# From Contemplative to Predictive Modeling (in actuarial science and risk management)

**Arthur Charpentier**, Florent Crouzet & Agathe Fernandes Machado

KU Leuven, 2024


SCOR The Art & Science of Risk | Foundation for science

# Back in June 2015...

In June 2015, Jan invited me to give a Actuarial Contact Program (ACP) talk in Leuven. The take-away slide started with



**Take-Away Conclusion**

"People rarely succeed unless they have fun in what they are doing " D. Carnegie

# Back in June 2015...

In June 2015, Jan invited me to give a Actuarial Contact Program (ACP) talk in Leuven. The take-away slide was claiming that data won't speak for themselves



## Take-Away Conclusion

"People rarely succeed unless they have fun in what they are doing " D. Carnegie

"the numbers have no way of speaking for themselves. We speak for them. ... Before we demand more of our data, we need to demand more of ourselves " N. Silver, in Silver (2012).

# "End of Theory"... and Models ?

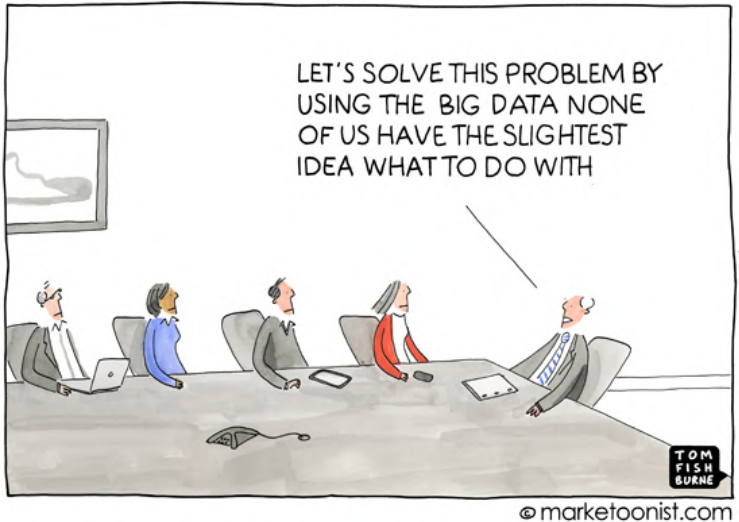
CHRIS ANDERSON SCIENCE JUN 23, 2008 12:00 PM

## The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

Illustration: Marian Bantjes "All models are wrong, but some are useful." So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies [...]

Source: https://www.wired.com/2008/06/pb-theory/

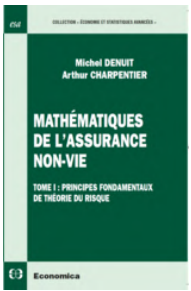
# ... Data can't speak for itself

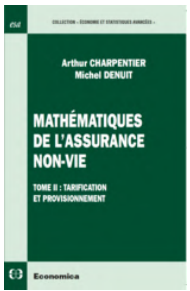


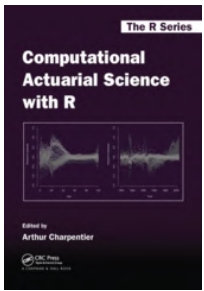

Source: https://marketoonist.com/2014/01/big-data.html

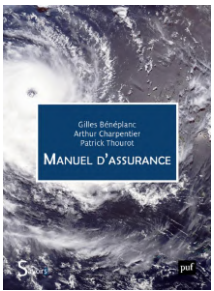
**Arthur Charpentier** Professor at Université du Québec à Montréal

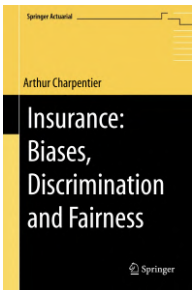
- Denuit and Charpentier (2004, 2005) Mathématiques de l'Assurance Non-Vie,
- Charpentier (2014) Computational Actuarial Science with R,
- Bénéplanc et al. (2022) Manuel d'Assurance,
- Charpentier (2024) Insurance: Biases, Discrimination and Fairness.



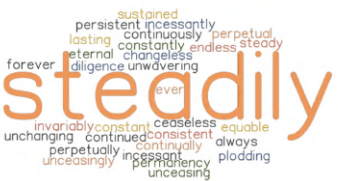



# Disclaimer

"Whereof what's past is prologue," William Shakespeare (1610), The Tempest

(that phrase stands for the idea that history sets the context for the present, see e.g. Murray and Sinnreich (2006))

> **contemplation**: noun, *con·tem·pla·tion*; the act of regarding steadily [MW]

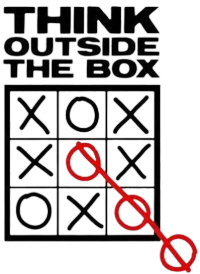



This is a long ongoing work, and, despite my efforts, it might contain errors of any type. Concepts and results presented in those slides are probably either extremely vague, or wrong. All apologies.

# Agenda

"Rara avis in terris nigroque simillima cygno," Decimus Iunius Iuvenalis (82 AD)

"No amount of observations of white swans can allow the inference that all swans are white, but the observation of a single black swan is sufficient to refute that conclusion," John Stuart Mill (1848)




- Statistics rely (a lot) on i.i.d. (stationary) assumption
- Machine learning focuses (mainly) on generalization
- Accuracy is based on strong stability assumptions
- Bias selection can impact interpretation of models
- In many actuarial applications, we know that there are ruptures or changes... what could we do about it ?

# Agenda

- what if we observe $\{x_1, \cdots, x_n\}$ drawn under $\mathbb{P}$ but we need to compute quantities under $\mathbb{Q}$ ?
- what if we were able to estimate $\mathbb{E}_{\mathbb{P}}[Y|\boldsymbol{X}]$ but we want $\mathbb{E}_{\mathbb{Q}}[Y|\boldsymbol{X}]$ ?
- what if $X_1 \not\perp\!\!\!\perp X_2$ under $\mathbb{P}$, can we have $X_1 \perp\!\!\!\perp X_2$ under $\mathbb{Q}$ (fairness)
- a lot of problems in actuarial science can be formalized like that
- brief introduction to "transfer learning"
  - instance transfer (reweighting observations)
  - feature transfer (mapping into a common space)
  - parameter transfer

# Climate, Finance and Insurance

As mentioned in Intergovernmental Panel on Climate Change, page 594

"What does the accuracy of a climate model's simulation of past or contemporary climate say about the accuracy of its projections of climate change? This question is just beginning to be addressed, exploiting the newly available ensembles of models..." Randall et al. (2007)

A standard financial disclaimer, see e.g.,

"Past performance is no guarantee of future returns," Brain (2010)

or in insurance (about wildfire losses in California)

"Looking backward has become less effective in predicting the future," Frazier (2021)

"History Doesn't Repeat Itself, but It Often Rhymes," Mark Twain (1874)

# Motivation, statistics, *rebus sic stantibus*

**Statistics** : *clausula rebus sic stantibus* ("with things thus standing")

> Statistics commonly deals with random samples. A random sample can be thought of as a set of objects that are chosen randomly. More formally, it is "a sequence of independent, identically distributed random data points". (...) Independent and identically distributed random variables are often used as an assumption, which tends to simplify the underlying mathematics. In practical applications of statistical modeling, however, the assumption may or may not be realistic $\mathbb{W}$

Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space,

Let $y_1, y_2, \cdots, y_n$ be $n$ i.i.d. samples of a random variable $Y$ distributed by $\mathbb{P}$

# Motivation, statistics, *rebus sic stantibus*

An important concept in actuarial science is the return period.

"**1.0.1. Conditions.** The aim of a statistical theory of extreme values is to analyze observed extremes and to forecast further extremes. (...) The essential condition in the analysis is the *clausula rebus sic stantibus*," Emil Gumbel (1958), Statistics of Extremes, page 1.

- *rebus sic stantibus* is Latin for "with things thus standing" ("in gelijkblijvende omstandigheden" or "les choses demeurant en l'état")
- *clausula rebus sic stantibus* is the legal doctrine allowing for a contract or a treaty to become inapplicable because of a fundamental change of circumstances,
- maxim *omnis conventio intelligitur rebus sic stantibus* for "every convention is understood with circumstances as they stand", by the Italian jurist Scipione Gentili (1563–1616).

# Motivation, statistics, *rebus sic stantibus*

"The distribution from which the extremes have been drawn and its parameters must remain constant in time (or space), or the influence that time (or space) exercises upon them must be taken into account or eliminated (...) This assumption, made in most statistical work, is hardly ever realized." Emil Gumbel (1958), Statistics of Extremes, page 1.

"**1.0.3. The Flood Problem**. Similar stationary time series may easily be obtained for annual droughts, largest precipitations, snowfalls, maxima and minima of atmospheric pressures and temperatures, and other meteorological phenomena." Emil Gumbel (1958), Statistics of Extremes, page 4.

Gumbel (1941a,b) discussed "the return period of flood flows", term used in Fuller (1914) Hazen (1930), on flood flows.

# Motivation, statistics, *rebus sic stantibus*

**Geometric distribution:** The probability that the first occurrence of success requires $k$ independent trials, each with success probability $p$, the probability that the $k$-th trial is the first success is

$$\mathbb{P}(X = k) = (1-p)^{k-1}p$$

for $k = 1, 2, 3, 4, \cdots$. And then, $\mathbb{E}_{\mathbb{P}}[X] = p^{-1}$.

| $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ | $Y_5$ | $Y_6$ | $Y_7$ | $\cdots$ | $Y_{k-1}$ | $Y_k$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | $\cdots$ | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | $\cdots$ | 0 | 0 |

Task Committee on Hydrology Handbook of Management Group D of ASCE (1996)

# Motivation, statistics, *rebus sic stantibus*

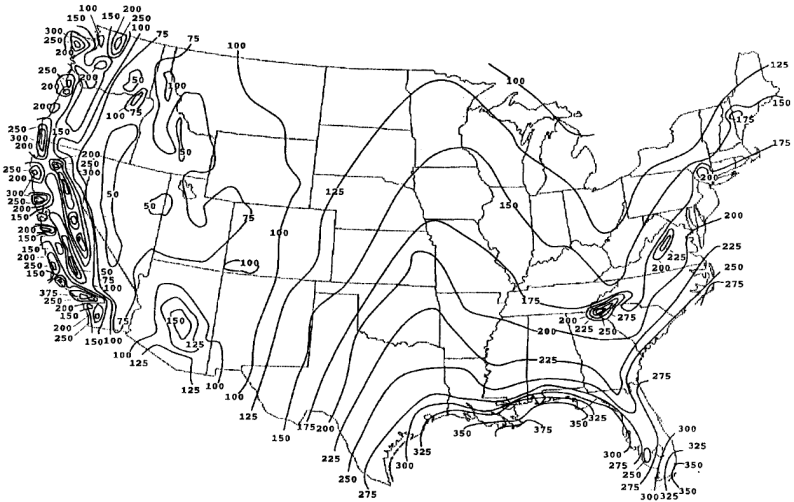


*Figure 2.13.—100-yr 24-hr Precipitation (from Hershfield, 1961)*

# Motivation, statistics, *rebus sic stantibus*

There is also a connection with the law of small numbers,

**Law of small numbers:** In a given period of $n$ years, the probability of a given number $r$ of events of a return period $\mu$ is given by the binomial distribution as follows,

$$\mathbb{P}(X = r) = \binom{n}{r} \mu^r (1-\mu)^{n-r}$$

and if $n \to \infty$ and $\mu \to 0$ in such a way that $n\mu \to \lambda$ then

$$\binom{n}{r} \mu^r (1-\mu)^{n-r} \to e^{-\lambda} \frac{\lambda^r}{r!}.$$

# Motivation, statistics, *rebus sic stantibus*

If $\mu = 1/T$, $\mathbb{P}(\text{no-occurrence in } [0, t]) = e^{-\mu t} = e^{-t/T}$.

$$e^{-1} = 0.3678794 \text{ and } 1 - e^{-1} = 0.6321206$$

> This means, for example, that there is a 63.2% probability of a flood larger than the 50-year return flood to occur within any period of 50 year $\mathbb{W}$

Is this only in the statistical literature ?

