + \frac{u_{1-\alpha/2}^2}{2}\right)$

$$\left[\widehat{p} \pm u_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right], \text{ and } \left[\tilde{p} \pm u_{1-\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}\right]$$

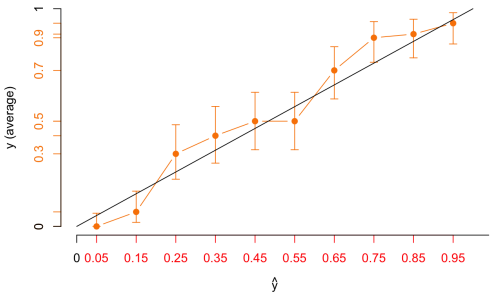


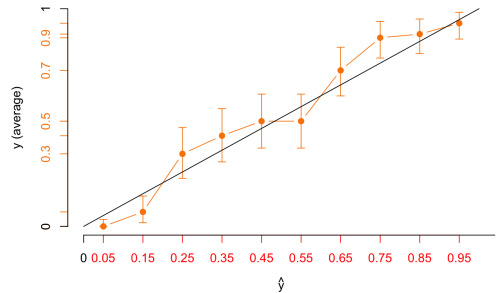


Calibration

- Exact, and Wilson (1927)

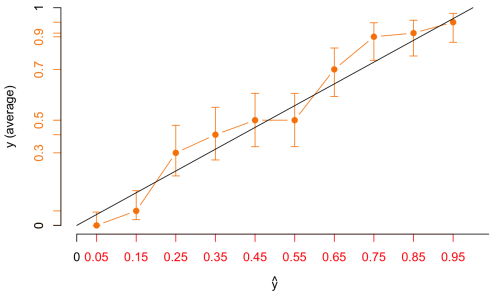
$$\left[\frac{1}{1+\frac{u^2_{1-\alpha/2}}{n}} \left(\hat{p} + \frac{u^2_{1-\alpha/2}}{2n} \right) \pm \frac{u_{1-\alpha/2}}{1+\frac{u^2_{1-\alpha/2}}{n}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{u^2_{1-\alpha/2}}{4n^2}} \right]$$

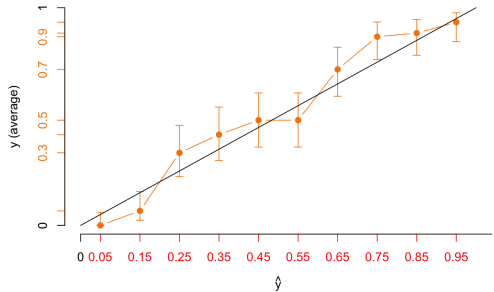




Calibration

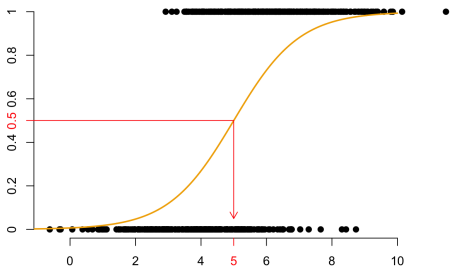
- Bayes and Probit (via `binom.confint` in `binom` package)

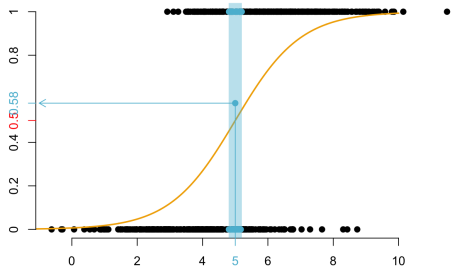




Calibration

- Given $p \in (0,1)$, consider $\mathcal{I}_p = \{i : \widehat{m}(\boldsymbol{x}_i) \in [p-h, p+h]\}$ for some $h > 0$, set $\overline{y}_p = \dfrac{1}{n_{\mathcal{I}_p}} \displaystyle\sum_{i \in \mathcal{I}_p} y_i$

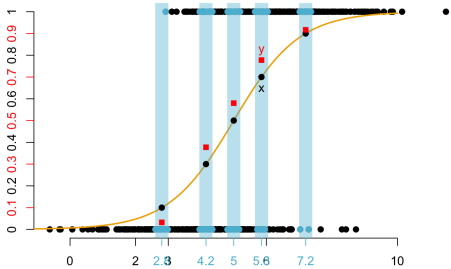


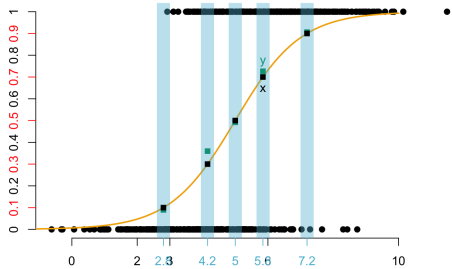


Calibration

- Given $p \in (0,1)$, compute $\overline{y}_p = \dfrac{1}{n_{\mathcal{I}_p}} \sum_{i \in \mathcal{I}_p} y_i$ (for appropriate bandwidth $h > 0$)

One could also consider some kermel based average... $\overline{y}_p$

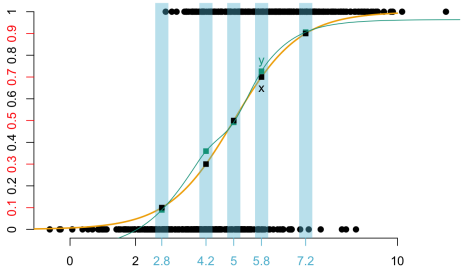


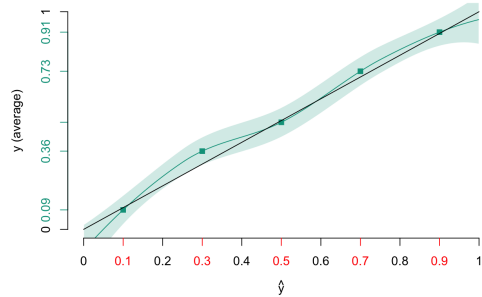


Calibration

- Compute $\overline{y}_p$ for various p, and plot $p \mapsto \overline{y}_p$, that is an estimate of $\mathbb{E}[Y|\widehat{m}(\boldsymbol{X}) = p]$.
- Add a confidence band around.

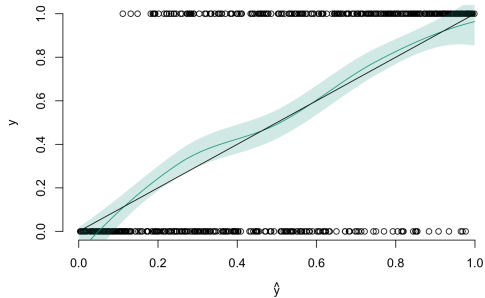
... but here, it works only because $\widehat{m}$ is smooth enough...

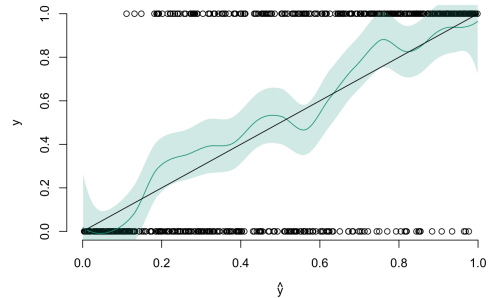




Calibration

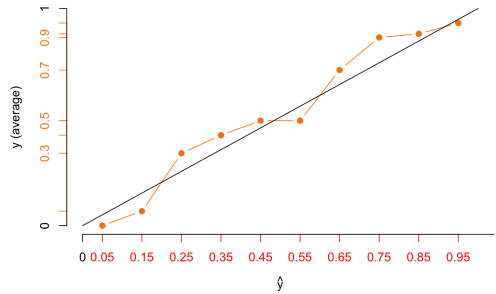
- More generally, consider the local regression of y_i's against $\widehat{y}_i = \widehat{m}(\boldsymbol{x}_i)$'s.





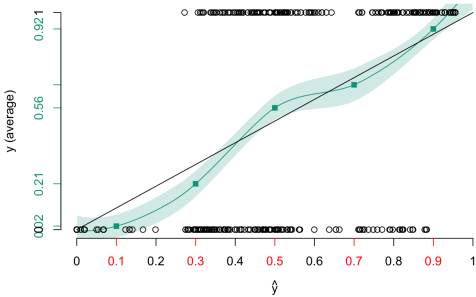
Calibration

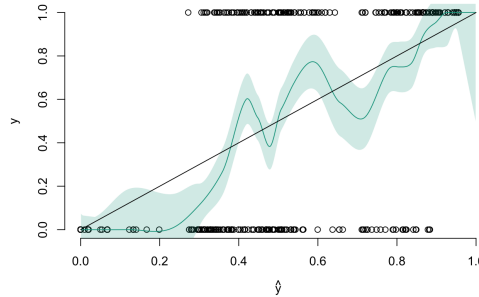
› Consider e.g., some random forest model for $\widehat{m}$, and then the calibration curve per decile,



Calibration

- Or, for our random forest model $\widehat{m}$, some local regression approach





Calibration

Definition 4.48: Calibration plot

The calibration plot associated with model m is the function $\widehat{y} \mapsto \mathbb{E}(Y|m(\boldsymbol{X}) = \widehat{y})$. The empirical version is some local regression on $\{y_i, m(\boldsymbol{x}_i)\}$.

Definition 4.49: Globally unbiased model m, Denuit et al. (2021)

Model m is globally unbiased if $\mathbb{E}[Y] = \mathbb{E}[m(\boldsymbol{X})]$.

Definition 4.50: Locally unbiased model m, Denuit et al. (2021)

Model m is locally unbiased at $\widehat{y}$ if $\mathbb{E}[Y|m(\boldsymbol{X}) = \widehat{y}] = \widehat{y}$.

Calibration

- For GLM, remember that

$$f(y_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\varphi} + c(y_i, \varphi)\right),$$

$$\frac{\partial \log \mathcal{L}_i}{\partial \beta_j} = \frac{\partial \log \mathcal{L}_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \log \mathcal{L}_i}{\partial \beta_j} = \frac{y_i - \mu_i}{\varphi} \cdot \frac{1}{V(\mu_i)} \cdot x_{i,j} \cdot \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-1}$$

- When g is the canonical link ($g_\star = b'^{-1}$ or $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \theta_i$)

$$\nabla \log \mathcal{L} = \boldsymbol{X}^\top(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \boldsymbol{0}$$

Proposition 4.33: Calibration of GLM

In the GLM framework with the canonical link function, $\widehat{m}(\boldsymbol{x}) = g_\star^{-1}(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})$ is globally unbiased (on the training dataset), but possibly locally biased.

Calibration

- Otherwise

$$\nabla \log \mathcal{L} = \boldsymbol{X}^\top \boldsymbol{\Omega}(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \boldsymbol{0},$$

where $\boldsymbol{\Omega}$ is a diagonal matrix ($\boldsymbol{\Omega} = \boldsymbol{W}\boldsymbol{\Delta}$, where $\boldsymbol{W} = \text{diag}((V(\mu_i)g'(\mu_i)^2)^{-1})$ and $\boldsymbol{\Delta} = \text{diag}(g'(\mu_i))$, so that we recognize Fisher information - corresponding to the Hessian matrix (up to a negative sign) – $\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X}$).

	training data					validation data				
	$\overline{y}$	GLM	CART	GAM	RF	$\overline{y}$	GLM	CART	GAM	RF
$\widehat{m}(\boldsymbol{x}, s)$	8.73	8.73	8.73	8.73	8.27	8.55	9.05	9.03	8.84	8.70
$\widehat{m}(\boldsymbol{x})$	8.73	8.73	8.73	8.73	8.29	8.55	9.05	9.03	8.84	8.73

Calibration

Definition 4.51: Brier score (binary classifier) Brier (1950)

Brier score is the mean squared error of probability estimate,

$$\text{BS} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{m}(\boldsymbol{x}_i), y_i\right)^2$$

- Let g be the calibration function, $g(\widehat{m}(\boldsymbol{x})) \approx p(\boldsymbol{x})$.
- Platt scaling (from Platt et al. (1999)), $g(s) = [1 + e^{-(ws+b)}]^{-1}$.

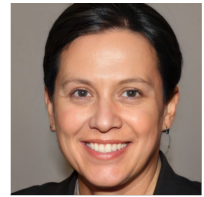
- "confidence" value given by `Picpurify`, using pictures generate by a GAN (a generative adversarial network, used in Hill and White (2020)).

Calibration

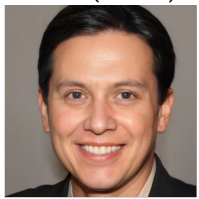
female (0.984)
male (0.016)

female (0.983)
male (0.017)

female (0.982)
male (0.018)

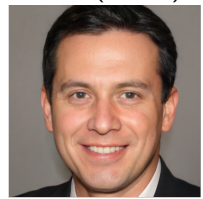


female (0.960)
male (0.040)



female (0.009)
male (0.991)

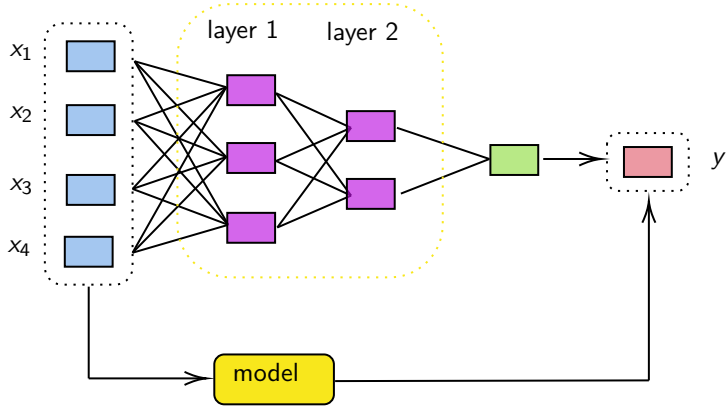
female (0.013)
male (0.987)



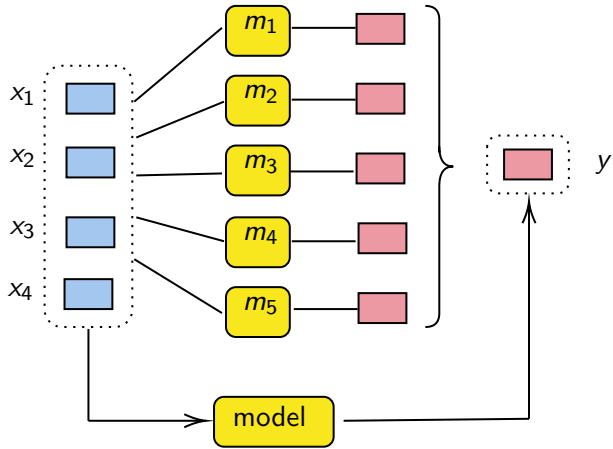
female (0.014)
male (0.986)

female (0.015)
male (0.985)

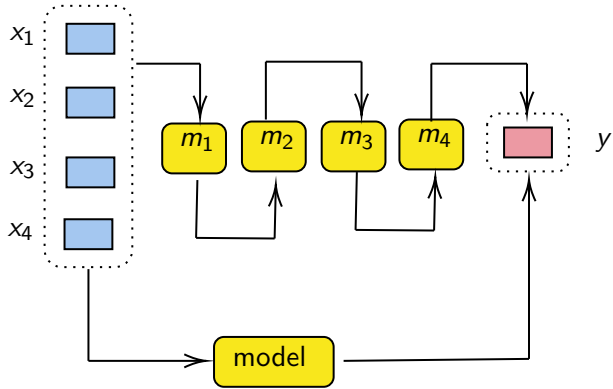
Standard modeling architecture



Standard modeling architecture



Standard modeling architecture

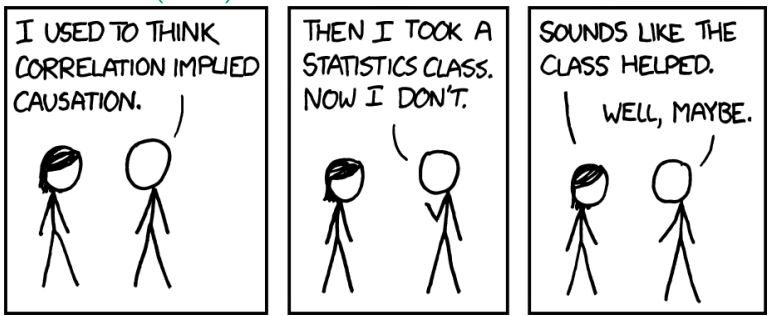


– Part 3 –

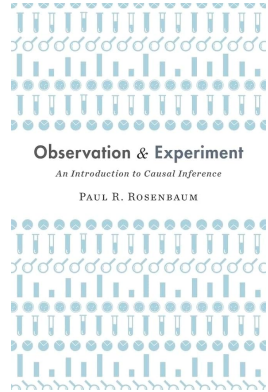
Data

Data (the two types)

"*It is often said, 'You cannot prove causality with statistics.' One of my professors, Frederick Mosteller, liked to counter, 'You can only prove causality with statistics.'* (...) *The title, 'Observation and Experiment,' marks the modern distinction between randomized experiments and observational studies.*"
Rosenbaum (2018)



Correlation, Randall Munroe, 2009 https://xkcd.com/552/



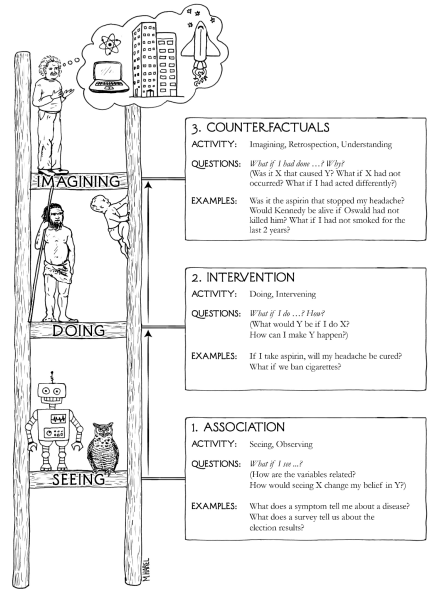
Data (the three rung ladder)

"Ladder of causation" from Pearl et al. (2009)

- 3. **Counterfactuals**
(Imagining, "*what if I had done...*")
- 2. **Intervention**
(Doing, "*what if I do...*")
- 1. **Association**
(Seeing, "*what if I see...*")

Picture source: Pearl and Mackenzie (2018)

What would be the impact of a treatment T on a variable of interest Y ?



Proxy

- "*OK, let's not use race, but should we use zip code, which of course is a proxy for race in our segregated society?*," O'Neil (2016).

Definition 4.52: Proxy, Merriam-Webster (2022)

A proxy is a person authorized to act for another (from a contracted form of the Middle English word *procuracie* (from French "procuration")).

Definition 4.53: Perfect proxy, Datta et al. (2017)

A variable X is a perfect proxy for Z if there exist functions $\varphi : \mathcal{X} \to \mathcal{Z}$ and $\psi : \mathcal{Z} \to \mathcal{Y}$ such that

$$\mathbb{P}[X = \psi(Z)] = \mathbb{P}[\varphi(X) = Z] = 1.$$

Proxy

Definition 4.54: Comonotonicity, Hoeffding (1940); Fréchet (1951)

Variables X and Y are comonotonic if $(X, Y) = (F_x^{-1}(U), F_y^{-1}(U))$ for some $U \sim \mathcal{U}([0, 1])$.

- See also Dhaene et al. (2002a,b) on comonotonic vectors.
- See also Prince and Schwarcz (2019), or Tschantz (2022) for discrimination by proxy.
- Range of possible situation between independence and perfect proxy.

Independence

Definition 4.55: Independence (dimension 2)

X and Y are independent, denoted $X \perp\!\!\!\perp Y$, if for any sets $\mathcal{A}, \mathcal{B} \subset \mathbb{R}$,

$$\mathbb{P}[X \in \mathcal{A}, Y \in \mathcal{B}] = \mathbb{P}[X \in \mathcal{A}] \cdot \mathbb{P}[Y \in \mathcal{B}].$$

Definition 4.56: Linear Independence (dimension 2)

Consider two random variables X and Y. $X \perp Y$ if and only if $\text{Cov}[X, Y] = 0$.

Independence

Definition 4.57: Correlation (dimension 2), Pearson (1895)

X and Y are two random variables

$$\text{Corr}[X,Y] = \frac{\text{Cov}[X,Y]}{\sqrt{\text{Var}[X]\cdot\text{Var}[Y]}}.$$

where $\text{Cov}[X,Y] = \mathbb{E}\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\big] = \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]$.

Independence

Proposition 4.34: Correlation bounds (dimension 2)

For any random variables X and Y (with finite variances), $r_{\min} \leq \text{Corr}[X, Y] \leq r_{\max}$, where

$$r_{\min} = \frac{\text{Cov}[F_x^{-1}(U), F_y^{-1}(1-U)]}{\sqrt{\text{Var}[X] \cdot \text{Var}[Y]}} \text{ and } r_{\max} = \frac{\text{Cov}[F_x^{-1}(U), F_y^{-1}(U)]}{\sqrt{\text{Var}[X] \cdot \text{Var}[Y]}}$$

- Maximal correlation is obtained when X and Y are comonotonic (minimal correlation when X and $-Y$ are comonotonic).
- Related to optimal transport, see also Knott and Smith (1984).

Independence

Proposition 4.35

Consider two random variables X and Y. $X \perp\!\!\!\perp Y$ if and only if for any functions $\varphi : \mathbb{R} \to \mathbb{R}$ and $\psi : \mathbb{R} \to \mathbb{R}$ (such that the expected values below exist and are well-defined) $\text{Cov}[\varphi(X), \psi(Y)] = 0$, i.e.,

$$\mathbb{E}[\varphi(X) \cdot \psi(Y)] = \mathbb{E}[\varphi(X)] \cdot \mathbb{E}[\psi(Y)].$$

Definition 4.58: Maximal Correlation, HGR

Consider two random variables X and Y,

$$r^\star(X, Y) = \max_{\varphi,\psi} \{\text{Corr}[\varphi(X), \psi(Y)]\}.$$

Independence

- HGR because of Hirschfeld (1935), Gebelein (1941) and Rényi (1959) (also Sarmanov (1958a,b)).

$$r^\star(X,Y) = \max_{\varphi\in\mathcal{F}_x,\ \psi\in\mathcal{G}_y} \mathbb{E}[\varphi(X)\psi(Y)],$$

where

$$\begin{cases} \mathcal{F}_x = \{\varphi : \mathcal{X} \to \mathbb{R} : \mathbb{E}[\varphi(X)] = 0 \text{ and } \mathbb{E}[\varphi^2(X)] = 1\} \\ \mathcal{G}_y = \{\psi : \mathcal{Y} \to \mathbb{R} : \mathbb{E}[\psi(Y)] = 0 \text{ and } \mathbb{E}[\psi^2(Y)] = 1\} \end{cases}$$

- See either `ccaPP` or `acepack` package,

```
> ccaPP::maxCorProj(x = x, y = y, method = "pearson")
> corstar = acepack::ace(x = x, y = y)
> cor(corstar$tx, corstar$ty)
```

Proposition 4.36

Consider two random variables X and Y. $X \perp\!\!\!\perp Y$ if and only if $r^\star(X, Y) = 0$.

Proof: Given a random variable X, its characteristic function is $\phi_X(t) = \mathbb{E}[e^{itX}]$. Recall that

$$\begin{cases} \phi_X(t) = \phi_Y(t), \ \forall t \in \mathbb{R} \text{ if and only if } X \overset{\mathcal{L}}{=} Y \\ \phi_{X,Y}(s,t) = \mathbb{E}[e^{i(sX+tY)}] = \phi_X(s) \cdot \phi_Y(t), \ \forall s,t \in \mathbb{R} \text{ if and only if } X \perp\!\!\!\perp Y \end{cases}$$

If $r^\star(X, Y) = 0$, let $s, t \in \mathbb{R}$ and consider $\varphi(x) = \phi_X(x) = \mathbb{E}[e^{ixX}]$ and $\psi(y) = \phi_Y(y) = \mathbb{E}[e^{isY}]$, then $\text{Cov}[e^{isX}, e^{itY}] = \text{Cov}[X'_s, Y'_t] = 0$, i.e. $\mathbb{E}[X'_s Y'_t] = \mathbb{E}[X'_s]\mathbb{E}[Y'_t]$,

$$\underbrace{\mathbb{E}[e^{i(sX+tY)}]}_{\phi_{XY}(s,t)} = \underbrace{\mathbb{E}[e^{isX}] \cdot \mathbb{E}[e^{itY}]}_{\phi_X(s)\cdot\phi_Y(t)}, \ \forall s,t \in \mathbb{R} \text{ i.e. } X \perp\!\!\!\perp Y.$$

Independence

Proposition 4.37

Consider two random variables X and Y such that (X, Y) is a Gaussian vector. Then $r^\star(X, Y) = |\text{Corr[X,Y]}|$.

- See Lancaster (1957, 1958), and Gauss-Hermite decomposition

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2\left[1-\rho^2\right]}\right) = \phi(x)\phi(y) \cdot \sum_{i=0}^{\infty} r^i H_i(x) H_i(y)$$

where H_i's are Hermite polynomial.

Independence

- Instead of

$$r^\star(X,Y) = \max_{\varphi\in\mathcal{F}_x,\ \psi\in\mathcal{G}_y} \mathbb{E}[\varphi(X)\psi(Y)],$$

where

$$\begin{cases} \mathcal{F}_x = \{\varphi : \mathcal{X} \to \mathbb{R} : \mathbb{E}[\varphi(X)] = 0 \text{ and } \mathbb{E}[\varphi^2(X)] = 1\} \\ \mathcal{G}_y = \{\psi : \mathcal{Y} \to \mathbb{R} : \mathbb{E}[\psi(Y)] = 0 \text{ and } \mathbb{E}[\psi^2(Y)] = 1\} \end{cases}$$

Definition 4.59: Constrained Maximal Correlation, Bach and Jordan (2002), Gretton et al. (2005)

Consider two random variables X and Y, as well as some Hilbert spaces $\bar{\mathcal{F}}_x \subset \mathcal{F}_x$ and $\bar{\mathcal{G}}_y \subset \mathcal{G}_y$,

$$\bar{r}^\star(X,Y) = \max_{\varphi\in\bar{\mathcal{F}}_x,\psi\in\bar{\mathcal{G}}_y} \{\text{Corr}[\varphi(X),\psi(Y)]\}.$$

Independence

- Kimeldorf and Sampson (1978) and Kimeldorf et al. (1982) suggested to consider for $\bar{\mathcal{F}}_x$ and $\bar{\mathcal{G}}_y$ as subsets of monotone functions.

$$\begin{cases} \bar{\mathcal{F}}x = \{\varphi \in \mathcal{F}_x : \varphi \text{ monotone}\} \\ \bar{\mathcal{G}}_y = \{\psi \in \mathcal{G}_y : \psi \text{ monotone}\} \end{cases}$$

- See Mourier (1953), Hannan (1961), Jensen and Mayer (1977) and Lin (1987).

to go further ➔ (for more details on RKHS issues)

Independence

Definition 4.60: Linear Independence

In a general context, consider two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, in $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_y}$, respectively. $\boldsymbol{X} \perp \boldsymbol{Y}$ if and only if for any $\boldsymbol{a} \in \mathbb{R}^{d_x}$ and $\boldsymbol{b} \in \mathbb{R}^{d_y}$

$$\text{Cov}[\boldsymbol{a}^\top \boldsymbol{X}, \boldsymbol{b}^\top \boldsymbol{Y}] = 0.$$

Definition 4.61: Independence

In a general context, consider two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$. $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$ if and only if for any $\mathcal{A} \subset \mathbb{R}^{d_x}$ and $\mathcal{B} \subset \mathbb{R}^{d_y}$,

$$\mathbb{P}[\{\boldsymbol{X} \in \mathcal{A}\} \cap \{\boldsymbol{Y} \in \mathcal{B}\}] = \mathbb{P}[\{\boldsymbol{X} \in \mathcal{A}\}] \cdot \mathbb{P}[\{\boldsymbol{Y} \in \mathcal{B}\}].$$

Independence

Proposition 4.38: Independence

Consider two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$. $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$ if and only if for any functions $\varphi : \mathbb{R}^{d_x} \to \mathbb{R}$ and $\psi : \mathbb{R}^{d_y} \to \mathbb{R}$ (such that the expected values below exist and are well-defined)

$$\mathbb{E}[\varphi(\boldsymbol{X})\psi(\boldsymbol{Y})] = \mathbb{E}[\varphi(\boldsymbol{X})] \cdot \mathbb{E}[\psi(\boldsymbol{Y})],$$

or equivalently

$$\text{Cov}[\varphi(\boldsymbol{X}), \psi(\boldsymbol{Y})] = 0.$$

Independence

Definition 4.62: Mutual Independence

Let $\boldsymbol{Y} = (Y_1, \cdots, Y_k)$ denote some random vector. All components of $\boldsymbol{Y}$ are (mutually) independent if for any $\mathcal{A}_1, \cdots, \mathcal{A}_k \subset \mathbb{R}$

$$\mathbb{P}\left[\{(Y_1, \cdots, Y_k) \in \bigcap_{i=1}^{k} \mathcal{A}_i\}\right] = \prod_{i=1}^{k} \mathbb{P}[\{Y_i \in \mathcal{A}_i\}].$$

Definition 4.63: Conditional Independence (dimension 2)

X and Y are independent conditionally on Z, denoted $X \perp\!\!\!\perp Y \mid Z$, if for any sets $\mathcal{A}, \mathcal{B}, \mathcal{C} \subset \mathbb{R}$,

$$\mathbb{P}[X \in \mathcal{A}, Y \in \mathcal{B} | Z \in \mathcal{C}] = \mathbb{P}[X \in \mathcal{A} | Z \in \mathcal{C}] \cdot \mathbb{P}[Y \in \mathcal{B} | Z \in \mathcal{C}].$$

Independence

Definition 4.64: Conditional Independence

In a general context, consider three random vectors $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$. $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y})|\boldsymbol{Z}$ if and only if for any $\mathcal{A} \subset \mathbb{R}^{d_x}$, $\mathcal{B} \subset \mathbb{R}^{d_y}$ and $\mathcal{C} \subset \mathbb{R}^{d_z}$,

$$\mathbb{P}[\{\boldsymbol{X} \in \mathcal{A}\} \cap \{\boldsymbol{Y} \in \mathcal{B}\}|\boldsymbol{Z} \in \mathcal{C}] = \mathbb{P}[\{\boldsymbol{X} \in \mathcal{A}\}|\boldsymbol{Z} \in \mathcal{C}] \cdot \mathbb{P}[\{\boldsymbol{Y} \in \mathcal{B}\}|\boldsymbol{Z} \in \mathcal{C}].$$

Proposition 4.39

Consider three random variables X, Y, and Z. If $X \perp Z$ and $Y \perp Z$, then $aX + bY \perp Z$, for any $a, b \in \mathbb{R}$.

Independence

Proposition 4.40: $X \perp Z,\ Y \perp Z \not\Longrightarrow \psi(X, Y) \perp Z$

Consider three random variables X, Y, and Z. If $X \perp Z$ and $Y \perp Z$, it does not imply that $\psi(X, Y) \perp Z$, for any $\psi : \mathbb{R}^2 \to \mathbb{R}$.

$$(X, Y, Z) = \begin{cases} (0,0,0) & \text{with probability } 1/4, \\ (0,1,1) & \text{with probability } 1/4, \\ (1,0,1) & \text{with probability } 1/4, \\ (1,1,0) & \text{with probability } 1/4. \end{cases}$$

Proposition 4.41

Consider a random vector $\boldsymbol{X}$ in $\mathbb{R}^k$, and a random variable Z.
$\boldsymbol{X} \perp Z$ does not imply that $\psi(\boldsymbol{X}) \perp Z$, for any $\psi : \mathbb{R}^k \to \mathbb{R}$.

Independence

Proposition 4.42

Consider three random variables X, Y, and Z. Even if $X \perp\!\!\!\perp Z$ and $Y \perp\!\!\!\perp Z$, it does not imply either that $\psi(X, Y) \perp Z$ or that $\psi(X, Y) \perp\!\!\!\perp Z$, for any $\psi : \mathbb{R}^2 \to \mathbb{R}$.

Proposition 4.43

Consider a random vector $\boldsymbol{X}$ in $\mathbb{R}^k$, and a random variable Z.
$\boldsymbol{X} \perp\!\!\!\perp Z$ does not imply either that $\psi(\boldsymbol{X}) \perp Z$ **or** $\psi(\boldsymbol{X}) \perp\!\!\!\perp Z$, for any $\psi : \mathbb{R}^k \to \mathbb{R}$.