See "generalization" in machine learning...

# Motivation, statistics and overfit

In "old school econometrics", consider model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$, or $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ (true residuals).

Ordinary least squares, $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ so that $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ ($\boldsymbol{H}$)

Estimated residuals (estimated residuals) are $\widehat{\boldsymbol{\varepsilon}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}$ (true residuals).

Standardized residuals (studentized residuals) are $\widehat{r}_i = \dfrac{\widehat{\varepsilon}_i}{\widehat{\sigma}\sqrt{1 - H_{i,i}}}$ (estimated residuals).

Externally studentized residual residuals (externally studentized residuals) are $\widehat{t}_i = \dfrac{\widehat{\varepsilon}_i}{\tilde{\sigma}\sqrt{1 - H_{i,i}}} = \dfrac{y_i - \widehat{y}_{(i)}}{s_i}$ (estimated residuals).

# Motivation, statistics and overfit

$\widehat{y}_{(i)}$ is the predicted value for $i$-th point (i.e., $\boldsymbol{x}_i$) when observation $(\boldsymbol{x}_i, y_i)$ is removed from the (training) dataset: leave-one-out cross-validation, or "Jackknife"

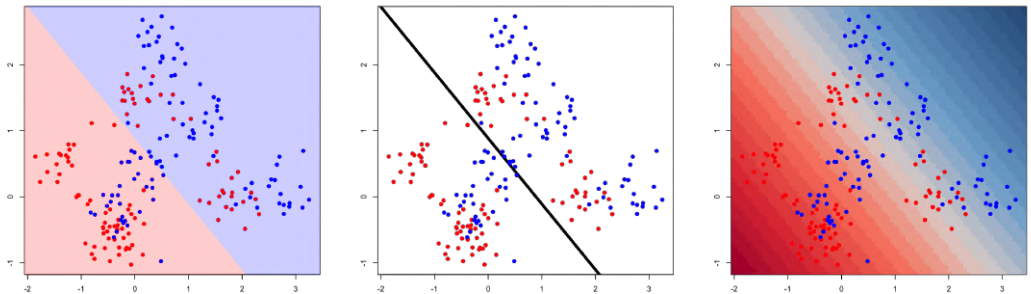
> Given a sample of size $n$, a jackknife estimator can be built by aggregating the parameter estimates from each subsample of size $(n-1)$ obtained by omitting one observation. The jackknife technique was developed by Maurice Quenouille (1924–1973) from 1949 and refined in 1956. John Tukey expanded on the technique in 1958 and proposed the name "jackknife" W

Related to the idea of "cross-validation", Stone (1974)

> Cross-validation is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set (...) It is often used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. It can also be used to assess the quality of a fitted model W

# Motivation, statistics and overfit

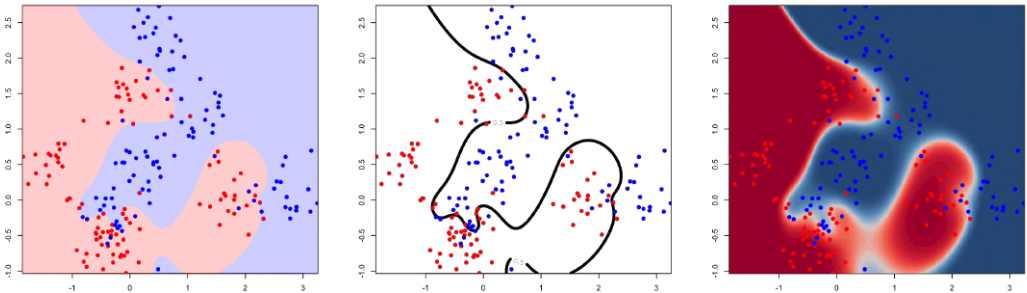
"Under-fit & Over-fit" (and Goldilocks)




Plan (vanilla) GLM (logistic) regression.

# Motivation, statistics and overfit

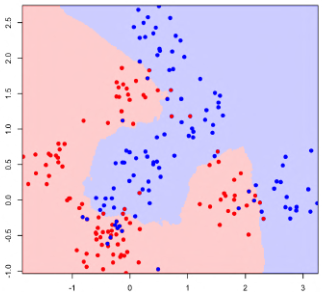
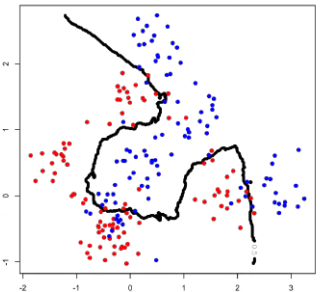
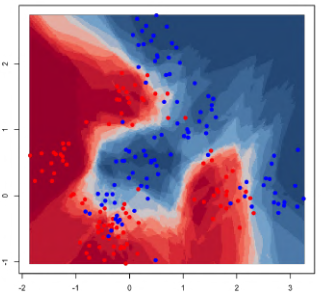
"Under-fit & Over-fit" (and Goldilocks)



GAM (logistic) regression, with splines.

# Motivation, statistics and overfit

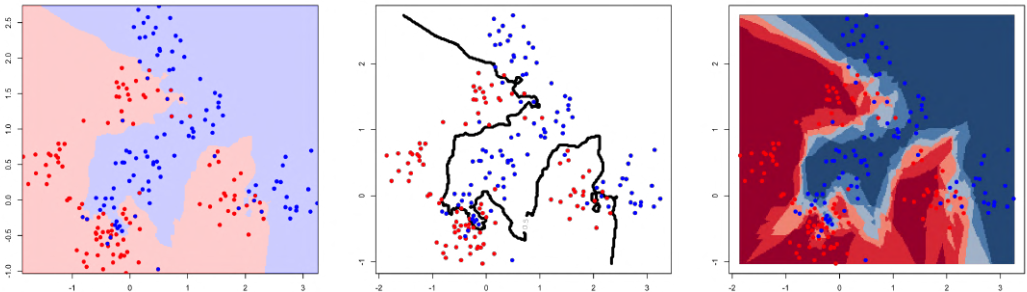
"Under-fit & Over-fit" (and Goldilocks)

$k$-nearest neighbors.

# Motivation, statistics and overfit

"Under-fit & Over-fit" (and Goldilocks)

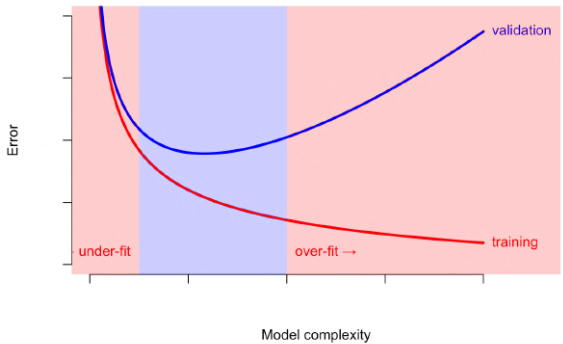


*k*-nearest neighbors.

# Motivation, statistics and overfit

"Under-fit & Over-fit" (and Goldilocks)

> Training error is shown in blue, validation error in red, both as a function of the number of training cycles. If the validation error increases (positive slope) while the training error steadily decreases (negative slope) then a situation of overfitting may have occurred. The best predictive and fitted model would be where the validation error has its global minimum. W



# Motivation, statistics and overfit

> In statistics and machine learning, the bias–variance tradeoff describes the relationship between a model's complexity, the accuracy of its predictions, and how well it can make predictions on previously unseen data that were not used to train the model. In general, as we increase the number of tunable parameters in a model, it becomes more flexible, and can better fit a training data set. It is said to have lower error, or bias. However, for more flexible models, there will tend to be greater variance to the model fit each time we take a set of samples to create a new training data set. It is said that there is greater variance in the model's estimated parameters. W

To compute biases and variances, we need some model...

> Overfitting is more likely to be a serious concern when there is little theory available to guide the analysis, W

# Motivation, machine learning, generalization

**Machine learning** : a key concept is generalization,

"The generalization performance of a learning method relates to its prediction capability on independent test data," section 7.1 Hastie et al. (2009)

> A central goal in designing a machine learning system is to guarantee that the learning algorithm will generalize, or perform accurately on new examples after being trained on a finite number of them. W

**Law of large numbers:** If $X, X_1, X_2, \cdots, X_n, \cdots$ are i.i.d. samples of a random variable distributed according to $\mathbb{P}$, then for any (small) positive non-zero value $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\mathbb{E}_{\mathbb{P}}[X] - \frac{1}{n}\sum_{i=1}^{n} X_i\right| > \epsilon\right] = 0$$

# Motivation, machine learning, generalization

To minimize the discrepancy between training and generalization errors, it is essential to understand the implications of the law of large numbers.
This understanding is facilitated by concentration inequalities, which provide a quantitative measure of how much random variables deviate from their expected values.

**Höffding's inequality:** If $X, X_1, X_2, \cdots, X_n, \cdots$ are i.i.d. samples of a random variable distributed according to $\mathbb{P}$, such that $\mathbb{P}(X_i \in [a, b]) = 1$, then for any (small) positive non-zero value $\epsilon > 0$:

$$\mathbb{P}\left[\left|\mathbb{E}_{\mathbb{P}}[X] - \frac{1}{n}\sum_{i=0}^{n} X_i\right| > \epsilon\right] \leq 2\exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right)$$

Note that we can write

$$\mathbb{P}\left[\left|\mathbb{E}_{\mathbb{P}}[X] - \frac{1}{n}\sum_{i=0}^{n} X_i\right| > (b-a)\sqrt{\frac{-1}{2n}\log(2\delta)}\right] \leq \delta$$

# Motivation, machine learning, generalization

For binary and 0/1 loss,

$$\text{for a given } m \in \mathcal{M},\ \mathcal{R}(m) - \widehat{\mathcal{R}}_n(m) \sim \frac{1}{\sqrt{n}}$$

and we have the following "worst case scenario"

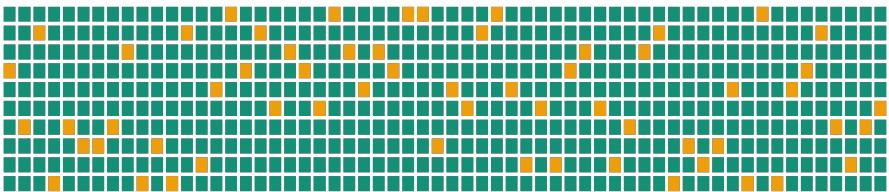
$$\sup_{m \in \mathcal{M}} \left\{ \mathcal{R}(m) - \widehat{\mathcal{R}}_n(m) \right\} \sim \frac{\log(\text{Card}(\mathcal{M}))}{\sqrt{n}}$$

see Bousquet et al. (2003). And if $\mathcal{M}$ is infinite, see Vapnik-Chervonenkis (VC) dimension, from Vapnik and Chervonenkis (1971).