Causation

Definition 4.65: Common cause, Reichenbach (1956)

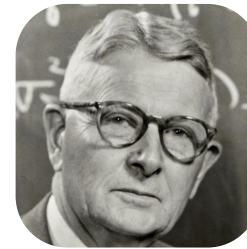
If X and Y are non-independent, $X \not\perp\!\!\!\perp Y$, then, either

$$\begin{cases} X \text{ causes } Y \\ Y \text{ causes } X \\ \text{there exists } Z \text{ such that } Z \text{ causes both } X \text{ and } Y. \end{cases}$$

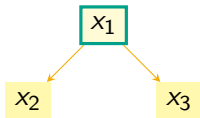
- See also Bollen and Pearl (2013)
- SCM, Goldberger (1972), Duncan (1975) or Bollen (1989)
- Bayesian network, Pearl (1985), Henrion (1988), Charniak (1991)
- Causal path diagrams and probabilistic DAGs, Spirtes et al. (1993)

Causation

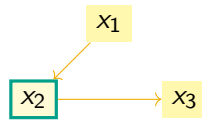
Sewall Wright (see Wright (1921, 1934)) use directed graphs to represent probabilistic cause and effect relationships among a set of variables, and developed path diagrams and path analysis



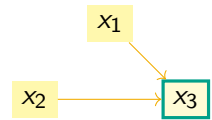
(a)
confounder



(b)
mediator



(c)
collider



Causation

Definition 4.66: Path

A path π from a node x_i to another node x_j is a sequence of nodes and edges starting at x_i and ending at x_j.

Definition 4.67: d-separation

A set of nodes $\boldsymbol{x}_i$ is said to be d-separated with another set of nodes $\boldsymbol{x}_j$ by $\boldsymbol{x}_c$ whenever every path from any $x_i \in \boldsymbol{x}_i$ to any $x_j \in \boldsymbol{x}_j$ is blocked by $\boldsymbol{x}_c$. We will simply denote $\boldsymbol{x}_i \perp_{\mathcal{G}} \boldsymbol{x}_j \mid \boldsymbol{x}_c$.

Proposition 4.44

Two nodes x_i and x_j are d-separated by $\boldsymbol{x}_c$ if and only members of $\boldsymbol{x}_c$ block all paths from x_i to x_j.

Causation

- Chain rule : $\begin{cases} \mathbb{P}[x_1, x_2, x_3, x_4] = \mathbb{P}[x_1] \times \mathbb{P}[x_2|x_1] \times \mathbb{P}[x_3|x_1, x_2] \times \mathbb{P}[x_4|x_1, x_2, x_3] \\ \mathbb{P}[x_1, x_2, x_3, x_4] = \mathbb{P}[x_4] \times \mathbb{P}[x_3|x_4] \times \mathbb{P}[x_2|x_3, x_4] \times \mathbb{P}[x_1|x_2, x_3, x_4] \end{cases}$

Definition 4.68: Directed acyclic graph, DAG (or causal graph)

A directed acyclic graph (DAG) $\mathcal{G}$ is a directed graph with no directed cycles.

Definition 4.69: Markov Property

Given a causal graph $\mathcal{G}$ with nodes $\boldsymbol{x}$, the joint distribution of $\boldsymbol{X}$ satisfies the (global) Markov property with respect to $\mathcal{G}$ if, for any disjoints $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_c$

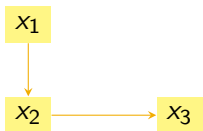
$$\boldsymbol{x}_1 \perp_{\mathcal{G}} \boldsymbol{x}_2 \mid \boldsymbol{x}_c \;\Rightarrow\; \boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{X}_2 \mid \boldsymbol{X}_c.$$

Proposition 4.45: Probabilistic graphical model

If $\boldsymbol{X}$ satisfies the (global) Markov property with respect to $\mathcal{G}$

$$\mathbb{P}[x_1, \cdots, x_n] = \prod_{i=1}^{n} \mathbb{P}[x_i | \text{parents}(x_i)]$$

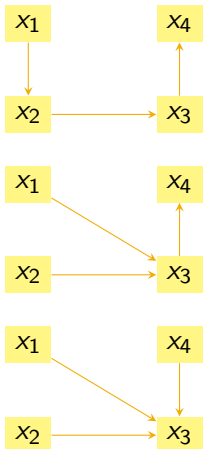
where parents(x_i) are nodes with edges directed towards x_i



› Path from x_1 to x_3 is blocked by x_2, i.e., $x_1 \perp_{\mathcal{G}} x_3 \mid x_2$, or $X_1 \perp\!\!\!\perp X_3 \mid X_2$. From the chain rule,

$$\mathbb{P}[x_1, x_2, x_3] = \mathbb{P}[x_1] \times \mathbb{P}[x_2|x_1] \times \underbrace{\mathbb{P}[x_3|x_2, x_1]}_{\mathbb{P}[x_3|x_2]}$$

Causation



- From the chain rule, for the causal graph on the left (top),

$$\mathbb{P}[x_1, x_2, x_3, x_4] = \mathbb{P}[x_1] \times \mathbb{P}[x_2|x_1] \times \mathbb{P}[x_3|x_2] \times \mathbb{P}[x_4|x_3]$$

- From the chain rule, for the causal graph on the left (middle),

$$\mathbb{P}[x_1, x_2, x_3, x_4] = \mathbb{P}[x_1] \times \mathbb{P}[x_2] \times \mathbb{P}[x_3|x_1, x_2] \times \mathbb{P}[x_4|x_3]$$

- From the chain rule, for the causal graph on the left (bottom),

$$\mathbb{P}[x_1, x_2, x_3, x_4] = \mathbb{P}[x_1] \times \mathbb{P}[x_2] \times \mathbb{P}[x_3|x_1, x_2, x_4] \times \mathbb{P}[x_4]$$

Intervention

- $\mathbb{P}[Y \in \mathcal{A}|X = x]$: how $Y \in \mathcal{A}$ is likely to occur if X happened to be equal to x
- Therefore, it is an observational statement.

- $P[Y \in \mathcal{A}|\text{do}(X = x)]$: how $Y \in \mathcal{A}$ is likely to occur if X is set to x
- It is here an intervention statement.

- Using causal graphs, intervention $\text{do}(X = x)$ means that all incoming edges to x are cut.

- If $P[Y \in \mathcal{A}|\text{do}(X = x)] \neq \mathbb{P}[Y \in \mathcal{A}|X = x]$, it means that X and Y are confounded, see Pearl (2009).

Intervention

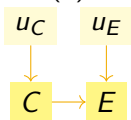
Definition 4.70: Structural Causal Models (SCM)

In a simple causal graph, with two nodes C (the cause) and E (the effect), the causal graph is $C \to E$, and the mathematical interpretation can be summarized in two assignments

$$\begin{cases} C = h_c(U_C) \\ E = h_e(C, U_E), \end{cases}$$

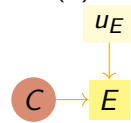
where U_C and U_E are two independent random variables, $U_C \perp\!\!\!\perp U_E$.

(a) observation



$$\begin{cases} C = h_c(U_C) \\ E = h_e(C, U_E) \end{cases}$$

(b) intervention



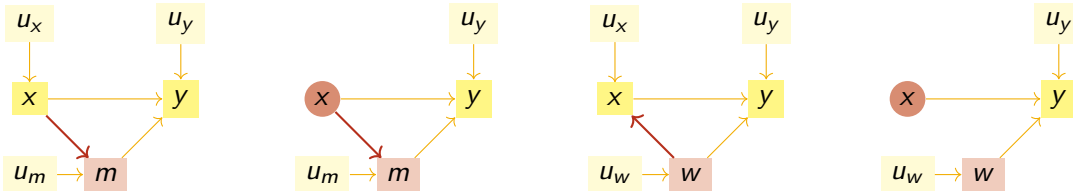
$$\begin{cases} C = c \quad (\text{or } \mathrm{do}(C = c)) \\ E_c^\star = h_e(c, U_E) \end{cases}$$

Intervention

(a) (b) (c) (d)

m mediator variable

w confounding variable



$$\begin{cases} \text{mediator}: & \mathbb{P}[Y_x^\star = 1] = \mathbb{P}[Y = 1|\text{do}(X = x)] = \mathbb{P}[Y = 1|X = x] \\ \text{confusion}: & \mathbb{P}[Y_x^\star = 1] = \mathbb{P}[Y = 1|\text{do}(X = x)] \neq \mathbb{P}[Y = 1|X = x]. \end{cases}$$

Intervention

› In fact, in the presence of a confounding factor, $\mathbb{P}[Y_x^\star = 1]$ which corresponds to $\mathbb{P}[Y = 1|\text{do}(X = x)]$ should be written

$$\sum_w \mathbb{P}[Y = 1|W = w, X = x] \cdot \mathbb{P}[W = w] = \mathbb{E}(\mathbb{P}[Y = 1|W, X = x]).$$

Causal Inference and counterfactuals

- Define potential outcomes to quantify the treatment effect, $\text{TE} = y^\star_{i,T\leftarrow 1} - y^\star_{i,T\leftarrow 0}$

$$\begin{cases} \text{observation} & : \ y^\star_{i,T\leftarrow 1} \text{ when } t_i = 1 \text{ is observed, and } \boldsymbol{x}_i \\ \text{counterfactual} & : \ y^\star_{i,T\leftarrow 0} \text{ when } t_i = 1 \text{ is observed, and } \boldsymbol{x}_i \end{cases}$$

- Here we want to observe counterfactuals $y^\star_{i,T\leftarrow t'}$ at the individual level.

	Gender	Name	Treatment	Outcome (Weight)				Height	⋯
			t_i	y_i	$y^\star_{i,T\leftarrow 0}$	$y^\star_{i,T\leftarrow 1}$	TE	x_i	⋯
1	H	Alex	0	75	75	64	11	172	⋯
2	F	Betty	1	52	67	52	15	161	⋯
3	F	Beatrix	1	57	71	57	14	163	⋯
4	H	Ahmad	0	78	78	61	17	183	⋯

- Different notations are used $y(1)$ and $y(0)$ in Imbens and Rubin (2015), y^1 and y^0 in Cunningham (2021), or $y_{t=1}$ and $y_{t=0}$ in Pearl and Mackenzie (2018).

Causal Inference and counterfactuals

- Define potential outcomes to quantify the treatment effect, $\text{TE} = y^\star_{i,T\leftarrow 1} - y^\star_{i,T\leftarrow 0}$

$$\begin{cases} \text{observation} & : \ y^\star_{i,T\leftarrow 1} \text{ when } t_i = 1 \text{ is observed, and } \boldsymbol{x}_i \\ \text{counterfactual} & : \ y^\star_{i,T\leftarrow 0} \text{ when } t_i = 1 \text{ is observed, and } \boldsymbol{x}_i \end{cases}$$

- Here we want to observe counterfactuals $y^\star_{i,T\leftarrow t'}$ at the individual level.

	Gender	Name	Treatment	Outcome (Weight)				Height	⋯
			t_i	y_i	$y^\star_{i,T\leftarrow 0}$	$y^\star_{i,T\leftarrow 1}$	TE	x_i	⋯
1	H	Alex	0	75	75	?	?	172	⋯
2	F	Betty	1	52	?	52	?	161	⋯
3	F	Beatrix	1	57	?	57	?	163	⋯
4	H	Ahmad	0	78	78	?	?	183	⋯

- Different notations are used $y(1)$ and $y(0)$ in Imbens and Rubin (2015), y^1 and y^0 in Cunningham (2021), or $y_{t=1}$ and $y_{t=0}$ in Pearl and Mackenzie (2018).

Causal Inference and counterfactuals

Definition 4.71: Average Treatment Effect, Holland (1986)

Given a treatment T, the average treatment effect on outcome Y is

$$\tau = \text{ATE} = \mathbb{E}\big[Y^\star_{t\leftarrow 1} - Y^\star_{t\leftarrow 0}\big].$$

Definition 4.72: Conditional Average Treatment Effect, Wager and Athey (2018)

Given a treatment T, the conditional average treatment effect on outcome Y, given some covariates $\boldsymbol{X}$,is

$$\tau(\boldsymbol{x}) = \text{CATE}(\boldsymbol{x}) = \mathbb{E}\big[Y^\star_{t\leftarrow 1} - Y^\star_{t\leftarrow 0}\big|\boldsymbol{X} = \boldsymbol{x}\big].$$

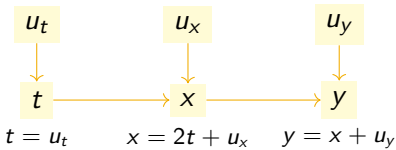
Definition 4.73: Individual Average Treatment Effect

Given a treatment T, the conditional average treatment effect on outcome Y, for individual i, given covariates $\boldsymbol{X}_i$, is

$$\text{IATE}(i) = \mathbb{E}\big[Y^\star_{i,t\leftarrow(1-t_i)} - Y^\star_{i,t\leftarrow t_i}\big].$$

Causal Inference and counterfactuals

- Pearl (2009) suggest to use a twin network representation of the counterfactual



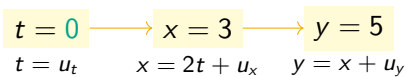
- Start with a simple structural causal model, e.g.,

$$\begin{cases} t = u_t \\ x = 2t + u_t \\ y = x + u_y \end{cases}$$

- Suppose we were able to estimate that model

Causal Inference and counterfactuals

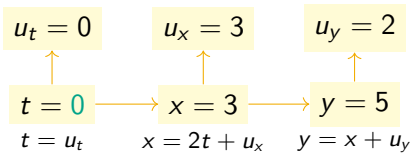
- Pearl (2009) suggest to use a twin network representation of the counterfactual



- Consider a single observation
- i.e. a triplet $(t_i, \boldsymbol{x}_i, y_i)$

Causal Inference and counterfactuals

- Pearl (2009) suggest to use a twin network representation of the counterfactual



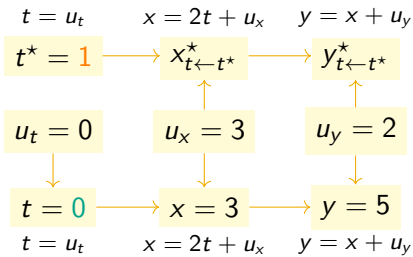
- Inverse the SCM

$$\begin{cases} u_t = t \\ u_t = x - 2t \\ u_y = y - x \end{cases}$$

- From that triplet $(t_i, \boldsymbol{x}_i, y_i)$
- Derive unobserved u's.

Causal Inference and counterfactuals

- Pearl (2009) suggest to use a twin network representation of the counterfactual

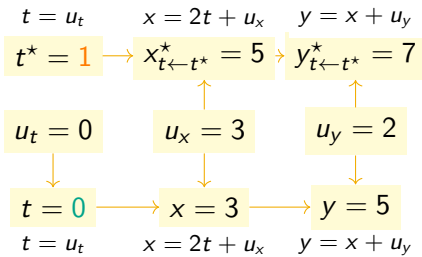


- Suppose that the same SCM holds in the counterfactual world

$$\begin{cases} t^\star \\ x^\star_{t\leftarrow t^\star} = 2t^\star + u_t \\ y^\star_{t\leftarrow t^\star} = x^\star_{t\leftarrow t^\star} + u_y \end{cases}$$

Causal Inference and counterfactuals

- Pearl (2009) suggest to use a twin network representation of the counterfactual

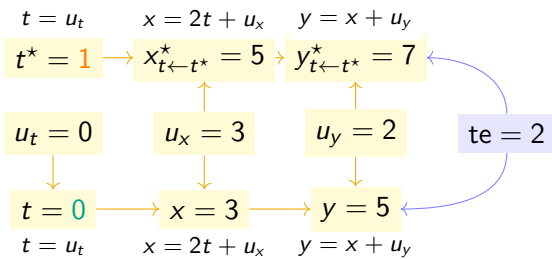


- Plugin u's obtained from $(t_i, \boldsymbol{x}_i, y_i)$

$$\begin{cases} t^\star \\ x^\star_{t\leftarrow t^\star} = 2t^\star + u_t \\ y^\star_{t\leftarrow t^\star} = x^\star_{t\leftarrow t^\star} + u_y \end{cases}$$

Causal Inference and counterfactuals

- Pearl (2009) suggest to use a twin network representation of the counterfactual

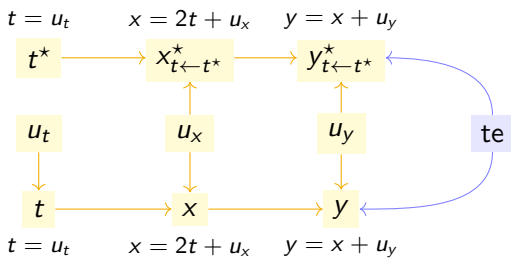


- We can compute the treatment effect

$$\text{te} = y - y^\star_{t\leftarrow t^\star}$$

Causal Inference and counterfactuals

- Pearl (2009) suggest to use a twin network representation of the counterfactual



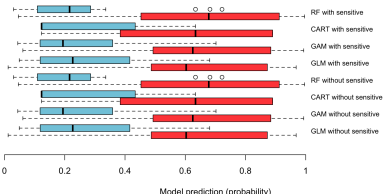
- Works well only if the SCM can be
- inverted (to derive the $\boldsymbol{u}$'s)

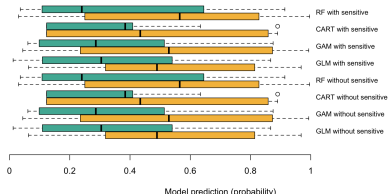
– Part 4 –

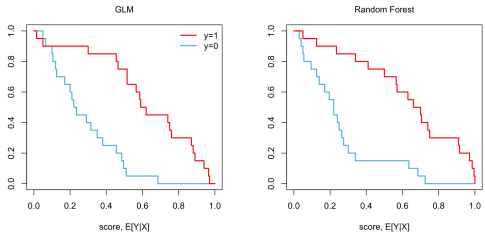
Group Fairness

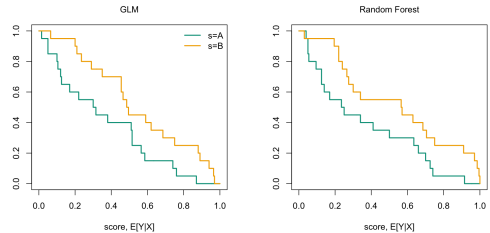
Group Fairness

- Back on `toydata2`, distributions of scores, $\widehat{m}(\boldsymbol{x}_i)$'s conditional on y_i and s_i









Group Fairness

Definition 5.1: Fairness through unawareness, Dwork et al. (2012)

A model m satisfies the fairness through unawareness criteria, with respect to sensitive attribute $s \in \mathcal{S}$ if $m : \mathcal{X} \to \mathcal{Y}$.

by Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold and Richard Zemel,





Group Fairness

- See introduction about the gender directive,

"*institutional messages of color blindness may therefore artificially depress formal reporting of racial injustice. Color-blind messages may thus appear to function effectively on the surface even as they allow explicit forms of bias to persist*," Apfelbaum et al. (2010)

Definition 5.2: Aware and unaware regression functions μ

The aware regression function is $\mu(\boldsymbol{x}, s) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, S = s]$
and the unaware regression function is $\mu(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$.

Historical Perspective: "Cultural Fairness" and "Statistical Discrimination"

Definition 5.3: Four definitions of cultural fairness, Darlington (1971)

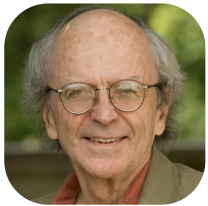
A test $(\widehat{y})$ is considered "culturally fair" if it fits the appropriate equation

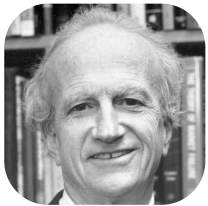
$$\begin{cases} \text{Cor}[S, \widehat{Y}] = \text{Cor}[S, Y]/\text{Cor}[Y, \widehat{Y}] \\ \text{Cor}[S, \widehat{Y}] = \text{Cor}[S, Y] \\ \text{Cor}[S, \widehat{Y}] = \text{Cor}[S, Y] \cdot \text{Cor}[Y, \widehat{Y}] \\ \text{Cor}[S, \widehat{Y}] = 0 \end{cases}$$

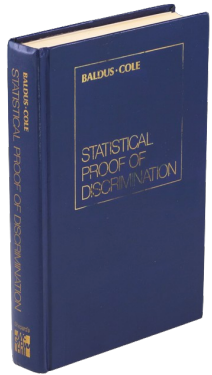
See also Thorndike (1971), Linn and Werts (1971), following Cleary (1968).

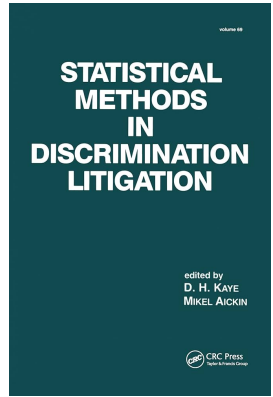
"Economics of Discrimination" and "Statistical Discrimination"

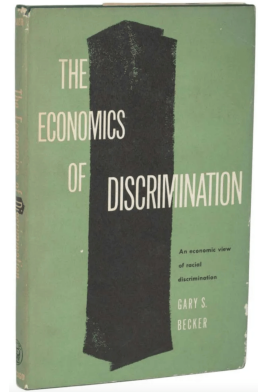
- See Becker (1957) or Baldus and Cole (1980), among (many) others.











Historical Perspective: Decomposition

$$\begin{cases} y_{A:i} = \boldsymbol{x}_{A:i}^\top \boldsymbol{\beta}_A + \varepsilon_{A:i} \quad \text{(group A)}, \ \overline{y}_A = \overline{\boldsymbol{x}}_A^\top \widehat{\boldsymbol{\beta}}_A \\ y_{B:i} = \boldsymbol{x}_{B:i}^\top \boldsymbol{\beta}_B + \varepsilon_{B:i} \quad \text{(group B)}, \ \overline{y}_B = \overline{\boldsymbol{x}}_B^\top \widehat{\boldsymbol{\beta}}_B. \end{cases}$$

- Using ordinary least squares estimates

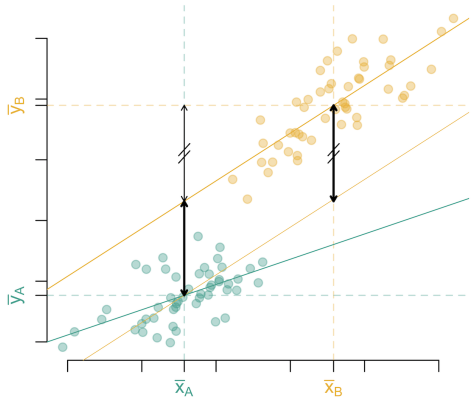
Definition 5.4: Kitagawa (1955), Oaxaca (1973), Blinder (1973)

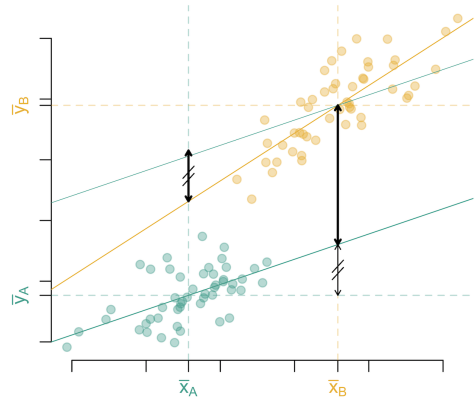
$$\overline{y}_A - \overline{y}_B = \underbrace{\left(\overline{\boldsymbol{x}}_A - \overline{\boldsymbol{x}}_B\right)^\top \widehat{\boldsymbol{\beta}}_B}_{\text{characteristics}} + \underbrace{\overline{\boldsymbol{x}}_A^\top \left(\widehat{\boldsymbol{\beta}}_A - \widehat{\boldsymbol{\beta}}_B\right)}_{\text{coefficients}}, \tag{5}$$

$$\overline{y}_A - \overline{y}_B = \underbrace{\left(\overline{\boldsymbol{x}}_A - \overline{\boldsymbol{x}}_B\right)^\top \widehat{\boldsymbol{\beta}}_A}_{\text{characteristics}} + \underbrace{\overline{\boldsymbol{x}}_B^\top \left(\widehat{\boldsymbol{\beta}}_A - \widehat{\boldsymbol{\beta}}_B\right)}_{\text{coefficients}}. \tag{6}$$

- Also Brown et al. (1980) and Conway and Roberts (1983).

Historical Perspective: Decomposition





$x_A(\widehat{\beta}_A - \widehat{\beta}_B)$ and $(\overline{x}_A - \overline{x}_B)\widehat{\beta}_B$ (as in Equation 5) on the left
$x_B(\widehat{\beta}_A - \widehat{\beta}_B)$ and $(\overline{x}_A - \overline{x}_B)\widehat{\beta}_A$ (as in Equation 6) on the right.

Independence and Demographic Parity

Definition 5.5: Independence, Barocas et al. (2017)

A model m satisfies the independence property if $m(\boldsymbol{Z}) \perp\!\!\!\perp S$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$.

by Solon Barocas, Moritz Hardt and Arvind Narayanan



› For classifiers, one might ask for independence $\widehat{Y} \perp\!\!\!\perp S$ (where $\widehat{y}$ is a class), as Darlington (1971).

Independence and Demographic Parity

Definition 5.6: Demographic Parity, Calders and Verwer (2010), Corbett-Davies et al. (2017)

A decision function $\widehat{y}$ – or a classifier m_t, taking values in $\{0,1\}$ – satisfies demographic parity, with respect to some sensitive attribute S if (equivalently)

$$\begin{cases} \mathbb{P}[\widehat{Y}=1|S=\mathtt{A}] = \mathbb{P}[\widehat{Y}=1|S=\mathtt{B}] = \mathbb{P}[\widehat{Y}=1] \\ \mathbb{E}[\widehat{Y}|S=\mathtt{A}] = \mathbb{E}[\widehat{Y}|S=\mathtt{B}] = \mathbb{E}[\widehat{Y}] \\ \mathbb{P}[m_t(\boldsymbol{Z})=1|S=\mathtt{A}] = \mathbb{P}[m_t(\boldsymbol{Z})=1|S=\mathtt{B}] = \mathbb{P}[m_t(\boldsymbol{Z})=1]. \end{cases}$$

by Toon Calders, Sicco Verwer, Sam Corbett-Davies, Emma Pierson, Sharad Goel, etc

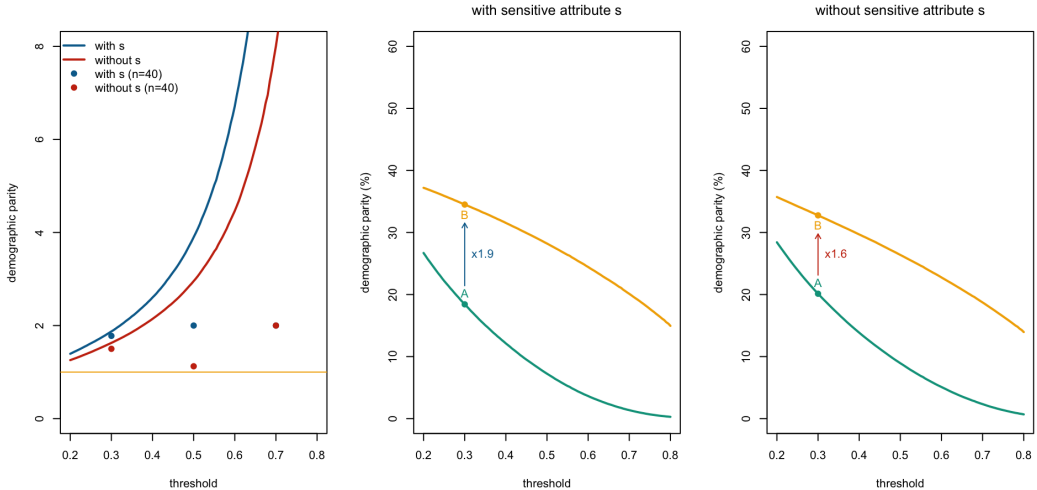
Independence and Demographic Parity

	unaware (without s)				aware (with s)			
	GLM	GAM	CART	RF	GLM	GAM	CART	RF
$n = 1000$, various t, ratio $\mathbb{P}[\widehat{Y} = 1 \vert S = \mathtt{B}]/\mathbb{P}[\widehat{Y} = 1 \vert S = \mathtt{A}]$								
$t = 30\%$	1.652	1.519	1.235	1.559	1.918	1.714	1.235	1.798
$t = 50\%$	1.877	2.451	2.918	2.404	2.944	3.457	2.918	2.180
$t = 70\%$	6.033	8.711	26.000	4.621	7.917	19.333	26.000	4.578

(`dem_parity` from R package `fairness`)

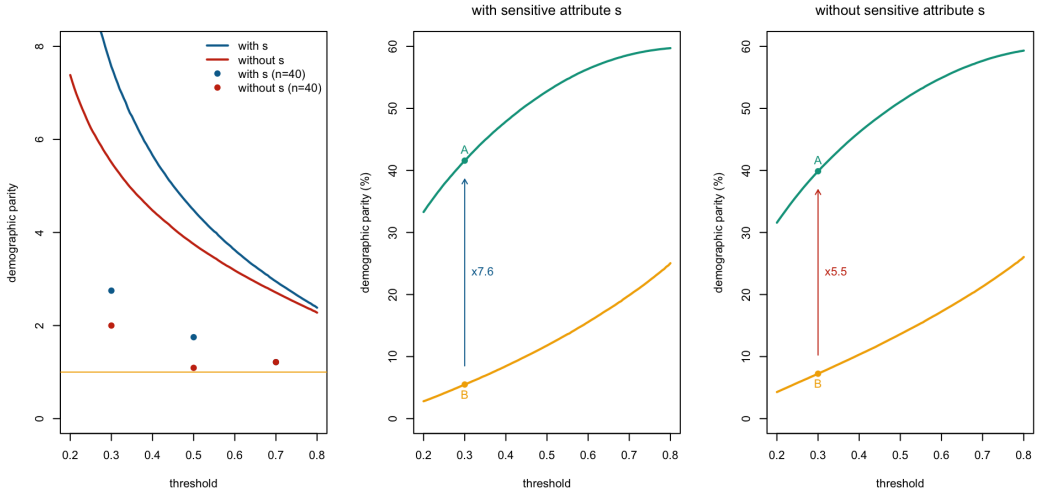
› On the left-hand side, evolution of the ratio ratio $\mathbb{P}[\widehat{Y} = 1 \vert S = \mathtt{B}]/\mathbb{P}[\widehat{Y} = 1 \vert S = \mathtt{A}]$. The horizontal line (at $y = 1$) corresponds to perfect demographic parity. In the middle $t \mapsto \mathbb{P}[m_t(\boldsymbol{X}) > t \vert S = \mathtt{B}]$ and $t \mapsto \mathbb{P}[m_t(\boldsymbol{X}) > t \vert S = \mathtt{A}]$ on the model with s, and on the right-hand side without s.

Independence and Demographic Parity



On the left-hand side, evolution of the ratio ratio $\mathbb{P}[\widehat{Y} = 1|S = \texttt{B}]/\mathbb{P}[\widehat{Y} = 1|S = \texttt{A}]$.

Independence and Demographic Parity



- On the left-hand side, evolution of the ratio ratio $\mathbb{P}[\widehat{Y} = 0|S = \mathtt{A}]/\mathbb{P}[\widehat{Y} = 0|S = \mathtt{B}]$

Independence and Demographic Parity

Definition 5.7: Weak Demographic Parity

A decision function $\widehat{y}$ satisfies weak demographic parity if

$$\mathbb{E}[\widehat{Y}|S = \texttt{A}] = \mathbb{E}[\widehat{Y}|S = \texttt{B}].$$

Definition 5.8: Strong Demographic Parity

A decision function $\widehat{y}$ satisfies demographic parity if $\widehat{Y} \perp\!\!\!\perp S$, i.e., for all A,

$$\mathbb{P}[\widehat{Y} \in \mathcal{A}|S = \texttt{A}] = \mathbb{P}[\widehat{Y} \in \mathcal{A}|S = \texttt{B}], \ \forall \mathcal{A} \subset \mathcal{Y}.$$

Independence and Demographic Parity

Proposition 5.1

A model m satisfies the strong demographic parity property if and only if

$$d_{\mathrm{TV}}(\mathbb{P}_{m|\mathtt{A}}, \mathbb{P}_{m|\mathtt{B}}) = d_{\mathrm{TV}}(\mathbb{P}_{\mathtt{A}}, \mathbb{P}_{\mathtt{B}}) = 0.$$

› $d_{\mathrm{TV}}(\mathbb{P}_{m|\mathtt{A}}, \mathbb{P}_{m|\mathtt{B}})$ could be seen as a measure of "unfairness", but for a non-binary sensitive attribute, a more general definition is necessary (see Denis et al. (2021)).