# Motivation, Time Series

On i.i.d. data, standard to use *k*-fold cross validation

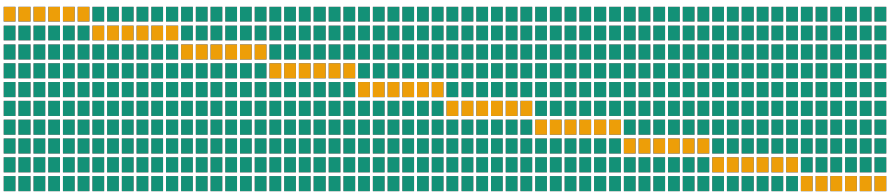
"This approach involves randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining k-1 folds," James et al. (2013)



# Motivation, Time Series

Standard approach, shuffle and then create $k$ groups (sorted)

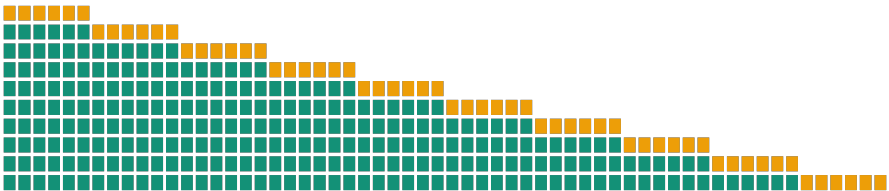
”there is a bias-variance trade-off associated with the choice of k in k-fold cross-validation. Typically, given these considerations, one performs k-fold cross-validation using k = 5 or k = 10, as these values have been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance,” James et al. (2013)



# Motivation, Time Series

With cross validation, it is possible to "adapt" and take into account the dynamics,

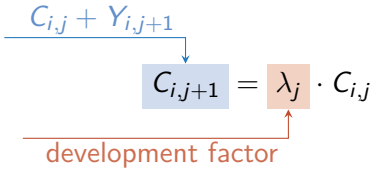
for time series, use "forward-validation", Hjorth (1982, 1994), or "rolling cross-validation, Bergmeir and Benítez (2012)



# Motivation, actuarial science

**Claims reserving** : assume that dynamic of past payments

- Astesan (1938)'s chain ladder
- $i$ accident year (cohort)
- $j$ development year
- $i+j$ calendar year




> The primary underlying assumption of the chain-ladder method is that historical loss development patterns are indicative of future loss development patterns. W

- Mack (1991)'s log-Poisson model, $Y_{i,j} \sim \mathcal{P}(\mu_{i,j})$, $\mu_{i,j} = e^{a_i+b_j}$
- Verrall (1996)'s additive version $Y_{i,j} \sim \mathcal{P}(\mu_{i,j})$, $\mu_{i,j} = e^{a(i)+b(j)}$

# Motivation, actuarial science

Observed incremental payment

$Y_{i,j}$, where $i + j \leq n$

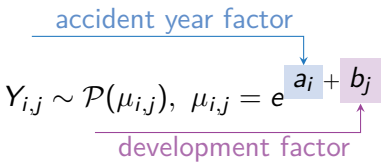
| | | | | |
|---|---|---|---|---|
| $Y_{1,0}$ | $Y_{1,1}$ | | $Y_{1,n-2}$ | $Y_{1,n-1}$ |
| $Y_{2,0}$ | $Y_{2,1}$ | | $Y_{2,n-2}$ | |
| | | | | |
| $Y_{n-1,0}$ | $Y_{n-1,1}$ | | | |
| $Y_{n,0}$ | | | | |

| | | | | |
|---|---|---|---|---|
| $Y_{1,0}$ | $Y_{1,1}$ | | $Y_{1,n-2}$ | $Y_{1,n-1}$ |
| $Y_{2,0}$ | $Y_{2,1}$ | | $Y_{2,n-2}$ | |
| | | | | |
| $Y_{n-1,0}$ | $Y_{n-1,1}$ | | | |
| $Y_{n,0}$ | | | | |

Future incremental payment

$Y_{i,j}$, where $i + j > n$, to be predicted

# Motivation, actuarial science




| | | | | | | |
|---|---|---|---|---|---|---|
| $a_1$ | $Y_{1,0}$ | $Y_{1,1}$ | | $Y_{1,n-2}$ | $Y_{1,n-1}$ | |
| $a_2$ | $Y_{2,0}$ | $Y_{2,1}$ | | $Y_{2,n-2}$ | | |
| | | | | | | |
| $a_{n-1}$ | $Y_{n-1,0}$ | $Y_{n-1,1}$ | | | | |
| $a_n$ | $Y_{n,0}$ | | | | | |

# Motivation, actuarial science

accident year factor

$$Y_{i,j} \sim \mathcal{P}(\mu_{i,j}),\ \mu_{i,j} = e^{a_i + b_j}$$

development factor

Once we have estimates, $\widehat{a}_i$ and $\widehat{b}_j$, $i, j = 1, \cdots, n$,

$$\widehat{Y}_{i,j} = e^{\widehat{a}_i + \widehat{b}_j}$$

| $b_0$ | $b_1$ | | $b_{n-1}$ | $b_n$ | |
|---|---|---|---|---|---|
| $Y_{1,0}$ | $Y_{1,1}$ | | $Y_{1,n-2}$ | $Y_{1,n-1}$ | |
| $Y_{2,0}$ | $Y_{2,1}$ | | $Y_{2,n-2}$ | | |
| | | | | | |
| $Y_{n-1,0}$ | $Y_{n-1,1}$ | | | | |
| $Y_{n,0}$ | | | | | |

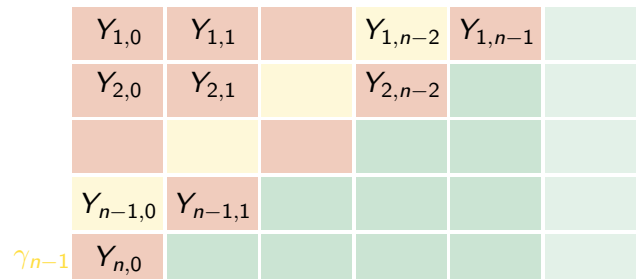
# Motivation, actuarial science

- Taylor (1977)'s separation method

"It is crucial to the logic underlying the chain-ladder method that the "exogeneous influences" should not be too great (...) Clearly, it would be preferable to separate, if possible, the basic stationary claim delay distribution from the exogeneous influences which are upsetting the stationarity"

$$Y_{i,j} \sim \mathcal{P}(\mu_{i,j}),\ \mu_{i,j} = e^{a_i + \gamma_{i+j}}$$

$$\widehat{Y}_{i,j} = e^{\widehat{a}_i + \widehat{\gamma}_{i+j}}$$



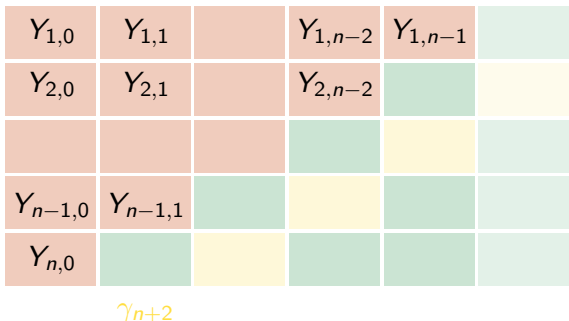
# Motivation, actuarial science

- Taylor (1977)'s separation method

$$Y_{i,j} \sim \mathcal{P}(\mu_{i,j}),\ \mu_{i,j} = e^{a_i + \gamma_{i+j}}$$

$$\widehat{Y}_{i,j} = e^{\widehat{a}_i + \widehat{\gamma}_{i+j}}$$

what is $\widehat{\gamma}_{i+j}$ when $i + j > n$ ?



# Motivation, actuarial science

- Quarg and Mack (2004)'s Munich chain ladder, learn from both (accumulated) payments and incurred estimates

| | | | | |
|---|---|---|---|---|
| $C_{1,0}$ | $C_{1,1}$ | | $C_{1,n-2}$ | $C_{1,n-1}$ |
| $C_{2,0}$ | $C_{2,1}$ | | $C_{2,n-2}$ | |
| | | | | |
| $C_{n-1,0}$ | $C_{n-1,1}$ | | | |
| $C_{n,0}$ | | | | |

| | | | | |
|---|---|---|---|---|
| $I_{1,0}$ | $I_{1,1}$ | | $I_{1,n-2}$ | $I_{1,n-1}$ |
| $I_{2,0}$ | $I_{2,1}$ | | $I_{2,n-2}$ | |
| | | | | |
| $I_{n-1,0}$ | $I_{n-1,1}$ | | | |
| $I_{n,0}$ | | | | |

# Motivation, actuarial science

- Quarg and Mack (2004)'s Munich chain ladder, learn from both (cumulated) payments and incurred estimates

| | | | | |
|---|---|---|---|---|
| $C_{1,0}$ | $C_{1,1}$ | | $C_{1,n-2}$ | $C_{1,n-1}$ |
| $C_{2,0}$ | $C_{2,1}$ | | $C_{2,n-2}$ | |
| | | | | |
| $C_{n-1,0}$ | $C_{n-1,1}$ | | | |
| $C_{n,0}$ | | | | |

| | | | | |
|---|---|---|---|---|
| $I_{1,0}$ | $I_{1,1}$ | | $I_{1,n-2}$ | $I_{1,n-1}$ |
| $I_{2,0}$ | $I_{2,1}$ | | $I_{2,n-2}$ | |
| | | | | |
| $I_{n-1,0}$ | $I_{n-1,1}$ | | | |
| $I_{n,0}$ | | | | |

Constraint, $C_{i,n-1} = I_{i,n-1}$, $\forall i$

# Motivation, actuarial science

**Static life table** :

- Denuit and Robert (2007) or Pitacco et al. (2009)

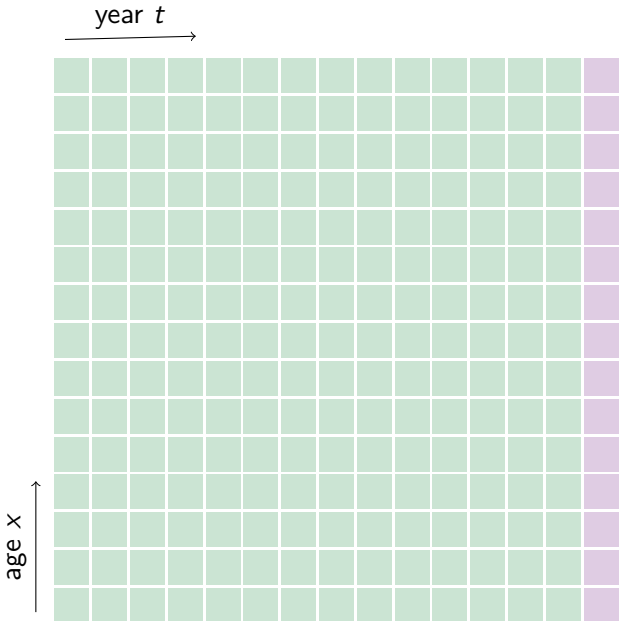
Exposure $E_{x,t}$

Deaths $D_{x,t} = E_{x,t} - E_{x+1,t+1}$

At $t$, fixed $q_x = \dfrac{D_{x,t}}{E_{x,t}} = \mu_x$

$L_{x+1} = L_x \cdot (1 - q_x)$

> The period life table represents mortality rates during a specific time period for a certain population $\mathbb{W}$



# Motivation, actuarial science

**Prospective (cohort) life table** :
Lee and Carter (1992) or
Pitacco et al. (2009)

rate of change per age

$$\log[\mu_{x,t}] = a_x + b_x \cdot \kappa_t$$

average rate — time index

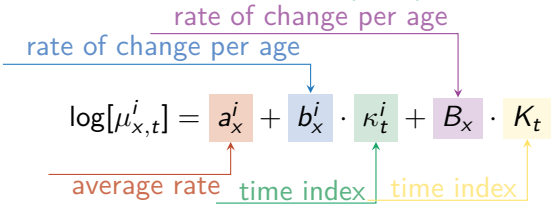
$$\mu_{x,t} = \frac{D_{x,t}}{E_{x,t}} = \frac{E_{x,t} - E_{x+1,t+1}}{E_{x,t}}$$

We need $\widehat{\kappa}_t$ for $t >$ today.

# Motivation, actuarial science

> A cohort life table, often referred to as a generation life table, is used to represent the overall mortality rates of a certain population's entire lifetime. W

**Multiple prospective life table** : Li and Lee (2005)




"Mortality patterns and trajectories in closely related populations are likely to be similar in some respects, and differences are unlikely to increase in the long run. It should therefore be possible to improve the mortality forecasts for individual countries by taking into account the patterns in a larger group," Li and Lee (2005)

# Motivation, climate change

A personal take on science and society

## World view

By Gavin Schmidt

## Why 2023's heat anomaly is worrying scientists

**Climate models struggle to explain why planetary temperatures spiked suddenly. More and better data are urgently needed.**

When I took over as the director of NASA's Goddard Institute for Space Studies, I inherited a project that tracks temperature changes since 1880. Using this trove of data, I've made climate predictions at the start of every year since 2016. It's humbling, and a bit worrying, to admit that no year has confounded climate scientists' predictive capabilities more than 2023 has.

For the past nine months, mean land and sea surface

> "If the anomaly does not stabilize by August, then the world will be in uncharted territory."

from stratospheric water vapour, and the ramping up of solar activity in the run-up to a predicted solar maximum. But these factors explain, at most, a few hundredths of a degree in warming (Schoeberl, M. R. *et al. Geophys. Res. Lett.* **50**, e2023GL104634; 2023). Even after taking all plausible explanations into account, the divergence between expected and observed annual mean temperatures in 2023 remains about 0.2 °C – roughly the gap between the previous and current annual record.

There is one more factor that could be playing a part. In 2020, new regulations required the shipping industry to use cleaner fuels that reduce sulfur emissions. Sulfur compounds in the atmosphere are reflective and influence several properties of clouds, thereby having

**Climate**, how to predict in "uncharted territory", Schmidt (2024)?

# Motivation, climate change

> A wildfire (or forest fire, bushfire) is an unplanned, uncontrolled and unpredictable fire in an area of combustible vegetation. W

**Climate risk in California (U.S.)**

"Why is it illegal in California to consider climate-informed catastrophe models when setting wildfire insurance premiums?" Frazier (2021)

Some general context:

California Code Of Regulations, title 10, Chapter 5 (Insurance Commissioner), § 2644 ("Determination of Reasonable Rates")

Cal. Code Regs. tit. 10 § 2644.4 (Projected Losses)

"Projected losses" means the insurer's historic losses per exposure, adjusted by catastrophe adjustment, as prescribed in section 2644.5.

# Motivation, climate change

Cal. Code Regs. tit. 10 § 2644.5 (Catastrophe Adjustment)

In those insurance lines and coverages where catastrophes occur, the catastrophic losses of any one accident year in the recorded period are replaced by a loading based on a multi-year, long-term average of catastrophe claims. The number of years over which the average shall be calculated shall be at least 20 years for homeowners multiple peril fire, and at least 10 years for private passenger auto physical damage. Where the insurer does not have enough years of data, the insurer's data shall be supplemented by appropriate data. The catastrophe adjustment shall reflect any changes between the insurer's historical and prospective exposure to catastrophe due to a change in the mix of business. There shall be no catastrophe adjustment for private passenger auto liability.

# Motivation, climate change

**Climate risk in France**,

> Subsidence is a general term for downward vertical movement of the Earth's surface, which can be caused by both natural processes and human activities. W

"To determine whether a drought episode is considered abnormal, the SWI established for a given month is compared to the indicators for that same month over the previous 50 years. It is considered "abnormal" if the indicator presents a return period greater or equal to 25 years," Charpentier et al. (2022)[1]

[1] not sure how to define properly a "25 year return period" for non-stationary time series, Olsen et al. (1998), Salas and Obeysekera (2014), Read and Vogel (2015), Du et al. (2015)

# Motivation, is-ought (and algorithmic fairness)

- **Fairness**, of predictive models, Charpentier (2024),

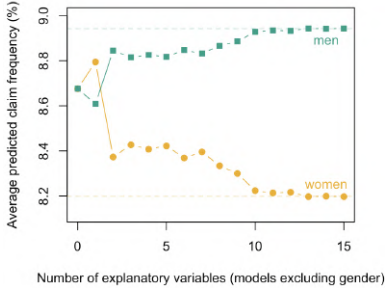
"is-ought" problem, David Hume (1739, 1748)

> The is–ought problem arises when one makes claims about what ought to be that are based solely on statements about what is (...) an ethical or judgmental conclusion cannot be inferred from purely descriptive factual statements. W

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

Logistic regression on $k$ variables excluding gender.

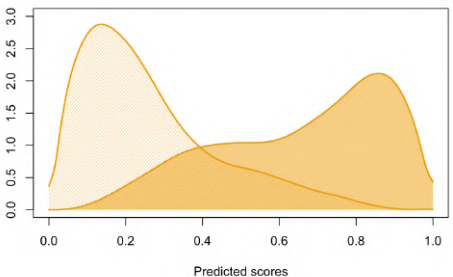
| | men | women |
|---|---|---|
| $k = 0$ | 8.68% | 8.68% |
| $k = 2$ | 8.85% | 8.37% |
| $k = 8$ | 8.87% | 8.33% |
| $k = 15$ | 8.94% | 8.20% |
| empirical | 8.94% | 8.20% |

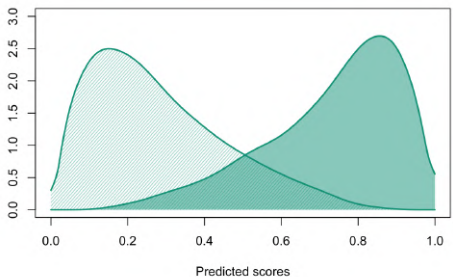


# Motivation, is-ought (and algorithmic fairness)

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

What if distributions of scores, conditional on $S_1$ or $S_2$ (two protected attributes) are significantly different ? e.g., $S_1 \in \{\text{man}, \text{woman}\}$ and $S_2 \in \{\text{white}, \text{black}\}$.



# Motivation, is-ought (and algorithmic fairness)

Mitigation of discrimination (or unequal treatment) has a lot to do with the "is-ought" problem

Supreme Court Justice Harry Blackmun stated, in 1978, "in order to get beyond racism, we must first take account of race. There is no other way. And in order to treat some persons equally, we must treat them differently," cited in Knowlton (1978), as mentioned in Lippert-Rasmussen (2020)

To quote another Supreme Court Justice, in 2007, John G. Roberts of the US Supreme Court submits: "The way to stop discrimination on the basis of race is to stop discriminating on the basis of race" Turner (2015) and Sabbagh (2007)

# Selection Biais & Observational Data

- Selection bias

> The phrase "selection bias" most often refers to the distortion of a statistical analysis, resulting from the method of collecting samples. If the selection bias is not taken into account, then some conclusions of the study may be false. W

Classical econometric problem, see Heckman (1974, 1976, 1979). Suppose

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

but either observed $z_i = 1$, or not $z_i = 0$. Suppose

$$z_i = \begin{cases} 1 \text{ if } z_i^\star = \boldsymbol{x}_i^\top \boldsymbol{\gamma} + u_i > 0 \\ 0 \text{ if } z_i^\star = \boldsymbol{x}_i^\top \boldsymbol{\gamma} + u_i \leq 0 \end{cases}$$

where $U \sim \mathcal{N}(0, 1)$.

# Selection Biais & Observational Data

$z_i = 1$ (not censored) when $u_i \geq -\boldsymbol{x}_i^\top \boldsymbol{\gamma}$, i.e.

$$\mathbb{P}(Z_i = 1) = \mathbb{P}(u_i \geq -\boldsymbol{x}_i^\top \boldsymbol{\gamma}) = 1 - \Phi(-\boldsymbol{x}_i^\top \boldsymbol{\gamma}) = \Phi(\boldsymbol{x}_i^\top \boldsymbol{\gamma}).$$

Suppose

$$\begin{pmatrix} U \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & \sigma^2 \end{pmatrix} \right)$$

i.e. $\mathbb{E}(Y_i | Y_i \text{ observed}) = \mathbb{E}(Y_i | Z_i^\star > 0) = \mathbb{E}(Y_i | U_i > -\boldsymbol{x}_i^\top \boldsymbol{\gamma})$,

$$\mathbb{E}(Y_i | Y_i \text{ observed}) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \mathbb{E}(\varepsilon_i | U_i > -\boldsymbol{x}_i^\top \boldsymbol{\gamma}) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \rho\sigma \frac{\phi(\boldsymbol{x}_i^\top \boldsymbol{\gamma})}{\Phi(\boldsymbol{x}_i^\top \boldsymbol{\gamma})}$$

If $\rho = 0$, no endogenous selection effects.

inverse Mills ratio

See Charpentier and Geoffard (2024) on bodily injury claims, in France (trial or insurance).

# Selection Biais & Observational Data

- Propensity score

> The "propensity" describes how likely a unit is to have been treated, given its covariate values. The stronger the confounding of treatment and covariates, and hence the stronger the bias in the analysis of the naive treatment effect, the better the covariates predict whether a unit is treated or not. By having units with similar propensity scores in both treatment and control, such confounding is reduced. $\mathbb{W}$

"The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates", Rosenbaum and Rubin (1983)

Suppose observed data are $\{(\boldsymbol{x}_i, a_i, y_i)\}_{i=1}^n$ drawn i.i.d (independent and identically distributed) from unknown distribution $\mathbb{P}$, where $A \in \{0, 1\}$, denotes either "control" (placebo) or "treated" (medicine).

# Selection Biais & Observational Data

Let $Y(a)$ (or $Y(x, a)$) denote "potential outcomes" (under control and treatment).

In many application, the quantity of interest is $TE$ (or $TE(x)$) the treatment effect, $TE = Y(1) - Y(0)$

| | Name | Treatment | Outcome (Weight) | | | | Gender | Height | ⋯ |
|---|---|---|---|---|---|---|---|---|---|
| | | $a_i$ | $y_i$ | $y_i(0)$ | $y_i(1)$ | $TE$ | $x_{1,i}$ | $x_{2,i}$ | ⋯ |
| 1 | Alex | 0 | 75 | 75 | ? | ? | H | 172 | ⋯ |
| 2 | Betty | 1 | 52 | ? | 52 | ? | F | 161 | ⋯ |
| 3 | Beatrix | 1 | 57 | ? | 57 | ? | F | 163 | ⋯ |
| 4 | Ahmad | 0 | 78 | 78 | ? | ? | H | 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).