Independence and Demographic Parity

Definition 5.9: Conditional demographic parity, Corbett-Davies et al. (2017)

We will have a conditional demographic parity if (at choice) for all $\boldsymbol{x}$,

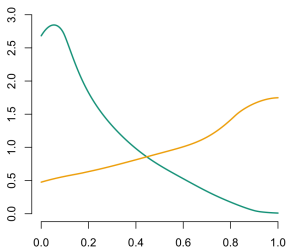
$$\begin{cases} \mathbb{P}[\widehat{Y}=1|\boldsymbol{X}_L=\boldsymbol{x},S=\mathtt{A}]=\mathbb{P}[\widehat{Y}=1|\boldsymbol{X}_L=\boldsymbol{x},S=\mathtt{B}],\ \forall y\in\{0,1\} \\ \mathbb{E}[\widehat{Y}|\boldsymbol{X}_L=\boldsymbol{x},S=\mathtt{A}]=\mathbb{E}[\widehat{Y}|\boldsymbol{X}_L=\boldsymbol{x},S=\mathtt{B}], \\ \mathbb{P}[\widehat{Y}\in\mathcal{A}|\boldsymbol{X}_L=\boldsymbol{x},S=\mathtt{A}]=\mathbb{P}[\widehat{Y}\in\mathcal{A}|\boldsymbol{X}_L=\boldsymbol{x},S=\mathtt{B}],\ \forall\mathcal{A}\subset\mathcal{Y}, \end{cases}$$

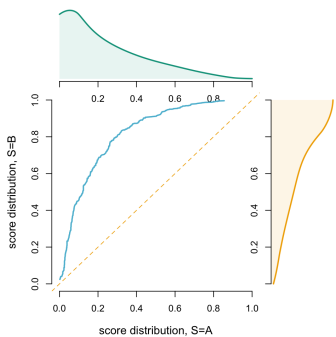
where L denotes a "legitimate" subset of unprotected covariates.

Independence and Demographic Parity

Proposition 5.2

A model m satisfies is strongly fair if and only if $W_2(\mathbb{P}_A, \mathbb{P}_B) = 0$.

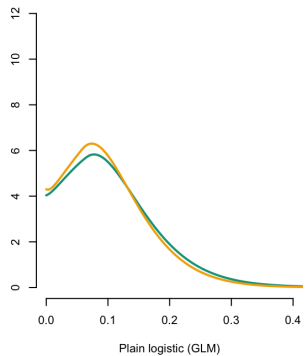


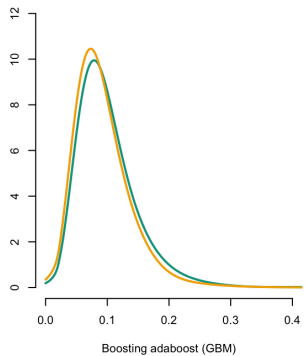


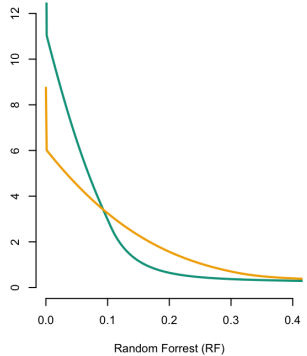

```
> model_glm = glm(y~x1
  +x2+x3, data=
  toydata2, family=
  binomial)
> pred_y_glm = predict
  (model_glm, type="
  response")
> sA = pred_y_glm[
  toydata2$sensitive
  =="A"]
> library(transport)
> wasserstein1d(sA,sB)
[1] 0.3860795
```

Independence and Demographic Parity

- On the `FrenchMotor` dataset, consider GLM, GBM and RF for claim occurence








```
> wasserstein1d(lA,lB)
[1] 0.007220468
```

```
> wasserstein1d(bA,bB)
[1] 0.008895917
```

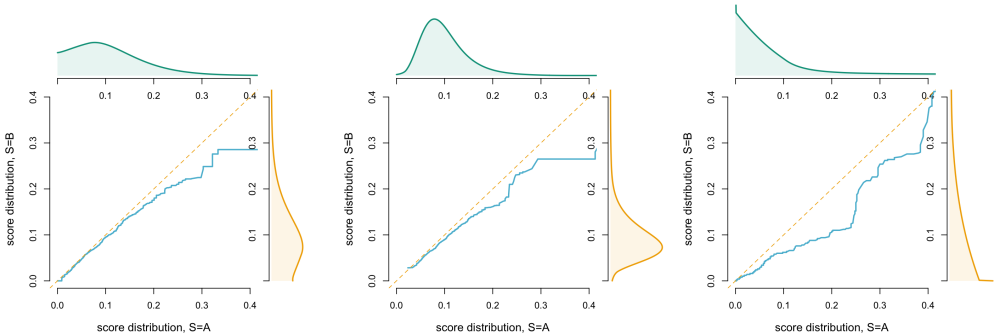
```
> wasserstein1d(fA,fB)
[1] 0.01001088
```

Independence and Demographic Parity

```
> wasserstein1d(lA,lB)
[1] 0.007220468
```

```
> wasserstein1d(bA,bB)
[1] 0.008895917
```

```
> wasserstein1d(fA,fB)
[1] 0.01001088
```



Independence and Demographic Parity

Definition 5.10: Unfairness, Denis et al. (2021); Chzhen and Schreuder (2022)

Given a model m, let $\mathbb{P}_m$ denote the distribution of $m(\boldsymbol{X}, S)$ and $\mathbb{P}_{m|s}$ denote the conditional distribution of $m(\boldsymbol{X}, S)$ given $S = s$, define

$$\begin{cases} \mathcal{U}_{\mathrm{TV}}(m) = \max\limits_{s\in\{\mathtt{A},\mathtt{B}\}} \left\{ d_{\mathrm{TV}}(\mathbb{P}_m, \mathbb{P}_{m|s}) \text{ or } \sum\limits_{s\in\{\mathtt{A},\mathtt{B}\}} d_{\mathrm{TV}}(\mathbb{P}_m, \mathbb{P}_{m|s}) \right. \\ \mathcal{U}_{\mathrm{KS}}(m) = \max\limits_{s\in\{\mathtt{A},\mathtt{B}\}} \left\{ d_{\mathrm{KS}}(\mathbb{P}_m, \mathbb{P}_{m|s}) \right\} \text{ or } \sum\limits_{s\in\{\mathtt{A},\mathtt{B}\}} d_{\mathrm{KS}}(\mathbb{P}_m, \mathbb{P}_{m|s}) \\ \mathcal{U}_{\mathrm{W}_k}(m) = \max\limits_{s\in\{\mathtt{A},\mathtt{B}\}} \left\{ W_k(\mathbb{P}_m, \mathbb{P}_{m|s}) \right\} \text{ or } \sum\limits_{s\in\{\mathtt{A},\mathtt{B}\}} W_k(\mathbb{P}_m, \mathbb{P}_{m|s}) \end{cases}$$

› In the original version, Chzhen and Schreuder (2022) suggested to use the one on the right.

Independence and Demographic Parity

- Those measures characterize strong demographic parity,

Proposition 5.3: Strong Demographic Parity

A model m is strongly fair if and only if $\mathcal{U}(m) = 0$.

Separation and Equalized Odds

Definition 5.11: Separation, Barocas et al. (2017)

A model $m : \mathcal{Z} \to \mathcal{Y}$ satisfies the separation property if $m(\boldsymbol{Z}) \perp\!\!\!\perp S \mid Y$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$.

by Solon Barocas, Moritz Hardt and Arvind Narayanan



Separation and Equalized Odds

Definition 5.12: True positive equality, (Weak) Equal Opportunity, Hardt et al. (2016)

A decision function $\widehat{y}$ – or a classifier $m_t(\cdot)$, taking values in $\{0,1\}$ – satisfies equal opportunity, with respect to some sensitive attribute S if

$$\begin{cases} \mathbb{P}[\widehat{Y}=1|S=\mathtt{A},Y=1]=\mathbb{P}[\widehat{Y}=1|S=\mathtt{B},Y=1]=\mathbb{P}[\widehat{Y}=1|Y=1] \\ \mathbb{P}[m_t(\boldsymbol{Z})=1|S=\mathtt{A},Y=1]=\mathbb{P}[m_t(\boldsymbol{Z})=1|S=\mathtt{B},Y=1]=\mathbb{P}[m_t(\boldsymbol{Z})=1|Y=1], \end{cases}$$

which corresponds to parity of true positives, in the two groups, $\{\mathtt{A},\mathtt{B}\}$.

Separation and Equalized Odds

Definition 5.13: Strong Equal Opportunity

A classifier $m(\cdot)$, taking values in $\{0, 1\}$, satisfies equal opportunity, with respect to some sensitive attribute S if

$$\mathbb{P}[m(\boldsymbol{X}, S) \in \mathcal{A} | S = \texttt{A}, Y = 1] = \mathbb{P}[m(\boldsymbol{X}, S) \in \mathcal{A} | S = \texttt{B}, Y = 1] = \mathbb{P}[m(\boldsymbol{X}, S) \in \mathcal{A} | Y = 1]$$

for all $\mathcal{A} \subset [0, 1]$.

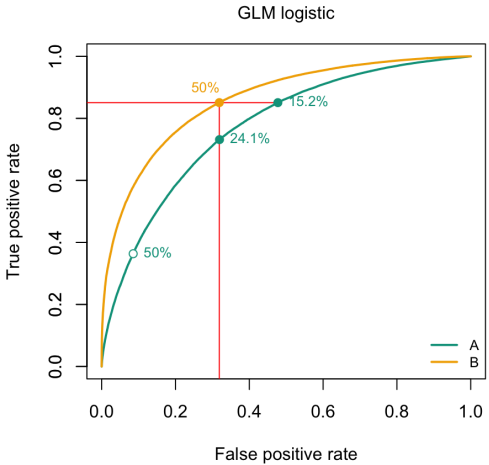
Separation and Equalized Odds

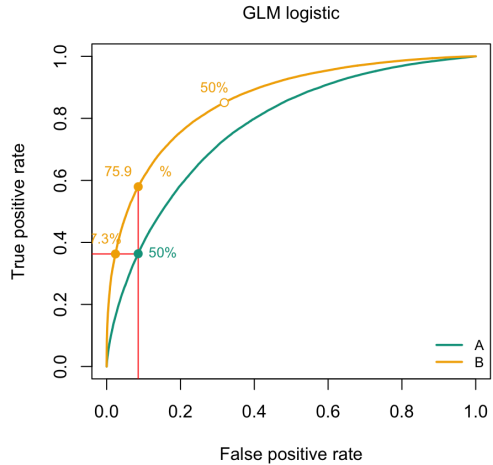
Definition 5.14: False positive equality, Hardt et al. (2016)

A decision function $\widehat{y}$ – or a classifier $m_t(\cdot)$, taking values in $\{0,1\}$ – satisfies parity of false positives, with respect to some sensitive attribute s, if

$$\begin{cases} \mathbb{P}[\widehat{Y}=1|S=\texttt{A},Y=0]=\mathbb{P}[\widehat{Y}=1|S=\texttt{B},Y=0]=\mathbb{P}[\widehat{Y}=1|Y=0] \\ \mathbb{P}[m_t(\boldsymbol{Z})=1|S=\texttt{A},Y=0]=\mathbb{P}[m_t(\boldsymbol{Z})=1|S=\texttt{B},Y=0]=\mathbb{P}[m_t(\boldsymbol{Z})=1|Y=0]. \end{cases}$$

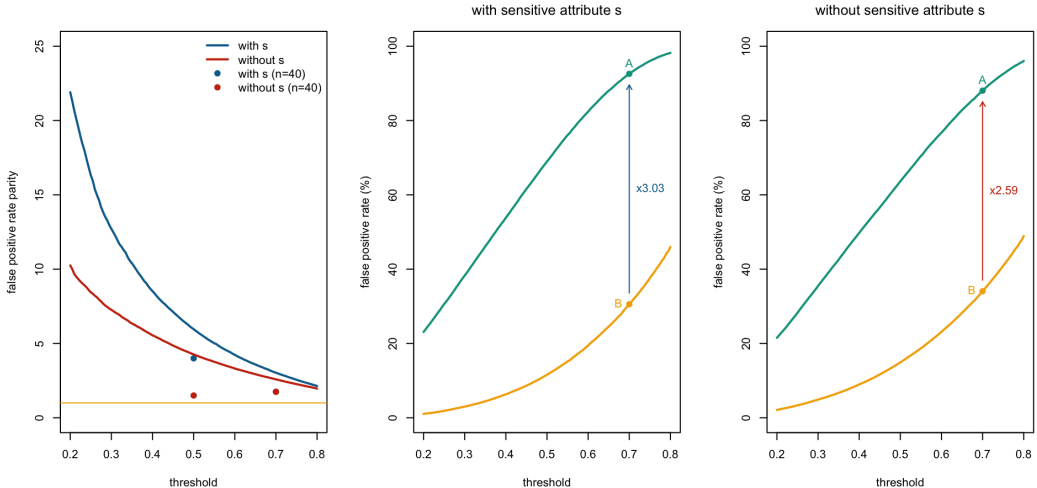
Separation and Equalized Odds





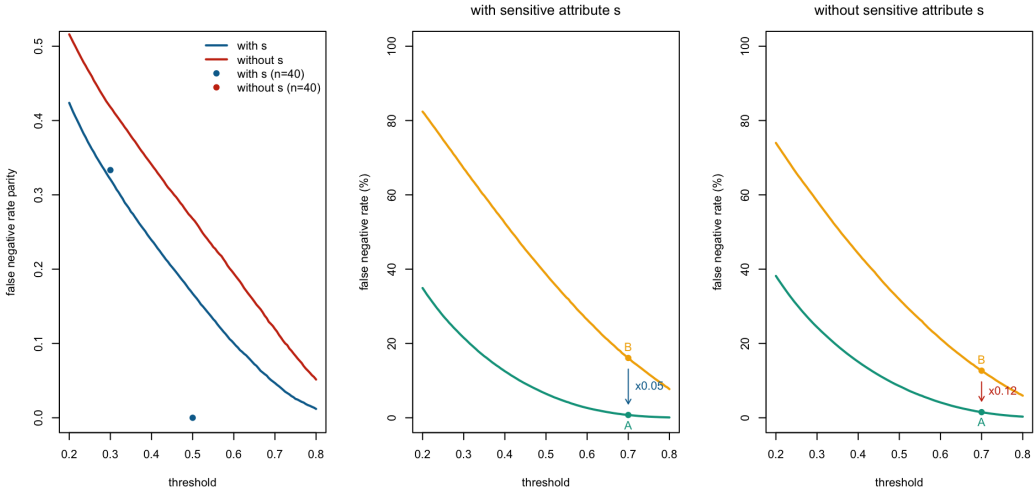
- ROC curves (TPR against FPR) for the logistic regression on `toydata2`.

Separation and Equalized Odds



- Evolution of the false positive rates, `fpr_parity` from `fairness`.

Separation and Equalized Odds



- Evolution of the false negative rates, `fnr_parity` from `fairness`.

Separation and Equalized Odds

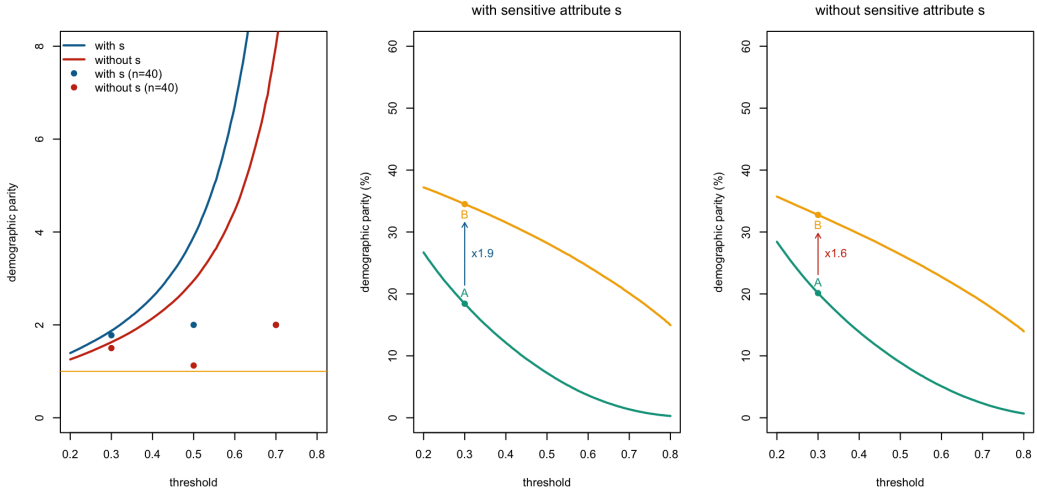
Definition 5.15: Equalized Odds, Hardt et al. (2016)

A decision function $\widehat{y}$ – or a classifier $m_t(\cdot)$ taking values in $\{0,1\}$ – satisfies equal odds constraint, with respect to some sensitive attribute S, if

$$\begin{cases} \mathbb{P}[\widehat{Y}=1|S=\texttt{A},Y=y]=\mathbb{P}[\widehat{Y}=1|S=\texttt{B},Y=y]=\mathbb{P}[\widehat{Y}=1|Y=y],\ \forall y\in\{0,1\} \\ \mathbb{P}[m_t(\boldsymbol{Z})=1|S=\texttt{A},Y=y]=\mathbb{P}[m_t(\boldsymbol{Z})=1|S=\texttt{B},Y=y],\ \forall y\in\{0,1\}, \end{cases},$$

which corresponds to parity of true positive and false positive, in the two groups.