# Selection Biais & Observational Data

When $a_i = 1$ is observed, and $\boldsymbol{x}_i$,

$$\begin{cases} \text{observation} & : y_i(1) \\ \text{counterfactual} & : y_i(0) \end{cases}$$

Following Holland (1986), given a "treatment" $T$ (here $A$), the average treatment effect on outcome $y$ is

$$\tau = \text{ATE} = \mathbb{E}\big[Y(1) - Y(0)\big],$$

and following Wager and Athey (2018), given a treatment $a$, the conditional average treatment effect on outcome $y$, given some covariates $\boldsymbol{x}$ ,is

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

# Selection Biais & Observational Data

- Strongly ignorable treatment assignment

Treatment assignment is said to be strongly ignorable if the potential outcomes are independent of treatment ($A$) conditional on background variables $\boldsymbol{X}$

$$(Y(0), Y(1)) \perp\!\!\!\perp A \mid \boldsymbol{X}$$

- Balancing score

Following Rubin (1973, 1974), a balancing score $b(\boldsymbol{X})$ is a function of the observed covariates $\boldsymbol{X}$ such that the conditional distribution of $\boldsymbol{X}$ given $b(\boldsymbol{X})$ is the same for treated ($A = 1$) and control ($A = 0$) units

$$A \perp\!\!\!\perp \boldsymbol{X} \mid b(\boldsymbol{X})$$

# Selection Biais & Observational Data

- Propensity score

$$e(\boldsymbol{x}) = \mathbb{P}(A = 1 | \boldsymbol{A} = \boldsymbol{x})$$

As proved in Rosenbaum and Rubin (1983),

- the propensity score $e(\boldsymbol{x})$ is a balancing score
- if treatment assignment is strongly ignorable given $\boldsymbol{x}$ then, it is also strongly ignorable given any balancing function (specifically, given the propensity score)

$$(Y(0), Y(1)) \perp\!\!\!\perp A \mid e(\boldsymbol{X}).$$

# Selection Biais & Observational Data

- Horvitz -Thompson theory

> One very early weighted estimator is the Horvitz–Thompson estimator of the mean. When the sampling probability is known, from which the sampling population is drawn from the target population, then the inverse of this probability is used to weight the observations. This approach has been generalized to many aspects of statistics under various frameworks. In particular, there are weighted likelihoods, weighted estimating equations, and weighted probability densities from which a majority of statistics are derived. $\mathbb{W}$

Suppose observed data are $\{(\boldsymbol{X}_i, A_i, Y_i)\}_{i=1}^n$ drawn i.i.d (independent and identically distributed) from unknown distribution $\mathbb{P}$, where $A \in \{0, 1\}$.

# Selection Biais & Observational Data

Suppose observed data are $\{(\boldsymbol{X}_i, A_i, Y_i)\}_{i=1}^n$ drawn i.i.d (independent and identically distributed) from unknown distribution $\mathbb{P}$, where $A \in \{0, 1\}$.

On can derive an Inverse Probability Weighted Estimator (*IPWE*)

- $\mu_a = \mathbb{E}\left[\dfrac{\mathbf{1}_{A=a}Y}{p(A=a|\boldsymbol{X})}\right]$ where $p(a|\boldsymbol{x}) = \mathbb{P}(A=a|\boldsymbol{X}=\boldsymbol{x}) = \dfrac{\mathbb{P}(A=a, \boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(\boldsymbol{X}=\boldsymbol{x})}$
- estimate $p(a|\boldsymbol{x})$ with $\hat{p}_n(a|\boldsymbol{x})$, using any propensity model (e.g., logistic regression model)
- $\hat{\mu}_{a,n}^{IPWE} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \dfrac{y_i 1_{A_i=a}}{\hat{p}_n(a_i|\boldsymbol{x}_i)}$

# Selection Biais & Observational Data

We make the following assumptions.

(A1) Consistency: $Y = Y(A)$

(A2) No un-measured confounders: $\{Y(0), Y(1)\} \perp\!\!\!\perp A|\boldsymbol{X}$.

More formally, for each bounded and measurable functions $f$ and $g$,

$$\mathbb{E}_{(A,Y)}\left[f(Y(\boldsymbol{X}, A))\, g(A) \,|\, \boldsymbol{X}\right] = \mathbb{E}_{Y}\left[f(Y(\boldsymbol{X}, A)) \,|\, \boldsymbol{X}\right] \cdot \mathbb{E}_{A}\left[g(A) \,|\, \boldsymbol{X}\right].$$

This means that treatment assignment is based solely on covariate data and independent of potential outcomes.

(A3) Positivity: $\mathbb{P}(A = a|\boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}_{A}[\mathbf{1}(A = a) \,|\, \boldsymbol{X} = \boldsymbol{x}] > 0$ for all $a$ and $\boldsymbol{x}$.

# Selection Biais & Observational Data

$$\mathbb{E}[Y^*(a)] = \mathbb{E}_{(X,Y)}[Y(X,a)] = \mathbb{E}_{(X,A,Y)}\left[\frac{Y\mathbf{1}(A=a)}{P(A=a|X)}\right]$$

from (A1)

$$\mathbb{E}_{(X,Y)}[Y(X,a)] = \mathbb{E}_X[\mathbb{E}_Y[Y(X,a)\,|\,X]].$$

then simply (by (A3) $\mathbb{E}_A[\mathbf{1}(A=a)\,|\,\boldsymbol{X}]>0$)

$$\mathbb{E}_Y[Y(\boldsymbol{X},a)\,|\,X] = \frac{\mathbb{E}_Y[Y(\boldsymbol{X},a)\,|\,X]\,\mathbb{E}_A[\mathbf{1}(A=a)\,|\,\boldsymbol{X}]}{\mathbb{E}_A[\mathbf{1}(A=a)\,|\,\boldsymbol{X}]} = \frac{\mathbb{E}_{(A,Y)}[Y(\boldsymbol{X},a)\mathbf{1}(A=a)\,|\,\boldsymbol{X}]}{\mathbb{E}[\mathbf{1}(A=a)\,|\,\boldsymbol{X}]}$$

i.e.

$$\mathbb{E}_Y[Y(\boldsymbol{X},a)\,|\,\boldsymbol{X}] = \mathbb{E}_{(A,Y)}\left[\frac{Y(\boldsymbol{X},a)\mathbf{1}(A=a)}{\mathbb{E}[\mathbf{1}(A=a)\,|\,\boldsymbol{X}]}\,\middle|\,\boldsymbol{X}\right]$$

The Inverse Probability Weighted Estimator (*IPWE*) is known to be unstable if some estimated propensities are too close to 0 or 1 (see calibration issues).

# Selection Biais & Observational Data

Augmented Inverse Probability Weighted Estimator (*AIPWE*), Cao et al. (2009)

$$\begin{aligned}\hat{\mu}_{a,n}^{AIPWE} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i 1_{A_i=a}}{\hat{p}_n(A_i|X_i)} - \frac{1_{A_i=a}-\hat{p}_n(A_i|X_i)}{\hat{p}_n(A_i|X_i)}\hat{Q}_n(X_i,a)\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1_{A_i=a}}{\hat{p}_n(A_i|X_i)}Y_i + (1-\frac{1_{A_i=a}}{\hat{p}_n(A_i|X_i)})\hat{Q}_n(X_i,a)\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Q}_n(X_i,a)\right) + \frac{1}{n}\sum_{i=1}^{n}\frac{1_{A_i=a}}{\hat{p}_n(A_i|X_i)}\left(Y_i-\hat{Q}_n(X_i,a)\right)\end{aligned}$$

here we need a regression estimator $\hat{Q}_n(\boldsymbol{x}, a)$ to predict outcome $Y$ based on covariates $\boldsymbol{X}$ and treatment $A$, for some subject $i$.

This approach is said to by "doubly robust" (with a second order bias)

# Post Stratification and Weights

Inspired from techniques used in sampling theory, use post-stratification techniques, which is standard when dealing with a "biased sample".

The regression function is defined a

$$\mu(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}}[Y|\boldsymbol{X}=\boldsymbol{x}] = \mathbb{E}\big[\mathbb{E}_{\mathbb{P}}[Y|\boldsymbol{X}=\boldsymbol{x}, A]\big] = \int_{\mathcal{A}} \mathbb{E}_{\mathbb{P}}[Y|\boldsymbol{X}=\boldsymbol{x}, A=a]\mathrm{d}\mathbb{P}[A=a].$$

Following Moodie and Stephens (2022), the later can be written

$$\mu(\boldsymbol{x}) = \int_{\mathcal{A}} \mathbb{E}_{\mathbb{P}}[Y \cdot W|\boldsymbol{X}=\boldsymbol{x}, A=a]\mathrm{d}\mathbb{P}[A=a|\boldsymbol{X}=\boldsymbol{x}] = \mathbb{E}_{\mathbb{P}}[Y \cdot W|\boldsymbol{X}=\boldsymbol{x}],$$

where $W$ is a version of the Radon-Nikodym derivative

$$W = \frac{\mathrm{d}\mathbb{P}[A=a]}{\mathrm{d}\mathbb{P}[A=a|\boldsymbol{X}=\boldsymbol{x}]},$$

corresponding to the change of measure that will give independence between $\boldsymbol{X}$ and $A$.

# Post Stratification and Weights

- Properties of $W$

We have the following interesting property: let $W$ be a version of the Radon-Nikodym derivative

$$W = \frac{\mathrm{d}\mathbb{P}[A = a]}{\mathrm{d}\mathbb{P}[A = a|\boldsymbol{X} = \boldsymbol{x}]},$$

then $\mathbb{E}_{\mathbb{P}}[W] = 1$, $\mathbb{E}_{\mathbb{P}}[A \cdot W] = \mathbb{E}_{\mathbb{P}}[A]$ and $\mathbb{E}_{\mathbb{P}}[\boldsymbol{X} \cdot W] = \mathbb{E}_{\mathbb{P}}[\boldsymbol{X}]$.
As proved in Fong et al. (2018),

$$\mathbb{E}_{\mathbb{P}}[W] = \iint w \mathrm{d}\mathbb{P}[A = a, \boldsymbol{X} = \boldsymbol{x}] = \iint w \mathrm{d}\mathbb{P}[A = a|\boldsymbol{X} = \boldsymbol{x}] \mathrm{d}\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}]$$

that can be written

$$\mathbb{E}_{\mathbb{P}}[W] = \iint \frac{\mathrm{d}\mathbb{P}[A = a]}{\mathrm{d}\mathbb{P}[A = a|\boldsymbol{X} = \boldsymbol{x}]} \mathrm{d}\mathbb{P}[A = a|\boldsymbol{X} = \boldsymbol{x}] \mathrm{d}\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}],$$

# Post Stratification and Weights

and therefore

$$\mathbb{E}_{\mathbb{P}}[W] = \iint \mathrm{d}\mathbb{P}[A = a]\mathrm{d}\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}] = 1.$$

Similarly

$$\mathbb{E}_{\mathbb{P}}[A \cdot W] = \iint sw\mathrm{d}\mathbb{P}[A = a, \boldsymbol{X} = \boldsymbol{x}] = \iint sw\mathrm{d}\mathbb{P}[A = a|\boldsymbol{X} = \boldsymbol{x}]\mathrm{d}\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}],$$

and

$$\mathbb{E}_{\mathbb{P}}[A \cdot W] = \iint s\mathrm{d}\mathbb{P}[A = a]\mathrm{d}\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}] = \int \mathbb{E}_{\mathbb{P}}[S]\mathrm{d}\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}_{\mathbb{P}}[S].$$

In statistics, this Radon-Nikodym derivative is related to the propensity score, as discussed in Freedman and Berk (2008), Li and Li (2019) and Karimi et al. (2022).

# Censoring, Kaplan-Meier and Reweighting

Classical problem in survival analysis, $t$ is (true) failure times, $a \in \{0, 1\}$ denotes censoring, and $c$ censored time.

$$a_i = \mathbf{1}(t_i \leq c_i) \text{ and we observe } y_i = \min\{t_i, c_i\}.$$

Let $S(t) = \mathbb{P}[T > t]$ and $K(t) = \mathbb{P}[C > t]$.
Observations are $(y_i, a_i)$. Suppose random censoring.
Following Kaplan and Meier (1958), recall that the survival function satisfies
$S(t+1) = q(t+1) \cdot S(t)$, for $t \in \mathbb{N}$

$$\begin{aligned} S(t) &= \text{Prob}(\tau > t \mid \tau > t-1)\,\text{Prob}(\tau > t-1) \\ &= (1 - \text{Prob}(\tau \leq t \mid \tau > t-1))\,\text{Prob}(\tau > t-1) \\ &= (1 - \text{Prob}(\tau = t \mid \tau \geq t))\,\text{Prob}(\tau > t-1) \\ &= q(t)S(t-1)\,, \end{aligned}$$

# Censoring, Kaplan-Meier and Reweighting

If we iterate $S(t) = q(t) \cdot S(t-1) = q(t) \cdot q(t-1) \cdot S(t-2) = \cdots$,

$$S(t) = \prod_{k=0}^{t} q(k) \text{ where } q(k) = 1 - \mathbb{P}[t = k | t \geq k]$$

$$\widehat{S}_{KM}(t) = \prod_{k=0}^{t} \widehat{q}(k) \text{ where } \widehat{q}(k) = 1 - \frac{d_k}{n_k} = 1 - \frac{\sum_{i=1}^{n} \mathbf{1}(y_i = k)}{\sum_{i=1}^{n} \mathbf{1}(y_i \geq k)},$$

$d(k)$ is the number of known deaths at time $k$ and $n(k)$ is the number of those persons who are alive (and not being censored) at time $k-1$.

Similarly, we can also derive the survival function for censoring times $K$, $\widehat{K}(t)$.

# Censoring, Kaplan-Meier and Reweighting

Recall that without censoring, $\widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(y_i \leq t)$. With censoring, Robins and Rotnitzky (1992) suggested

$$\widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{a_i \mathbf{1}(y_i \leq t)}{\widehat{K}(y_i^-)}$$

where $a_i = \mathbf{1}(t_i \leq c_i)$. Satten and Datta (2001) proved that $\widehat{F}(t) = 1 - \widehat{S}_{KM}(t)$.

# Calibration (when "probabilities" are badly assessed)

In many applications, we need to properly assess $\mathbb{P}(Y = 1 | \boldsymbol{X} = \boldsymbol{x})$

> model calibration can be also used to refer to Bayesian inference about the value of a model's parameters, given some data set, or more generally to any type of fitting of a statistical model. As Philip Dawid puts it, "*a forecaster 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.*" $\mathbb{W}$, see Dawid (1982).

Prediction $\widehat{Y}$ of $Y$ is a well-calibrated prediction if $\mathbb{E}_{\mathbb{P}}[Y|\widehat{Y} = p] = \widehat{y}$, for all $p \in (0, 1)$.

"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 (when "probabilities" are badly assessed)

"When we speak of the 'probability of death', the exact meaning of this expression can be defined in the following way only. We must not think of an individual, but of a certain class as a whole, e.g., 'all insured men forty-one years old living in a given country and not engaged in certain dangerous occupations'. A probability of death is attached to the class of men or to another class that can be defined in a similar way. We can say nothing about the probability of death of an individual even if we know his condition of life and health in detail. The phrase 'probability of death', when it refers to a single person, has no meaning for us at all," von Mises (1928, 1939).

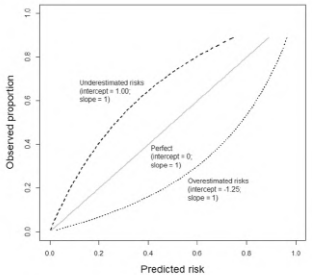
# Calibration (when "probabilities" are badly assessed)

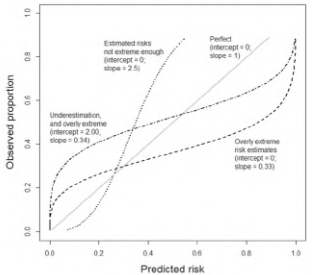
As explained in Van Calster et al. (2019), "among patients with an estimated risk of 20%, we expect 20 in 100 to have or to develop the event".

- 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 (when "probabilities" are badly assessed)

$$\text{BS} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{s}(\boldsymbol{x}_i) - y_i\right)^2$$

Calibration curve is defined as

$$\mathrm{g} : \begin{cases} [0,1] \to [0,1] \\ p \mapsto \mathrm{g}(p) := \mathbb{E}_{\mathbb{P}}[Y \mid \hat{s}(\boldsymbol{x}) = p] \end{cases}$$

- Quantile Bins

Set $\widehat{y}_i = \hat{s}(\boldsymbol{x}_i)$, sorted $\widehat{y}_1 \leq \widehat{y}_2 \leq \cdots \leq \widehat{y}_n$, partition $\mathcal{I}_1, \cdots, \mathcal{I}_{10}$ of $\{1, 2, \cdots, n\}$.
As in Pakdaman Naeini et al. (2015), consider scatter plot

$$(u, v_k), \text{ where } u_k = \frac{1}{n_k}\sum_{i \in \mathcal{I}_k} \widehat{y}_i \text{ and } v_k = \frac{1}{n_k}\sum_{i \in \mathcal{I}_k} y_i$$

# Calibration (when "probabilities" are badly assessed)

- Local Regression

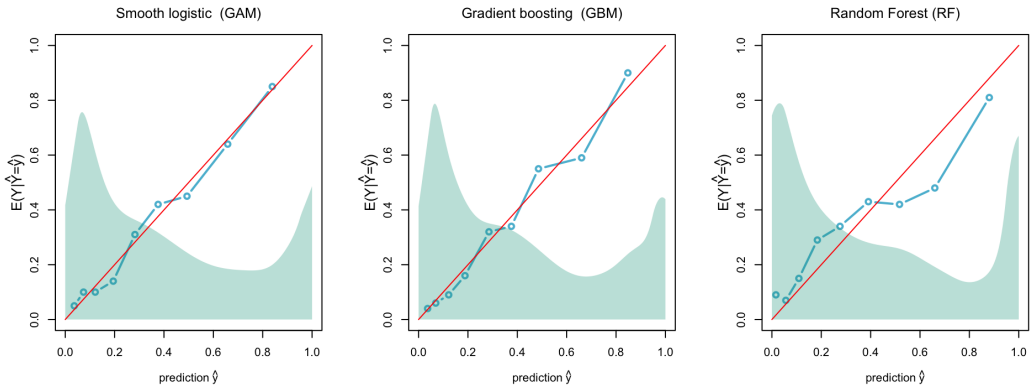
Local regression of $\{(\hat{s}(\boldsymbol{x}_i), y_i)\}$

$$\widehat{\mathrm{g}}_\alpha(p) = \frac{1}{n_{\mathcal{I}}} \sum_{i \in \mathcal{I}} y_i \text{ where } \mathcal{I} = \{i : |\hat{s}(\boldsymbol{x}_i) - p| \leq \alpha\}.$$

as suggested in Denuit et al. (2021)
One could also consider some kernel based local regression (of degree 1 or 2)

# Calibration (when "probabilities" are badly assessed)

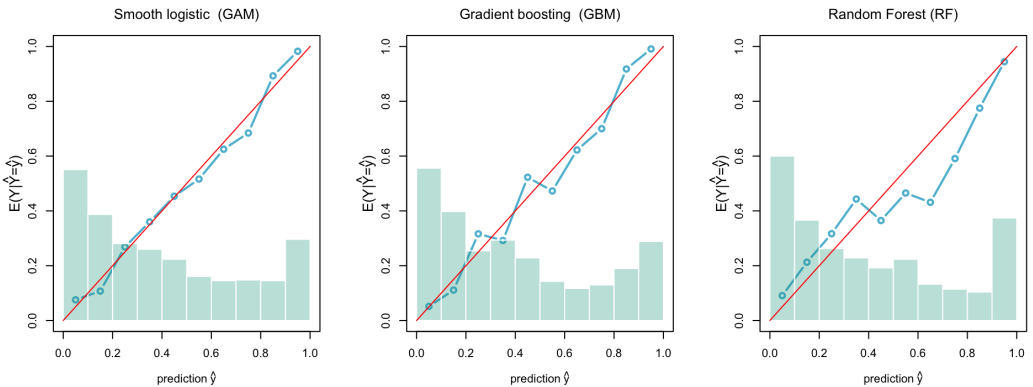
Calibration scatterplot per quantile bins




(see also Machado et al. (2024a,b))

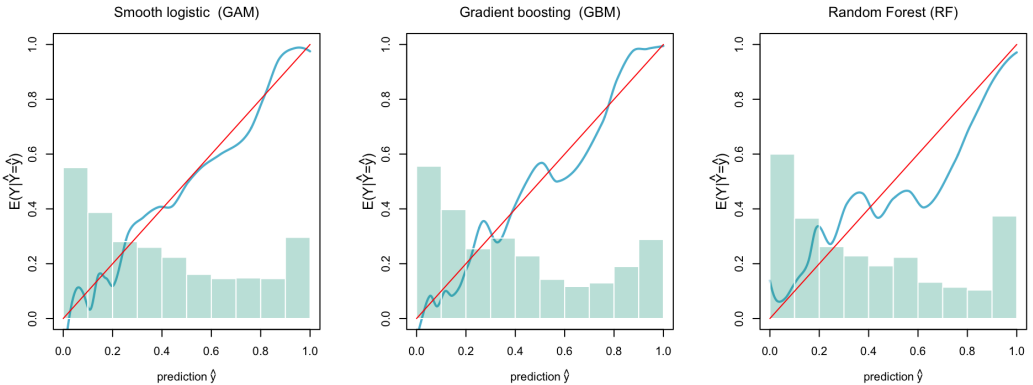
# Calibration (when "probabilities" are badly assessed)

Local regression scatterplot per bins, $[0; 0.1)$, $[0.1; 0.2)$, $[0.2; 0.3)$, $[0.3; 0.4)$, etc



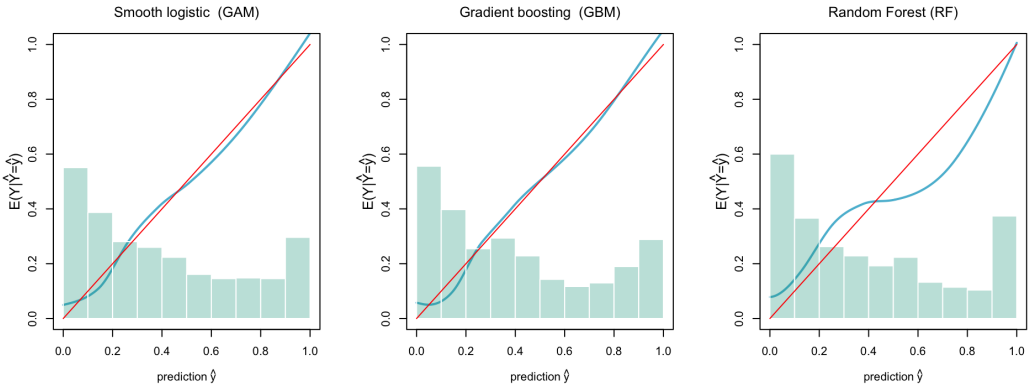
# Calibration (when "probabilities" are badly assessed)

Calibration scatterplot per local regression (small bandwidth)



# Calibration (when "probabilities" are badly assessed)

Local regression scatterplot per local regression (larger bandwidth)



# Selection Bias and Importance Sampling

Importance sampling is a classical technique for Monte Carlo simulations.
Monte Carlo is based on the law of large numbers: if we can draw i.i.d. copies of a random variable $X_i$'s, under probability $\mathbb{P}$, then

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

Much more can be obtained, since the empirical distribution $\mathbb{P}_n$ (associated with sample $\{x_1, \cdots, x_n\}$) converges to $\mathbb{P}$ as $n \to \infty$ (see, e.g., Van der Vaart (2000)).