Separation and Equalized Odds



- Evolution of the equalized odds metrics

Separation and Equalized Odds

› One can also consider any kind of standard metrics on confusion matrices, such as ϕ (introduced in Yule (1912)), usually named "Matthews correlation coefficient"

Definition 5.16: ϕ-fairness, Chicco and Jurman (2020)

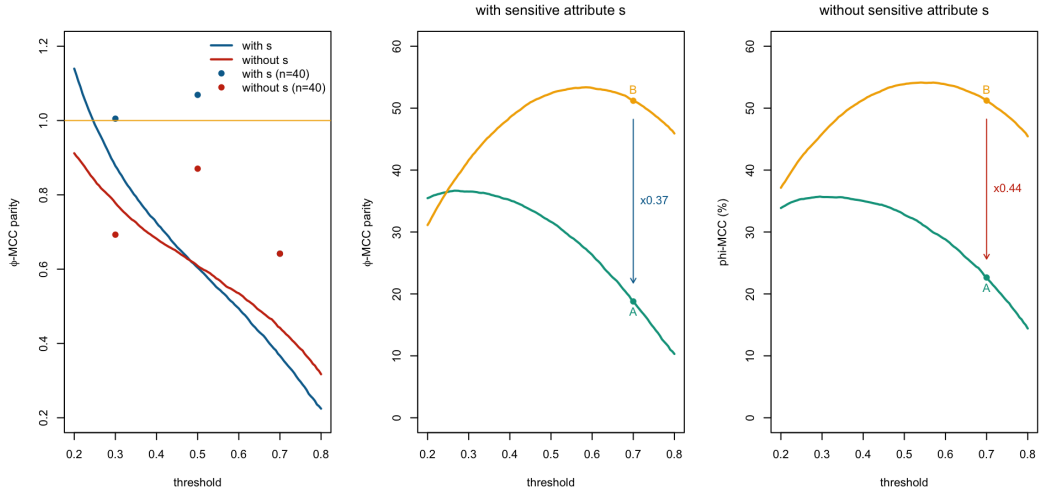
We will have ϕ-fairness if $\phi_{\mathtt{A}} = \phi_{\mathtt{B}}$, where ϕ_s denotes Matthews correlation coefficient for the s group,

$$\phi_s = \frac{\mathrm{TP}_s \cdot \mathrm{TN}_s - \mathrm{FP}_s \cdot \mathrm{FN}_s}{\sqrt{(\mathrm{TP}_s + \mathrm{FP}_s)(\mathrm{TP}_s + \mathrm{FN}_s) \cdot (\mathrm{TN}_s + \mathrm{FP}_s)(\mathrm{TN}_s + \mathrm{FN}_s)}}, \ s \in \{\mathtt{A}, \mathtt{B}\}.$$

› but one could consider the F_1-score (as defined in Van Rijsbergen (1979)), Fowlkes–Mallows or Jaccard indices (in Fowlkes and Mallows (1983) or Jaccard (1901)).

› .. or AUC as we will considered later on.

Separation and Equalized Odds



- Evolution of the ϕ-fairness metric

Separation and Equalized Odds

Definition 5.17: Class Balance, Kleinberg et al. (2016)

We will have class balance in the weak sense if

$$\mathbb{E}[m(\boldsymbol{X})|Y=y,S=\texttt{A}]=\mathbb{E}[m(\boldsymbol{X})|Y=y,S=\texttt{B}],\ \forall y\in\{0,1\},$$

or in the strong sense if

$$\mathbb{P}[m(\boldsymbol{X})\in\mathcal{A}|Y=y,S=\texttt{A}]=\mathbb{P}[m(\boldsymbol{X})\in\mathcal{A}|Y=y,S=\texttt{B}],\ \forall\mathcal{A}\subset[0,1],\ \forall y\in\{0,1\}.$$

Separation and Equalized Odds

Definition 5.18: Similar Mistreatement, Zafar et al. (2019)

We will have similar mistreatment, or "*lack of disparate mistreatment*," if

$$\begin{cases} \mathbb{P}[\widehat{Y} = Y | S = \mathtt{A}] = \mathbb{P}[\widehat{Y} = Y | S = \mathtt{B}] = \mathbb{P}[\widehat{Y} = Y] \\ \mathbb{P}[m_t(\boldsymbol{X}) = Y | S = \mathtt{A}] = \mathbb{P}[m_t(\boldsymbol{X}) = Y | S = \mathtt{B}] = \mathbb{P}[m_t(\boldsymbol{X}) = Y]. \end{cases}$$

Definition 5.19: Equality of ROC curves, Vogel et al. (2021)

Let $\text{FRP}_s(t) = \mathbb{P}[m(\boldsymbol{X}) > t | Y = 0, S = s]$ and $\text{TPR}_s(t) = \mathbb{P}[m(\boldsymbol{X}) > t | Y = 1, S = s]$, where $s \in \{\mathtt{A}, \mathtt{B}\}$. Set $\Delta_{TPR}(t) = \text{TPR}_{\mathtt{B}} \circ \text{TPR}_{\mathtt{A}}^{-1}(t) - t$ et $\Delta_{FRP}(t) = \text{FPR}_{\mathtt{B}} \circ \text{FPR}_{\mathtt{A}}^{-1}(t) - t$. We will have fairness with respect to ROC curves if $\|\Delta_{TPR}\|_\infty = \|\Delta_{FPR}\|_\infty = 0$.

Separation and Equalized Odds

Definition 5.20: AUC Fairness, Borkan et al. (2019)

We will have AUC fairness if $\text{AUC}_A = \text{AUC}_B$, where AUC_s is the AUC associated with model m within the s group.

	unaware (without s)				aware (with s)			
	GLM	GAM	CART	RF	GLM	GAM	CART	RF
ratio of AUC	0.837	0.839	0.913	0.768	0.857	0.860	0.913	0.763

Sufficiency and Calibration

- Inspired by Cleary (1968), define

Definition 5.21: Sufficiency, Barocas et al. (2017)

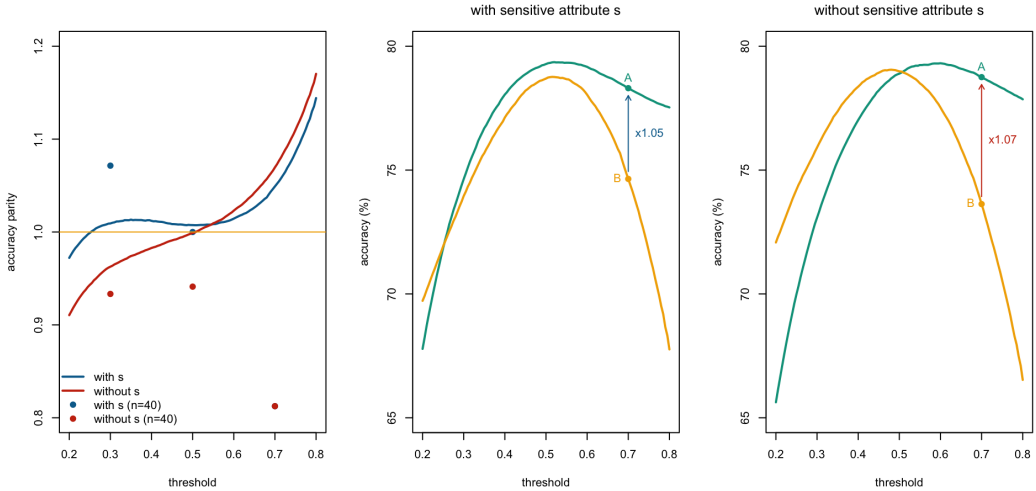
A model $m : \mathcal{Z} \to \mathcal{Y}$ satisfies the sufficiency property if $Y \perp\!\!\!\perp S \mid m(\boldsymbol{Z})$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$.

Definition 5.22: Calibration Parity, Accuracy Parity, Kleinberg et al. (2016), Zafar et al. (2019)

Calibration parity is met if

$$\mathbb{P}[Y = 1 | m(\boldsymbol{X}) = t, S = \texttt{A}] = \mathbb{P}[Y = 1 | m(\boldsymbol{X}) = t, S = \texttt{B}], \ \forall t \in [0, 1].$$

Sufficiency and Calibration



- Evolution of accuracy, in groups A and B.

Sufficiency and Calibration

Definition 5.23: Good Calibration, Kleinberg et al. (2017), Verma and Rubin (2018)

Fairness of good calibration is met if

$$\mathbb{P}[Y = 1 | m(\boldsymbol{X}) = t, S = \mathtt{A}] = \mathbb{P}[Y = 1 | m(\boldsymbol{X}) = t, S = \mathtt{B}] = t, \ \forall t \in [0, 1].$$

Definition 5.24: Non-Reconstruction of Protected Attribute, Kim (2017)

If we cannot tell from the result ($\boldsymbol{x}$, $m(\boldsymbol{x})$, y and $\widehat{y}$) whether the subject was a member of a protected group or not, we will talk about fairness by non-reconstruction of the protected attribute

$$\mathbb{P}[S = \mathtt{A} | \boldsymbol{X}, m(\boldsymbol{X}), \widehat{Y}, Y] = \mathbb{P}[S = \mathtt{B} | \boldsymbol{X}, m(\boldsymbol{X}), \widehat{Y}, Y].$$

Relaxation and Approximate Fairness

Definition 5.25: Disparate Impact, Feldman et al. (2015)

A decision function $\widehat{Y}$ has a disparate impact, for a given threshold τ, if,

$$\min\left\{\frac{\mathbb{P}[\widehat{Y}=1|S=\mathtt{A}]}{\mathbb{P}[\widehat{Y}=1|S=\mathtt{B}]}, \frac{\mathbb{P}[\widehat{Y}=1|S=\mathtt{B}]}{\mathbb{P}[\widehat{Y}=1|S=\mathtt{A}]}\right\} < \tau \text{ (usually 80\%)}.$$

› The 80% rule was suggested by the "Technical Advisory Committee on Testing", from the State of California Fair Employment Practice Commission (FEPC) in 1971, or the 1978 "Uniform Guidelines on Employee Selection Procedures", a document used by the U.S. Equal Employment Opportunity Commission (EEOC), see Biddle (2017).

Relaxation and Approximate Fairness

- We have defined (Definition 5.10) unfairness as

$$\mathcal{U}_k(m) = \max_{s\in\{\texttt{A},\texttt{B}\}} \{W_k(\mathbb{P}_m, \mathbb{P}_{m|s})\},$$

so that m is (strongly) fair if and only if $\mathcal{U}_k(m) = 0$.

- Chzhen and Schreuder (2022) introduced the notion of Relative Improvement

Definition 5.26: ε-Approximate Fairness

Model m is ε-approximately fair if $\mathcal{U}_k(m) \leq \varepsilon \cdot \mathcal{U}_k(m^\star)$, where $m^\star$ is Bayes regressor, for some $\epsilon \geq 0$.

Three different concepts ?

$$\begin{cases} \text{Independence (Definition 5.5)}: & m(\boldsymbol{Z}) \perp\!\!\!\perp S \\ \text{Separation (Definition 5.11)}: & m(\boldsymbol{Z}) \perp\!\!\!\perp S \mid Y \\ \text{Sufficiency (Definition 5.21)}: & Y \perp\!\!\!\perp S \mid m(\boldsymbol{Z}) \end{cases}$$

- Independence assumes no differences among groups, regardless of accuracy
- Separation minimizes differences among groups by not trying to maximize accuracy
- Sufficiency maximizes accuracy by not trying to minimize differences among groups

See Kleinberg et al. (2016) or Chouldechova (2017).

Impossibility theorems

› Unless very specific properties are assumed on $\mathbb{P}$, there is no prediction function $m(\cdot)$ that can satisfy at the same time two fairness criteria.

$$\begin{cases} \text{Independence (Definition 5.5)}: & m(\boldsymbol{Z}) \perp\!\!\!\perp S \\ \text{Separation (Definition 5.11)}: & m(\boldsymbol{Z}) \perp\!\!\!\perp S \mid Y \\ \text{Sufficiency (Definition 5.21)}: & Y \perp\!\!\!\perp S \mid m(\boldsymbol{Z}) \end{cases}$$

Proposition 5.4

Suppose that a model m satisfies the independence condition (5.5) and the sufficiency property (5.21), with respect to a sensitive attribute s, then necessarily, $Y \perp\!\!\!\perp S$.

› Therefore, unless the sensitive attribute s has no impact on the outcome y, there is no model m which satisfies independence and sufficiency simultaneously.

Impossibility theorems

- From the sufficiency property , $S \perp\!\!\!\perp Y \mid m(\boldsymbol{Z})$, then, for $s \in \mathcal{S}$ and $\mathcal{A} \subset \mathcal{Y}$,

$$\mathbb{P}[S = s, Y \in \mathcal{A}] = \mathbb{E}\big[\mathbb{P}[S = s, Y \in \mathcal{A} | m(\boldsymbol{Z})]\big],$$

can be written

$$\mathbb{P}[S = s, Y \in \mathcal{A}] = \mathbb{E}\big[\mathbb{P}[S = s | m(\boldsymbol{Z})] \cdot \mathbb{P}[Y \in \mathcal{A} | m(\boldsymbol{Z})]\big].$$

And from the independence property (5.21), $m(\boldsymbol{Z}) \perp\!\!\!\perp S$, we can write the first component $\mathbb{P}[S = s | m(\boldsymbol{Z})] = \mathbb{P}[S = s]$, almost surely, and therefore

$$\mathbb{P}[S = s, Y \in \mathcal{A}] = \mathbb{E}\big[\mathbb{P}[S = s] \cdot \mathbb{P}[Y \in \mathcal{A} | m(\boldsymbol{Z})]\big] = \mathbb{P}[S = s] \cdot \mathbb{P}[Y \in \mathcal{A}],$$

for all $s \in \mathcal{S}$ and $\mathcal{A} \subset \mathcal{Y}$, corresponding to the independence between S and Y.

Impossibility theorems

Proposition 5.5

Consider a classifier m_t taking values in $\mathcal{Y} = \{0, 1\}$. Suppose that m_t satisfies the independence condition (5.5) and the separation property (5.11), with respect to a sensitive attribute s, then necessarily either $m_t(\boldsymbol{Z}) \perp\!\!\!\perp Y$ or $Y \perp\!\!\!\perp S$ (possibly both).

› Because m_t satisfies the independence condition (5.5), $m_t(\boldsymbol{Z}) \perp\!\!\!\perp S$, and the separation property (5.11), $m_t(\boldsymbol{Z}) \perp\!\!\!\perp S \mid Y$, them, for $\widehat{y} \in \mathcal{Y}$ and for $s \in \mathcal{S}$,

$$\mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y}] = \mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y} | S = s] = \mathbb{E}[\mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y} | Y, S = s]],$$

that we can write

$$\mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y}] = \sum_y \mathbb{P}\big[m_t(\boldsymbol{Z}) = \widehat{y} | Y = y, S = s\big] \cdot \mathbb{P}\big[Y = y | S = s\big],$$

Impossibility theorems

or

$$\mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y}] = \sum_y \mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y}|Y = y] \cdot \mathbb{P}[Y = y|S = s],$$

almost surely. Furthermore, we can also write

$$\mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y}] = \sum_y \mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y}|Y = y] \cdot \mathbb{P}[Y = y],$$

so that, if we combine the two expressions, we get

$$\sum_y \mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y}|Y = y] \cdot \Big(\mathbb{P}[Y = y|S = s] - \mathbb{P}[Y = y]\Big) = 0,$$

almost surely. And since we assumed that y was a binary variable, $\mathbb{P}[Y = 0] = 1 - \mathbb{P}[Y = 1]$, as well as $\mathbb{P}[Y = 0|S = s] = 1 - \mathbb{P}[Y = 1|S = s]$, and therefore

$$\mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y}|Y = 1] \cdot \Big(\mathbb{P}[Y = 1|S = s] - \mathbb{P}[Y = 1]\Big)$$

Impossibility theorems

or

$$-\mathbb{P}\big[m_t(\boldsymbol{Z}) = \widehat{y}|Y=0\big] \cdot \Big(\mathbb{P}\big[Y=0|S=s\big] - \mathbb{P}\big[Y=0\big]\Big)$$

can be written

$$\mathbb{P}\big[m_t(\boldsymbol{Z}) = \widehat{y}|Y=0\big] \cdot \Big(\mathbb{P}\big[Y=1|S=s\big] - \mathbb{P}\big[Y=1\big]\Big).$$

Thus, either $\mathbb{P}\big[Y=1|S=s\big] - \mathbb{P}\big[Y=1\big]$ almost surely, or $\mathbb{P}\big[m_t(\boldsymbol{Z}) = \widehat{y}|Y=0\big] = \mathbb{P}\big[m_t(\boldsymbol{Z}) = \widehat{y}|Y=1\big]$ (or both).