Now, assume that we have an algorithm to draw efficiently i.i.d. copies of a random variable $X_i$'s, under probability $\mathbb{P}$, and we want to compute $\mathbb{E}_{\mathbb{Q}}[h(X)]$.

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

# Selection Bias and Importance Sampling

The term is on the left is

$$\widehat{\mu}^{IS} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}\mathbb{Q}(x_i)}{\mathrm{d}\mathbb{P}(x_i)} h(x_i)$$

and if the likelihood ratio is known only up to a multiplicative constant, define a "self-normalized importance sampling" estimate, as coined in Neddermeyer (2009) and Owen (2013),

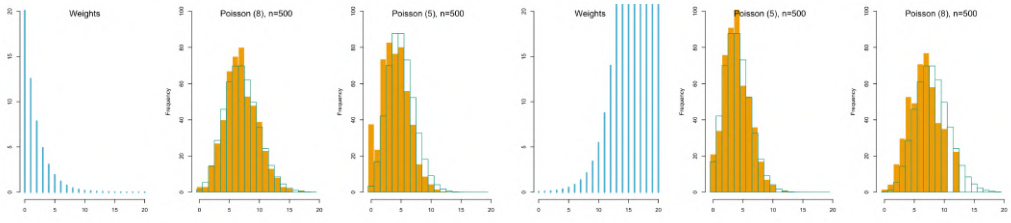
$$\widehat{\mu}^{IS\prime} = \frac{\sum_{i=1}^{n} \omega_i h(x_i)}{\sum_{i=1}^{n} \omega_i} \text{ with } \omega_i \propto \frac{\mathrm{d}\mathbb{Q}(x_i)}{\mathrm{d}\mathbb{P}(x_i)}.$$

$$\widehat{\mu}^{IS\prime} = \frac{\sum_{i=1}^{n} \omega_i h(x_i)}{\sum_{i=1}^{n} \omega_i} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \omega_i h(X_i)}_{\to \mathbb{E}_{\mathbb{P}}[X]} \underbrace{\frac{n}{\sum_{i=1}^{n} \omega_i h(x_i)}}_{\to 1} \to \mathbb{E}_{\mathbb{Q}}[X], \text{ as } n \to \infty.$$

# Selection Bias and Importance Sampling

**Law of large numbers:** If $X, X_1, X_2, \cdots, X_n, \cdots$ are i.i.d. samples of a random variable distributed according to $\mathbb{P}$, then for any (small) positive non-zero value $\epsilon > 0$, if $\mathbb{Q} \ll \mathbb{P}$:

$$\lim_{n\to\infty} \mathbb{Q}\left[\left|\mathbb{E}_{\mathbb{Q}}[X] - \frac{1}{n}\sum_{i=1}^{n} \omega_i X_i\right| > \epsilon\right] = 0, \text{ where } \omega_i \propto \frac{\mathrm{d}\mathbb{Q}(x_i)}{\mathrm{d}\mathbb{P}(x_i)}$$



# Selection Bias and Importance Sampling

(1) suppose we can generate Poisson distribution $\mathcal{P}(8)$, we want some Poisson $\mathcal{P}(5)$,
(2) suppose we can generate Poisson distribution $\mathcal{P}(5)$, we want some Poisson $\mathcal{P}(8)$.

In our context, one can define the "importance sampling estimator" of $\mathbb{E}_{\mathbb{P}}[Y(1)]$, as

$$\widehat{\mu}^{IS}(Y(1)) = \frac{1}{n_1}\sum_{t_i=1}\frac{y_i}{e(\boldsymbol{x}_i)}\frac{n_t}{n} = \frac{1}{n}\sum_{t_i=1}\frac{y_i}{e(\boldsymbol{x}_i)},$$

and a "self-normalized importance sampling" estimate for $\mathbb{E}_{\mathbb{P}}[Y(1)]$ is

$$\widehat{\mu}^{IS\prime}(Y(1)) = \frac{\sum_{t_i=1}\omega_i y_i}{\sum_{t_i=1}\omega_i}, \text{ where } \omega_i = \frac{1}{e(\boldsymbol{x}_i)}.$$

The "self-normalized importance sampling" estimate for $\tau$

$$\widehat{\tau}^{IS\prime} = \frac{\sum_{t_i=1}\omega_i y_i}{\sum_{t_i=1}\omega_i} - \frac{\sum_{t_i=0}\omega_i' y_i}{\sum_{t_i=0}\omega_i'}-, \text{ where } \omega_i = \frac{1}{e(\boldsymbol{x}_i)} \text{ and } \omega_i' = \frac{1}{1-e(\boldsymbol{x}_i)}.$$

# Distorting Scores to Mitigate Unfairness

When dealing with fairness and discrimination, Charpentier (2024), we want to insure, if $\widehat{y} = m(\boldsymbol{x}, s)$, either

- that model $m$ satisfies the **independence property** if $m(\boldsymbol{X}, S) \perp\!\!\!\perp S$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$ — demographic parity
- that model satisfies the **separation property** if $m(\boldsymbol{X}, S) \perp\!\!\!\perp S \mid Y$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$ — equalized odds
- that model satisfies the **sufficiency property** if $Y \perp\!\!\!\perp S \mid m(\boldsymbol{X}, S)$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$ — calibration

For demographic parity, maybe $\widehat{Y} \not\perp\!\!\!\perp S$ under $\mathbb{P}$, but $\widehat{Y} \perp\!\!\!\perp S$ under $\mathbb{Q}$.
Mitigation means distorting scores

# Distorting Scores to Mitigate Unfairness

When dealing with fairness and discrimination, Charpentier (2024), we want to insure, e.g., **demographic parity**, $\widehat{Y} \perp\!\!\!\perp S$ (where $S$ is a categorical sensitive attribute),

$$\begin{cases} \text{weak version : } \mathbb{E}_{\mathbb{P}}[\widehat{Y}|S=A] = \mathbb{E}_{\mathbb{P}}[\widehat{Y}|S=B] \\ \text{strong version : } (\widehat{Y}|S=A) \overset{\mathcal{L}}{=} (\widehat{Y}|S=B) \end{cases}$$

We need a "distance" between two probability measures, $\mathbb{P}_A$ and $\mathbb{P}_B$, e.g., Wasserstein,

$$\mathcal{W}_2^2(\mathbb{P}_A, \mathbb{P}_B) = \min_{T:[0,1]\to[0,1]} \mathbb{E}\big[\big(\overbrace{T(X)}^{Y\sim\mathbb{P}_B} - \underbrace{X}_{X\sim\mathbb{P}_A}\big)^2\big] = \mathbb{E}\left[\big(\overbrace{F_B^{-1}\circ F_A(X)}^{Y\sim\mathbb{P}_B} - \underbrace{X}_{X\sim\mathbb{P}_A}\big)^2\right]$$

# Distorting Scores to Mitigate Unfairness

optimal transport mapping

quantile of level $p$ in group B

$$T^\star(x) = F_B^{-1} \circ F_A(x)$$

probability $p$ associated with $x$ in group A

$T^\star$ is a monotonic (nondecreasing) mapping.

$$T^\star = \underset{T:[0,1]\to[0,1]}{\text{argmin}} \int_0^1 \big(T(x) - x\big)^2 dF_A(x)$$

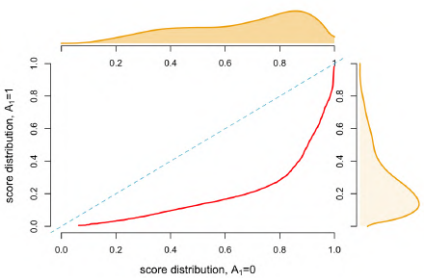
i.e. $\underset{T:[0,1]\to[0,1]}{\text{argmin}} \mathbb{E}\big[\big(T(X) - X\big)^2\big]$ where $X \sim F_A$,

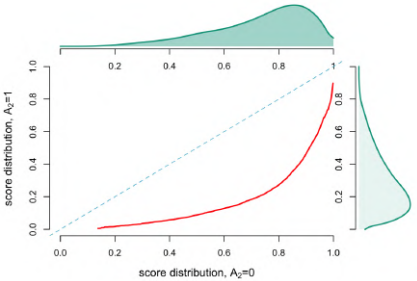
$Y$ with $Y \sim F_B$

# Distorting Scores

$$\mathbb{E}\left[\left(T^\star(X)-X\right)^2\right] = \min_{T:[0,1]\to[0,1]} \mathbb{E}\left[\left(T(X)-X\right)^2\right]$$

$$\mathcal{W}_2(\mathbb{P}_A,\mathbb{P}_B)^2 = \int_{\mathcal{X}} |F_B^{-1}\circ F_A(x)-x|^2 \mathrm{d}\mathbb{P}_A(x) = \int_0^1 |F_B^{-1}(u)-F_A^{-1}(u)|^2 \mathrm{d}u$$

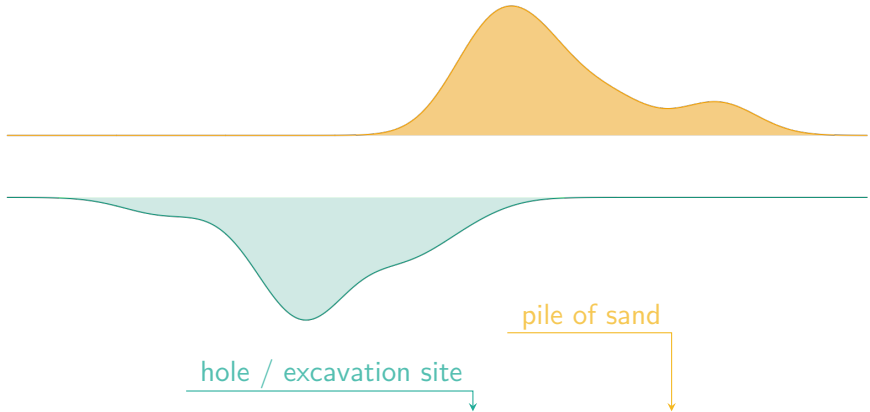



Mapping $T$ is associated to a "push-forward" operator
$\mathbb{P}_B(\mathcal{S}) = T_\#\mathbb{P}_A(\mathcal{S}) = \mathbb{P}_0(T^{-1}(\mathcal{S})),\ \forall \mathcal{S}\subset\mathbb{R}.$

# Distorting Scores

Optimal transport distance is interesting because it provides a constructive mapping, point by point (as in the original Monge (1781)'s problem),




Monge (1781), "Mémoire sur la théorie des déblais et des remblais"

# Distorting Scores

Optimal transport distance is interesting because it provides a constructive mapping, point by point (as in the original Monge (1781)'s problem),




Monge (1781), "Mémoire sur la théorie des déblais et des remblais"

# Distorting Scores

One can also consider **barycenters of measures**,

$$\mathbb{Q}^\star = \underset{\mathbb{Q}}{\operatorname{argmin}} \sum_j \omega_j \cdot \mathcal{W}_2(\mathbb{Q}, \mathbb{P}_j)^2$$

as in Agueh and Carlier (2011). This optimal distribution $\mathbb{Q}^\star$ is the distribution of $m^\star(\boldsymbol{X}, S)$

$$\begin{cases} m^\star(\boldsymbol{x}, s = A) = \mathbb{P}[S = A] \cdot m(\boldsymbol{x}, s = A) \\ \qquad\qquad + \mathbb{P}[S = B] \cdot F_B^{-1} \circ F_A(m(\boldsymbol{x}, s = A)) \\ m^\star(\boldsymbol{x}, s = B) = \mathbb{P}[S = A] \cdot F_A^{-1} \circ F_B(m(\boldsymbol{x}, s = B)) \\ \qquad\qquad + \mathbb{P}[S = B] \cdot m(\boldsymbol{x}, s = B). \end{cases}$$

score in group A

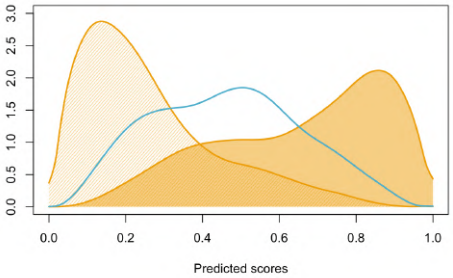
$p = F_A(m(\boldsymbol{x}, s = A))$

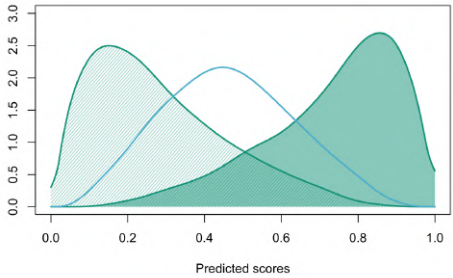
$$m^\star(\boldsymbol{x}, s = A) = \mathbb{P}[S = A] \cdot m(\boldsymbol{x}, s = A) + \mathbb{P}[S = B] \cdot F_B^{-1} \circ F_A(m(\boldsymbol{x}, s = A))$$

weights

# Distorting Scores

For example, on distributions of scores, conditional on $S_1$ and $S_2$,




See Charpentier (2024) for additional properties...