- Of course, the previous proposition holds only when y is a binary variable.

Impossibility theorems

Proposition 5.6

Consider a classifier m_t taking values in $\mathcal{Y} = \{0, 1\}$. Suppose that m_t satisfies the sufficiency condition (5.21) and the separation property (5.11), with respect to a sensitive attribute s, then necessarily either $\mathbb{P}[m_t(\boldsymbol{Z}) = 1|Y = 1] = 0$ or $Y \perp\!\!\!\perp S$ or $m_t(\boldsymbol{Z}) \perp\!\!\!\perp Y$.

› Suppose that m_t satisfies the sufficiency condition (5.21) and the separation property (5.11), respectively $Y \perp\!\!\!\perp S \mid m_t(\boldsymbol{Z})$ and $m_t(\boldsymbol{Z}) \perp\!\!\!\perp S \mid Y$. For all $s \in \mathcal{S}$, we can write, using Bayes formula

$$\mathbb{P}[Y = 1|S = s, m_t(\boldsymbol{Z}) = 1] = \frac{\mathbb{P}[m_t(\boldsymbol{Z}) = 1|Y = 1, S = s] \cdot \mathbb{P}[Y = 1|S = s]}{\mathbb{P}[m_t(\boldsymbol{Z}) = 1|S = s]},$$

Impossibility theorems

i.e.,

$$\mathbb{P}[Y=1|S=s,m_t(\boldsymbol{Z})=1]=\frac{\mathbb{P}[m_t(\boldsymbol{Z})=1|Y=1]\cdot\mathbb{P}[Y=1|S=s]}{\sum_{y\in\{0,1\}}\mathbb{P}[m_t(\boldsymbol{Z})=1|Y=y]\cdot\mathbb{P}[Y=1|S=s]},$$

that should not depend on s (from the sufficiency property). So a similar property holds if $S=s'$. Observe further that $\mathbb{P}[m_t(\boldsymbol{Z})=1|Y=1]$ is the *true positive rate* (TPR) while $\mathbb{P}[m_t(\boldsymbol{Z})=1|Y=0]$ is the *false positive rate* (TPR). Let $p_s=\mathbb{P}[Y=1|S=s]$, so that

$$\mathbb{P}[Y=1|S=s,m_t(\boldsymbol{Z})=1]=\frac{\text{TPR}}{p_s\cdot\text{TPR}+(1-p_s)\cdot\text{FPR}}.$$

Impossibility theorems

› Suppose that Y and S are not independent (otherwise $Y \perp\!\!\!\perp S$ as stated in the proposition), i.e., there are s and s' such that $p_s = \mathbb{P}[Y = 1|S = s] \neq \mathbb{P}[Y = 1|S = s'] = p_{s'}$. Hence, $p_s \neq p_{s'}$, but at the same time

$$\frac{\text{TPR}}{p_s \cdot \text{TPR} + (1 - p_s) \cdot \text{FPR}} = \frac{\text{TPR}}{p_{s'} \cdot \text{TPR} + (1 - p_{s'}) \cdot \text{FPR}}.$$

Supposes that $\text{TPR} \neq 0$ (otherwise $\text{TPR} = \mathbb{P}[m_t(\boldsymbol{Z}) = 1|Y = 1] = 0$ as stated in the proposition), then

$$(p_s - p_{s'}) \cdot \text{TPR} = (p_s - p_{s'}) \cdot \text{FPR} \neq 0,$$

and therefore $m_t(\boldsymbol{Z}) \perp\!\!\!\perp Y$.

– Part 5 –

Individual Fairness

Individual Fairness

Definition 6.1: Similarity Fairness, Luong et al. (2011), Dwork et al. (2012)

Consider two metrics, one on $\mathcal{Y} \times \mathcal{Y}$ (or for a classifier $[0,1]$ and not $\{0,1\}$) noted D_y, and one on $\mathcal{X}$ noted D_x, such that we will have similarity fairness on a database of size n if we have the following property (called Lipschitz property)

$$D_y(m(\boldsymbol{x}_i, s_i), m(\boldsymbol{x}_j, s_j)) \leq L \cdot D_x(\boldsymbol{x}_i, \boldsymbol{x}_j), \ \forall i, j = 1, \cdots, n,$$

for some $L < \infty$.

Individual Fairness

Definition 6.2: Local individual fairness, Petersen et al. (2021)

Consider two metrics, one on $\mathcal{Y}$ ($[0,1]$ for a classifier and not $\{0,1\}$) noted D_y, and one on $\mathcal{X}$ noted D_x, model m is locally individually fair if

$$\mathbb{E}_{(\boldsymbol{X},S)}\left[\limsup_{\boldsymbol{x}':D_x(\boldsymbol{X},\boldsymbol{x}')\to 0}\frac{D_y(m(\boldsymbol{X},S),m(\boldsymbol{x}',S))}{D_x(\boldsymbol{X},\boldsymbol{x}')}\right]\leq L<\infty.$$

Individual Fairness

Definition 6.3: Proxy Based Fairness, Kilbertus et al. (2017)

A decision making process $\widehat{y}$ exhibits no proxy discrimination with respect to sensitive attribute s if

$$\mathbb{E}\big[\widehat{Y}\big|\text{do}(S=\texttt{A})\big]=\mathbb{E}\big[\widehat{Y}\big|\text{do}(S=\texttt{B})\big].$$

Definition 6.4: Fairness on Average Treatment Effect, Kusner et al. (2017)

We achieve fairness on average treatment effect (counterfactual fairness on average)

$$\text{ATE}=\mathbb{E}\big[Y^\star_{S\leftarrow\texttt{A}}-Y^\star_{S\leftarrow\texttt{B}}\big]=0.$$

Individual Fairness

- A decision satisfies counterfactual 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*."

Definition 6.5: Counterfactual Fairness, Kusner et al. (2017)

We achieve counterfactual fairness for an individual with characteristics $\boldsymbol{x}$ if

$$\text{CATE}(\boldsymbol{x}) = \mathbb{E}\big[Y^{\star}_{S\leftarrow \text{A}} - Y^{\star}_{S\leftarrow \text{B}} \big| \boldsymbol{X} = \boldsymbol{x}\big] = 0.$$

References

Abraham, K. (1986). *Distributing risk: Insurance, legal theory and public policy*. Yale University Press,.

Aczél, J. (1948). On mean values. *Bulletin of the American Mathematical Society*, 54(4):392–400.

Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126.

Agueh, M. and Carlier, G. (2011). Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924.

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142.

Amari, S.-i. (2016). *Information geometry and its applications*, volume 194. Springer.

Andrus, M., Spitzer, E., Brown, J., and Xiang, A. (2021). What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 249–260.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica*, May 23.

References

Apfelbaum, E. P., Pauker, K., Sommers, S. R., and Ambady, N. (2010). In blind pursuit of racial equality? *Psychological science*, 21(11):1587–1592.

Austin, R. (1983). The insurance classification controversy. *University of Pennsylvania Law Review*, 131(3):517–583.

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

Avraham, R., Logue, K. D., and Schwarcz, D. (2013). Understanding insurance antidiscrimination law. *Southern California Law Review*, 87:195.

Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48.

Bailey, R. A. and Simon, L. J. (1959). An actuarial note on the credibility of experience of a single private passenger car. *Proceedings of the Casualty Actuarial Society*, XLVI:159.

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

Banerjee, A., Merugu, S., Dhillon, I. S., Ghosh, J., and Lafferty, J. (2005). Clustering with bregman divergences. *Journal of machine learning research*, 6(10).

Barbour, V. (1911). Privateers and pirates of the west indies. *The American Historical Review*, 16(3):529–566.

References

Barocas, S., Hardt, M., and Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, 1:2017.

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

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

Bauschke, H. H., Borwein, J. M., et al. (1997). Legendre functions and the method of random bregman projections. *Journal of convex analysis*, 4(1):27–67.

Bauschke, H. H. and Lewis, A. S. (2000). Dykstras algorithm with bregman projections: A convergence proof. *Optimization*, 48(4):409–427.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202.

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

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, R. D. (2018). Mine: Mutual information neural estimation.

Bellemare, M. G., Dabney, W., and Munos, R. (2017a). A distributional perspective on reinforcement learning. *arXiv:1707.06887*.

References

Bellemare, M. G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., and Munos, R. (2017b). The cramer distance as a solution to biased wasserstein gradients. *arXiv:1705.10743*.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. (2015). Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138.

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.

Biddle, D. (2017). *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Routledge.

Billingsley, P. (2017). *Probability and measure*. John Wiley & Sons.

Blanpain, N. (2018). L'espérance de vie par niveau de vie-méthode et principaux résultats. *INSEE Document de Travail*, F1801.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *The Journal of Human Resources*, 8(4):436–455.

Bollen, K. A. (1989). *Structural equations with latent variables*, volume 210. John Wiley & Sons.

References

Bollen, K. A. and Pearl, J. (2013). Eight myths about causality and structural equation models. In *Handbook of causal analysis for social research*, pages 301–328. Springer.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Bourdieu, P. (2018). Distinction a social critique of the judgement of taste. In *Inequality Classic Readings in Race, Class, and Gender*, pages 287–318. Routledge.

Box, G. E., Luceño, A., and del Carmen Paniagua-Quinones, M. (2011). *Statistical control by monitoring and adjustment*, volume 700. John Wiley & Sons.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Boyle, J. P. and Dykstra, R. L. (1986). A method for finding projections onto the intersection of convex sets in hilbert spaces. In *Advances in Order Restricted Statistical Inference: Proceedings of the Symposium on Order Restricted Statistical Inference held in Iowa City, Iowa, September 11–13, 1985*, pages 28–47. Springer.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217.

References

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

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.

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

Brown, R. S., Moon, M., and Zoloth, B. S. (1980). Incorporating occupational attainment in studies of male-female earnings differentials. *Journal of Human Resources*, pages 3–28.

Brualdi, R. A. (2006). *Combinatorial matrix classes*, volume 13. Cambridge University Press.

Bures, D. (1969). An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w^*-algebras. *Transactions of the American Mathematical Society*, 135:199–212.

Calders, T. and Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21(2):277–292.

References

Carlier, G., Duval, V., Peyré, G., and Schmitzer, B. (2017). Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418.

Carlini, N. and Wagner, D. (2017). Towards evaluating the robustness of neural networks.

Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Duxbury Advanced Series.

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.

Censor, Y. and Reich, S. (1998). The dykstra algorithm with bregman projections. *Communications in Applied Analysis*, 2(3):407–420.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions, intl. *Journal of Mathematical Models and Methods in Applied Sciences, Issue*, 4.

Chambert-Loir, A. (2023). *Information Theory: Three Theorems by Claude Shannon*, volume 144. Springer Nature.

Charniak, E. (1991). Bayesian networks without tears. *AI magazine*, 12(4):50–50.

Charpentier, A. (2014). Mesures de risque. In Droesbeke, J.-J. and Saporta, G., editors, *Approches statistiques du risque*. Éditions Technip.

Charpentier, A. (2023). *Insurance: biases, discrimination and fairness*. Springer Verlag.

References

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

Cheney-Lippold, J. (2017). We are data. In *We Are Data*. New York University Press.

Chicco, D. and Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Chzhen, E. and Schreuder, N. (2022). A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50(4):2416–2442.

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, 5(2):115–124.

Cohen, I. and Goldszmidt, M. (2004). Properties and benefits of calibrated classifiers. In *8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, volume 3202, pages 125–136. Springer.

Conway, D. A. and Roberts, H. V. (1983). Reverse regression, fairness, and employment discrimination. *Journal of Business & Economic Statistics*, 1(1):75–85.

References

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. *arXiv*, 1701.08230.

Cramér, H. (1928a). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74.

Cramér, H. (1928b). On the composition of elementary errors: second paper: statistical applications. *Scandinavian Actuarial Journal*, 1928(1):141–180.

Crossney, K. B. (2016). Redlining. *https://philadelphiaencyclopedia.org/essays/redlining/*.

Csiszár, I. (1964). Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 8:85–108.

Csiszár, I. (1967). On information-type measure of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318.

Cunningham, S. (2021). *Causal inference*. Yale University Press.

Cuturi, M. and Doucet, A. (2014). Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR.

Da Silva, N. (2023). *La bataille de la Sécu: une histoire du système de santé*. La fabrique éditions.

References

Dall'Aglio, G. (1956). Sugli estremi dei momenti delle funzioni di ripartizione doppia. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 10(1-2):35–74.

Dantzig, G. B. and Thapa, M. N. (1997). *Linear programming: Introduction*, volume 1. Springer.

Darlington, R. B. (1971). Another look at "cultural fairness" 1. *Journal of educational measurement*, 8(2):71–82.

Datta, A., Fredrikson, M., Ko, G., Mardziel, P., and Sen, S. (2017). Proxy non-discrimination in data-driven systems. *arXiv*, 1707.08120.

Dawid, A. P. (1982). The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

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.

de La Fontaine, J. (1668). *Fables*. Barbin.

De Pril, N. and Dhaene, J. (1996). Segmentering in verzekeringen. *DTEW Research Report 9648*, pages 1–56.

De Wit, G. and Van Eeghen, J. (1984). Rate making and society's sense of fairness. *ASTIN Bulletin: The Journal of the IAA*, 14(2):151–163.

References

Dedecker, J. and Merlevède, F. (2007). The empirical distribution function for dependent variables: asymptotic and nonasymptotic results in. *ESAIM: Probability and Statistics*, 11:102–114.

Delobelle, P., Temple, P., Perrouin, G., Frénay, B., Heymans, P., and Berendt, B. (2021). Ethical adversaries: Towards mitigating unfairness with adversarial machine learning. *ACM SIGKDD Explorations Newsletter*, 23(1):32–41.

Denis, C., Elie, R., Hebiri, M., and Hu, F. (2021). Fairness guarantee in multi-class classification. *arXiv*, 2109.13642.

Denuit, M. and Charpentier, A. (2004). *Mathématiques de l'assurance non-vie: Tome I Principes fondamentaux de théorie du risque*. Economica.

Denuit, M., Charpentier, A., and Trufin, J. (2021). Autocalibration and tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics & Economics*.

Devroye, L., Mehrabian, A., and Reddad, T. (2018). The total variation distance between high-dimensional gaussians with the same mean. *arXiv*, 1810.08693.

Dhaene, J., Denuit, M., Goovaerts, M. J., Kaas, R., and Vyncke, D. (2002a). The concept of comonotonicity in actuarial science and finance: applications. *Insurance: Mathematics and Economics*, 31(2):133–161.

References

Dhaene, J., Denuit, M., Goovaerts, M. J., Kaas, R., and Vyncke, D. (2002b). The concept of comonotonicity in actuarial science and finance: theory. *Insurance: Mathematics and Economics*, 31(1):3–33.

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

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580.

Duncan, O. D. (1975). *Introduction to structural equation models*. Academic Press.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

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.

Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860.

Feeley, M. and Simon, J. (1994). Actuarial justice: The emerging new criminal law. *The futures of criminology*, 173:174.

References

Feeley, M. M. and Simon, J. (1992). The new penology: Notes on the emerging strategy of corrections and its implications. *Criminology*, 30(4):449–474.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268.

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.

Fourcade, M. and Healy, K. (2013). Classification situations: Life-chances in the neoliberal era. *Accounting, Organizations and Society*, 38(8):559–572.

Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553–569.

Fox, E. T. (2013). *'Piratical Schemes and Contracts': Pirate Articles and Their Society 1660-1730*. PhD Thesis, University of Exeter.

François, P. (2022). Catégorisation, individualisation. retour sur les scores de crédit. *hal*, 03508245.

Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré*, volume 10, pages 215–310.

References

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Annales de l'Université de Lyon, 3^e serie, Sciences, Sect.A*, 14 : 53 – –77.

Freeman, S. (2007). *Rawls*. Routledge.

Frisch, K. R. (1955). The logarithmic potential method of convex programming. *Memorandum, University Institute of Economics, Oslo*, 5(6).

Galichon, A. (2016). *Optimal transport methods in economics*. Princeton University Press.

Gandy, O. H. (2016). *Coming to terms with chance: Engaging rational discrimination and cumulative disadvantage*. Routledge.

Gangbo, W. (1999). The monge mass transfer problem and its applications. *Contemporary Mathematics*, 226:79–104.

Garrioch, D. (2011). Mutual aid societies in eighteenth-century paris. *French History & Civilization*, 4.

Gebelein, H. (1941). Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379.

Ginsburg, M. (1940). Roman military clubs and their social functions. In *Transactions and Proceedings of the American Philological Association*, volume 71, pages 149–156. JSTOR.

References

Givens, C. R. and Shortt, R. M. (1984). A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240.

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.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Gneiting, T. and Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, 310(5746):248–249.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica*, pages 979–1001.

Goldman, A. (1979). *Justice and Reverse Discrimination*. Princeton University Press.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1):107–114.

References

Goodfellow, I., McDaniel, P., and Papernot, N. (2018). Making machine learning robust against adversarial inputs. *Communications of the ACM*, 61(7):56—66.

Goodfellow, I., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *ArXiv*, 1412.6572.

Gouic, T. L., Loubes, J.-M., and Rigollet, P. (2020). Projection to fairness in statistical learning. *arXiv*, 2005.11720.

Gowri, A. (2014). *The Irony of Insurance: Community and Commodity*. PhD thesis, University of Southern California.

Gretton, A., Smola, A., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. (2005). Kernel constrained covariance for dependence measurement. In *International Workshop on Artificial Intelligence and Statistics*, pages 112–119. PMLR.

Grove, K. and Karcher, H. (1973). How to conjugate c 1-close group actions. *Mathematische Zeitschrift*, 132(1):11–20.

Hacking, I. (1990). *The taming of chance*. Number 17. Cambridge University Press.

Han, X., Baldwin, T., and Cohn, T. (2022). Towards equal opportunity fairness through adversarial learning. *arXiv*, 2203.06317.

References

Hannan, E. J. (1961). The general theory of canonical correlation and its relation to functional analysis. *Journal of the Australian Mathematical Society*, 2(2):229–242.

Harari, Y. N. (2018). *21 Lessons for the 21st Century*. Random House.

Harcourt, B. E. (2015). Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter*, 27(4):237–243.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Hardy, G. H., Littlewood, J. E., Pólya, G., Pólya, G., et al. (1952). *Inequalities*. Cambridge university press.

Havens, H. V. (1979). Issues and needed improvements in state regulation of the insurance business. *U.S. General Acounting Office*.

He, X. D., Kou, S., and Peng, X. (2022). Risk measures: robustness, elicitability, and backtesting. *Annual Review of Statistics and Its Application*, 9:141–166.

Heimer, C. A. (1985). *Reactive Risk and Rational Action*. University of California Press.

Hellinger, E. (1909). Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 1909(136):210–271.

References

Henrion, M. (1988). Propagating uncertainty in bayesian networks by probabilistic logic sampling. In *Machine intelligence and pattern recognition*, volume 5, pages 149–163. Elsevier.

Hey, R. (1814). Xviii. propositions containing some properties of tangents to circles; and of trapeziums inscribed in circles, and non-inscribed. together with propositions on the elliptic representations of circles, upon a plane surface, by perspective. *Philosophical Transactions of the Royal Society of London*, (104):348–396.

Higham, N. J. (2008). *Functions of matrices: theory and computation*. SIAM.

Hill, K. and White, J. (2020). Designed to deceive: do these people look real to you? *The New York Times*, 11(21).

Hirschfeld, H. O. (1935). A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4):520–524.

Hoeffding, W. (1940). Masstabinvariante korrelationstheorie. *Schriften des Mathematischen Instituts und Instituts fur Angewandte Mathematik der Universitat Berlin*, 5:181–233.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hoffman, F. L. (1896). *Race traits and tendencies of the American Negro*, volume 11. American Economic Association.

References

Hoffman, F. L. (1918). *Mortality from respiratory diseases in dusty trades (inorganic dusts)*. Number 231. US Government Printing Office.

Hoffman, F. L. (1931). Cancer and smoking habits. *Annals of surgery*, 93(1):50.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Hubbard, G. N. (1852). *De l'organisation des sociétés de bienfaisance ou de secours mutuels et des bases scientifiques sur lesquelles elles doivent être établies*. Paris, Guillaumin.

Huttegger, S. M. (2013). In defense of reflection. *Philosophy of Science*, 80(3):413–433.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Ismay, P. (2018). *Trust among strangers: friendly societies in modern Britain*. Cambridge University Press.

References

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise de Sciences Naturelles*, 37:547–579.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

Jensen, D. and Mayer, L. (1977). Some variational results and their applications in multiple inference. *The Annals of Statistics*, pages 922–931.

Jordan, C. (1881). Sur la serie de fourier. *Camptes Rendus Hebdomadaires de l'Academie des Sciences*, 92:228–230.

Kantorovich, L. and Rubinstein, G. (1958). On the space of completely additive functions. *Vestnic Leningrad Univ., Ser. Mat. Mekh. i Astron.*, 13(7):52–59. In Russian.

Kantorovich, L. V. (1942). On the translocation of masses. In *Doklady Akademii Nauk USSR*, volume 37, pages 199–201.

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

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta psychologica*, 77(3):217–273.

References

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *arXiv*, 1706.02744.

Kim, P. T. (2017). Auditing algorithms for discrimination. *University of Pennsylvania Law Review*, 166:189.

Kimeldorf, G., May, J. H., and Sampson, A. R. (1982). Concordant and discordant monotone correlations and their evaluation by nonlinear optimization. *Studies in the Management Sciences*, 19:117–130.

Kimeldorf, G. and Sampson, A. R. (1978). Monotone dependence. *The Annals of Statistics*, pages 895–903.

Kimeldorf, G. and Wahba, G. (1971). Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95.

Kitagawa, E. M. (1955). Components of a difference between two rates. *Journal of the american statistical association*, 50(272):1168–1194.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv*, 1609.05807.

References

Knott, M. and Smith, C. S. (1984). On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43:39–49.

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

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Koenker, R., Chernozhukov, V., He, X., and Peng, L. (2017). *Handbook of quantile regression*. CRC press.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91.

Kolmogorov, A. N. (1930). *Sur la notion de la moyenne*. G. Bardi, tip. della R. Accad. dei Lincei.

Komiyama, J. and Shimao, H. (2017). Two-stage algorithm for fairness-aware machine learning. *arXiv*, 1710.04924.

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

Krüger, F. and Ziegel, J. F. (2021). Generic conditions for forecast dominance. *Journal of Business & Economic Statistics*, 39(4):972–983.

References

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Kuhn, H. W. (1956). Variants of the hungarian method for assignment problems. *Naval research logistics quarterly*, 3(4):253–258.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

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

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. (2020). Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740.

Lancaster, H. O. (1957). Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika*, 44(1/2):289–292.

Lancaster, H. O. (1958). The Structure of Bivariate Distributions. *The Annals of Mathematical Statistics*, 29(3):719 – 736.

References

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the compas recidivism algorithm. *ProPublica*, 23-05.

Laskov, P. and Lippmann, R. (2010). Machine learning in adversarial environments.

Leeson, P. T. (2009). The calculus of piratical consent: the myth of the myth of social contract. *Public Choice*, 139:443–459.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Li, K. C.-W. (1996). The private insurance industry's tactics against suspected homosexuals: redlining based on occupation, residence and marital status. *American Journal of Law & Medicine*, 22(4):477–502.

Li, X., Cui, Z., Wu, Y., Gu, L., and Harada, T. (2021). Estimating and improving fairness with adversarial learning. *arXiv*, 2103.04243.

Lichtenstein, S., Fischhoff, B., and Phillips, L. D. (1977). Calibration of probabilities: The state of the art. *Decision making and change in human affairs*, pages 275–324.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

References

Lin, P.-E. (1987). Measures of association between vectors. *Communications in Statistics-Theory and Methods*, 16(2):321–338.

Linn, R. L. and Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement*, 8(1):1–4.

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

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. (2018). Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR.

Luong, B. T., Ruggieri, S., and Turini, F. (2011). k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 502–510.

Ma, L. and Koenker, R. (2006). Quantile regression methods for recursive structural equation models. *Journal of Econometrics*, 134(2):471–506.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018). Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR.

Massey, D. S. (2007). *Categorically unequal: The American stratification system*. Russell Sage Foundation.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. Chapman & Hall.

References

Mcdonald, S. (2015). Indirect gender discrimination and the 'test-achats ruling': an examination of the uk motor insurance market. In *Royal Economic Society Conf., Manchester*.

Meinshausen, N. and Ridgeway, G. (2006). Quantile regression forests. *Journal of machine learning research*, 7(6).

Merriam-Webster (2022). *Dictionary*. .

Möbius, A. F. (1827). *Der barycentrische Calcul, ein Hülfsmittel zur analytischen Behandlung der Geometrie (etc.)*. Leipzig: J.A. Barth.

Mourier, E. (1953). Eléments aléatoires dans un espace de banach. In *Annales de l'institut Henri Poincaré*, volume 13, pages 161–244.

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*.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600.

Murphy, A. H. and Epstein, E. S. (1967). Verification of probabilistic predictions: A brief review. *Journal of Applied Meteorology and Climatology*, 6(5):748–755.

References

Nagumo, M. (1930). Über eine klasse der mittelwerte. In *Japanese journal of mathematics*, volume 7, pages 71–79. The Mathematical Society of Japan.

Nathan, A. (1952). *College Geometry: An Introduction to the Modern Geometry of the Triangle and the Circle*. Barnes & Noble.

Neumann, J. v. and Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton University Press.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, pages 819–847.

Nielsen, F. (2022). The many faces of information geometry. *Notices of the American Mathematical Society*, 69(1):36–45.

Nielsen, F. and Nock, R. (2013). On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Processing Letters*, 21(1):10–13.

Oakes, D. (1985). Self-calibrating priors do not exist. *Journal of the American Statistical Association*, 80(390):339–339.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, 14(3):693–709.

References

Olkin, I. and Pukelsheim, F. (1982). The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Pardo, L. (2018). *Statistical inference based on divergence measures*. CRC press.

Parlett, B. and Landis, T. (1982). Methods for scaling to doubly stochastic form. *Linear Algebra and its Applications*, 48:53–79.

Parthasarathy, T. (1970). On games over the unit square. *SIAM Journal on Applied Mathematics*, 19(2):473–476.

Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the 7th conference of the Cognitive Science Society, University of California, Irvine, CA, USA*, pages 15–17.

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

Pearl, J. et al. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

References

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the royal society of London*, 58(347-352):240–242.

Perry, W. L. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.

Petersen, F., Mukherjee, D., Sun, Y., and Yurochkin, M. (2021). Post-processing for individual fairness. *Advances in Neural Information Processing Systems*, 34:25944–25955.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

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

Polyanskiy, Y. and Wu, Y. (2022). *Information theory: From coding to learning*. Cambridge University Press.

Portnoy, S. and Koenker, R. (1997). The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 12(4):279–300.

Prince, A. E. and Schwarcz, D. (2019). Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Review*, 105:1257.

References

Prokhorov, Y. V. (1956). Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications*, 1(2):157–214.

Proschan, M. A. and Presnell, B. (1998). Expect the unexpected from conditional expectation. *The American Statistician*, 52(3):248–252.

Rao, C. R. and Mitra, S. K. (1972). Generalized inverse of a matrix and its applications. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*, volume 6, pages 601–621. University of California Press.

Reichenbach, H. (1956). *The direction of time*, volume 65. University of California Press.

Rényi, A. (1959). On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pages 547–562. University of California Press.

Rhynhart, R. (2020). Mapping the legacy of structural racism in philadelphia. *Philadelphia, Office pf the Controller*.

Rizzo, M. L. and Székely, G. J. (2016). Energy distance. *wiley interdisciplinary reviews: Computational statistics*, 8(1):27–38.

References

Roberts, H. V. (1968). On the meaning of the probability of rain. In *first national conference on statistical meteorology*.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

Rosenbaum, P. (2018). *Observation and experiment*. Harvard University Press.

Ross, S. M. (2014). *Introduction to probability models*. Academic press.

Rothstein, W. G. (2003). *Public health and the risk factor: A history of an uneven medical revolution*, volume 3. Boydell & Brewer.

Rouvroy, A., Berns, T., and Carey-Libbrecht, L. (2013). Algorithmic governmentality and prospects of emancipation. *Réseaux*, 177(1):163–196.

Rudin, W. (1966). *Real and Complex Analysis*. McGraw-hill New York.

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

Santambrogio, F. (2015). *Optimal transport for applied mathematicians*, volume 55. Springer.

Sarmanov, O. (1958a). Maximum correlation coefficient (non-symmetrical case). *Doklady Akademii Nauk SSSR*, 121(1):52–55.

References

Sarmanov, O. V. (1958b). The maximum correlation coefficient (symmetrical case). *Doklady Akademii Nauk SSSR*, 120(4):715–718.

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.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Shurbet, G., Lewis, T., and Boullion, T. (1974). Quadratic matrix equations. *The Ohio Journal of Science*, 74.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail-but some don't*. Penguin.

Simon, J. (1987). The emergence of a risk society-insurance, law, and the state. *Socialist Review*, (95):60–89.

Simon, J. (1988). The ideological effects of actuarial practices. *Law & Society Review*, 22:771.

Sinkhorn, R. (1962). On the factor spaces of the complex doubly stochastic matrices. *Notices of the American Mathematical Society*, 9:334–335.

References

Sinkhorn, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879.

Sinkhorn, R. (1966). A relationship between arbitrary positive matrices and stochastic matrices. *Canadian Journal of Mathematics*, 18:303–306.

Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281.

Spirtes, P., Glymour, C. N., and Scheines, R. (1993). *Causation, prediction, and search*. Springer Verlag.

Squires, G. D. (2003). Racial profiling, insurance style: Insurance redlining and the uneven development of metropolitan areas. *Journal of Urban Affairs*, 25(4):391–410.

Squires, G. D. and Velez, W. (1988). Insurance redlining and the process of discrimination. *The Review of Black Political Economy*, 16(3):63–75.

Stone, D. A. (1993). The struggle for the soul of health insurance. *Journal of Health Politics, Policy and Law*, 18(2):287–317.

References

Struyck, N. (1912). *Les oeuvres de Nicolas Struyck (1687-1769): qui se rapportent au calcul des chances, à la statistique général, la statistique des décès et aux rentes viagèter*. Société générale néerlandaise d'assurances sur la vie et de rentes viagères.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks.

Székely, G. J. (2003). E-statistics: The energy of statistical samples. *Bowling Green State University, Department of Mathematics and Statistics Technical Report*, 3(05):1–18.

Takatsu, A. (2008). On wasserstein geometry of the space of gaussian measures. *arXiv*, 0801.2250.

Takatsu, A. and Yokota, T. (2012). Cone structure of l2-wasserstein spaces. *Journal of Topology and Analysis*, 4(02):237–253.

The Zebra (2022). Car insurance rating factors by state. *https://www.thezebra.com/*.

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement*, 8(2):63–70.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tikhonov, A. N. (1943). On the stability of inverse problems. In *Doklady Akademii Nauk*, volume 5, pages 195–198.

References

Topkis, D. M. (1998). *Supermodularity and complementarity*. Princeton university press.

Tschantz, M. C. (2022). What is proxy discrimination? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1993–2003.

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

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. *Statistics: applications and new directions (Calcutta, 1981)*, pages 579–604.

Vallender, S. (1974). Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786.

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.

Van Rijsbergen, C. (1979). Information retrieval: theory and practice. In *Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems*, volume 79.

Vapnik, V. (1991). Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4.

References

Verboven, K. (2011). Introduction: Professional collegia: Guilds or social clubs? *Ancient Society*, pages 187–195.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Vogel, R., Bellet, A., Clémen, S., et al. (2021). Learning fair scoring functions: Bipartite ranking under roc-based fairness constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 784–792. PMLR.

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

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

von Neumann, J. (1928). Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320.

Wadsworth, C., Vera, F., and Piech, C. (2018). Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv*, 1807.00199.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

References

Wasserstein, L. N. (1969). Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72.

Watson, D. S., Gultchin, L., Taly, A., and Floridi, L. (2021). Local explanations via necessity and sufficiency: Unifying theory and practice. *Uncertainty in Artificial Intelligence*, pages 1382–1392.

Wilkie, D. (1997). Mutuality and solidarity: assessing risks and sharing losses. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1357):1039–1044.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Wortham, L. (1986). The economics of insurance classification: The sound of one invisible hand clapping. *Ohio State Law Journal*, 47:835.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20.

Wright, S. (1934). The method of path coefficients. *The annals of mathematical statistics*, 5(3):161–215.

Wu, Y., Zhang, L., Wu, X., and Tong, H. (2019). Pc-fairness: A unified framework for measuring causality-based fairness. *Advances in Neural Information Processing Systems*, 32.

Xu, H., Liu, X., Li, Y., Jain, A., and Tang, J. (2021). To be robust or to be fair: Towards fairness in adversarial training. In *International conference on machine learning*, pages 11492–11501. PMLR.

References

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, 75(6):579–652.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. *arXiv*, 1507.05259.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2019). Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1):2737–2778.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018a). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018b). Mitigating Unwanted Biases with Adversarial Learning. *Association for the Advancement of Artificial Intelligence*.

Zolotarev, V. M. (1976). Metric distances in spaces of random variables and their distributions. *Mathematics of the USSR-Sbornik*, 30(3):373.