# On Counterfactual 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*'," Kusner et al. (2017)

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

$$\text{ATE} = \mathbb{E}\big[Y(A) - Y(B)\big] = 0.$$

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

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

# A little bit of geometry ?

Optimal transport exists in any dimension (but no intuitive results based on quantiles)

> In geometry, a geodesic is a curve representing in some sense the shortest path (arc) between two points in a surface, or more generally in a Riemannian manifold (...) It is a generalization of the notion of a "straight line". W

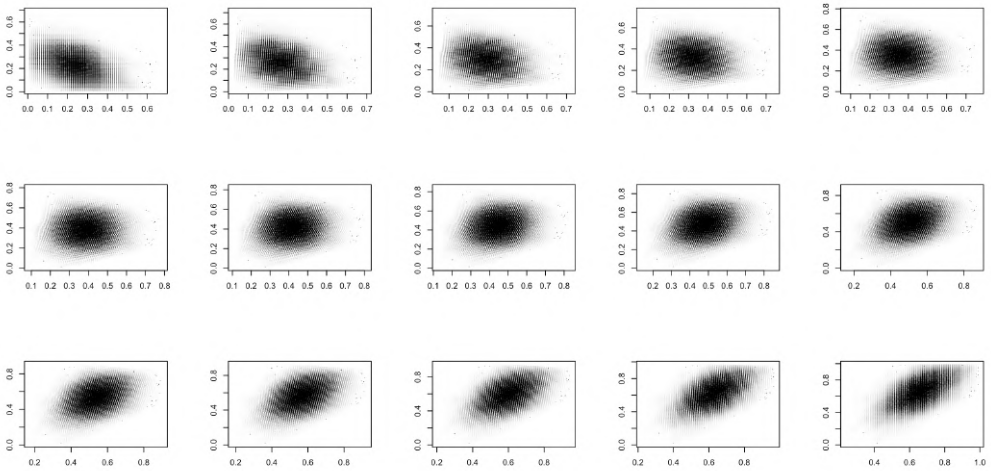
Formally, consider some metric space, $(E, d)$. A constant speed geodesic between two points $x_0, x_1 \in E$ is a continuous curve $x : [0, 1] \to E$ such that for every $s, t \in (0, 1)$, $d(x_s, x_t) = |s - t| d(x_0, x_1)$.

Given two measures $\mathbb{P}_0$ and $\mathbb{P}_1$, and an optimal transport map $T^\star$ such that $\mathbb{P}_1 = T^\star_\# \mathbb{P}_0$, then

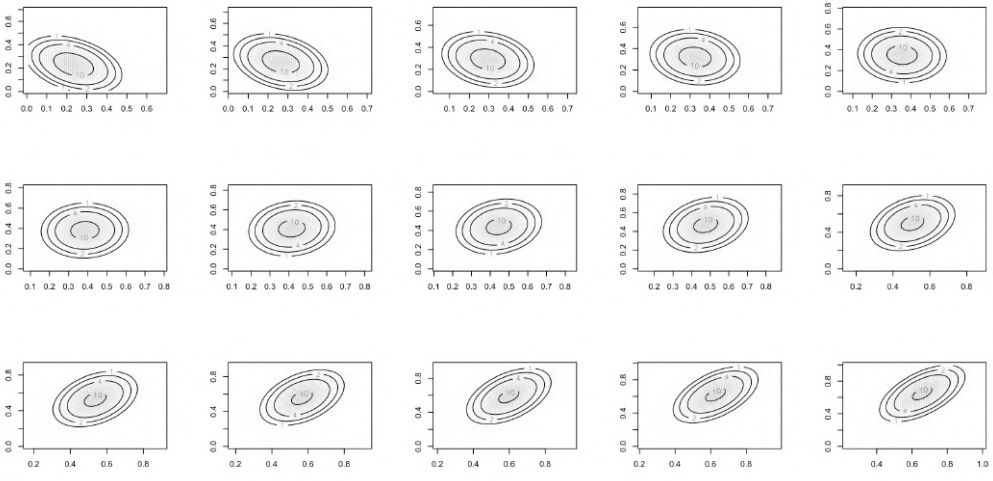
$$\mathbb{P}_t = ((1 - t)\mathbb{I} + tT^\star)_\# \mathbb{P}_0, \text{ for any } t \in (0, 1).$$

(that's what we used with barycenters)

# A little bit of geometry ?



# A little bit of geometry ?



# "Distances" between distributions (or measures)

Integral probability metrics (IPMs, Müller (1997)) are distances on the space of distributions over a set $\mathcal{X}$, defined by a class $\mathcal{F}$ of real-valued functions on $\mathcal{X}$ as

$$D_{\mathcal{F}}(p,q) = \sup_{f\in\mathcal{F}} \big|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]\big|.$$

$X \sim p$ $\qquad$ $Y \sim q$

Discussed also in Dedecker and Merlevède (2007)

For two distributions $p$ and $q$, the total variation distance (Jordan (1881); Rudin (1966)) between $p$ and $q$ is

$$d_{\mathrm{TV}}(p,q) = \sup_{\mathcal{A}} \{|p(\mathcal{A}) - q(\mathcal{A})|\}.$$

Equivalently,

$$d_{\mathrm{TV}}(p,q) = \frac{1}{2} \sup_{f:\mathbb{R}^k\to\{0,1\}} \left\{ \int f \mathrm{d}p - \int f \mathrm{d}q \right\}$$

# "Distances" between distributions (or measures)

(see e.g. https://djalil.chafai.net/blog/, with $f : \mathbb{R}^k \to \{-1, 1\}$, $f = \mathbf{1}_{\mathcal{A}} - \mathbf{1}_{\mathcal{A}^c}$)
Thus, it is an IPM with $\mathcal{F} = \{f : \mathcal{X} \to \{0, 1\}\}$, so that $\mathcal{F}$ is a set of indicator functions for any event.
For two distributions $p$ and $q$, Kolmorov-Smirnov distance (Kolmogorov (1933); Smirnov (1948)) between $p$ and $q$ is

$$d_{\mathrm{KS}}(p, q) = \sup_{t \in \mathbb{R}} \{|p((-\infty, t]) - q((-\infty, t])|\} = \sup_{t \in \mathbb{R}} \{|F_p(t) - F_q(t)|\} = \|F_p - F_q\|_\infty,$$

where $F_p$ and $F_q$ are the respective cumulative distribution functions.

# "Distances" between distributions (or measures)

For two discrete distributions $p$ and $q$, Kullback–Leibler divergence (Kullback and Leibler (1951) ) of $p$, with respect to $q$ is

$$D_{\mathrm{KL}}(p\|q) = \sum_i p(i) \log \frac{p(i)}{q(i)},$$

and for absolutely continuous distributions,

$$D_{\mathrm{KL}}(p\|q) = \int_{\mathbb{R}} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x \text{ or } \int_{\mathbb{R}^k} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \, \mathrm{d}\boldsymbol{x},$$

in higher dimension.

# "Distances" between distributions (or measures)

Consider two measures on $p$ and $q$ on $\mathbb{R}$. Then define Cramér distance (Cramér (1928a,b) and Székely (2003))

$$C_k(p,q) = \Big( \int_{-\infty}^{\infty} |F_p(x) - F_q(x)|^k \mathrm{d}x \Big)^{1/k}, \text{ for } k \geq 1$$

$C_2$ is named "energy-distance" in Székely (2003) and Rizzo and Székely (2016), and "continuous ranked probability score" in Gneiting et al. (2007).
It is an Integral Probability Metrics (IPM), since

$$C_k(p,q) = \sup_{f \in \mathcal{F}_{k'}} \big| \mathbb{E}[f(X)] - \mathbb{E}[f(Y)] \big|.$$

$k^{-1} + k'^{-1} = 1$ ; $X \sim p$ ; $Y \sim q$

where $\mathcal{F}_{k'}$ is the set of absolutely continuous functions such that $\|\nabla f\|_{k'} \leq 1$.
For example, if $k = 1$, $\|\nabla f\|_\infty \leq 1$ (corresponding to 1-Lipschitz functions).

# "Distances" between distributions (or measures)

Consider two measures on $p$ and $q$ on $\mathbb{R}$. Then define Wasserstein distance (Wasserstein (1969))

$$\mathcal{W}_k(p,q) = \Big( \int_0^1 |F_p^{-1}(u) - F_q^{-1}(u)|^k \mathrm{d}u \Big)^{1/k}, \text{ for } k \geq 1$$

Consider two measures on $p$ and $q$ on $\mathbb{R}$.

$$\mathcal{W}_2(p,q)^2 = \int_0^1 |F_p^{-1}(u) - F_q^{-1}(u)|^2 \mathrm{d}u \text{ while } C_2(p,q) = \int_{-\infty}^{\infty} |F_p(x) - F_q(x)|^2 \mathrm{d}x.$$

Consider two Gaussian distributions, then

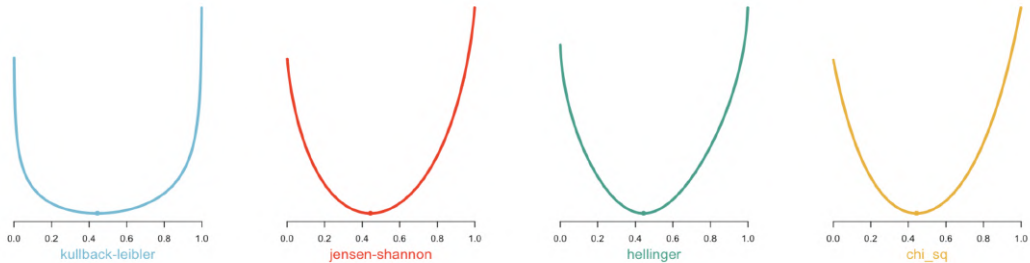
$$\mathcal{W}_2(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \boldsymbol{\sigma}_2^2))^2 = (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2,$$

and for two Bernoulli distributions, if $p_1 \leq p_2$, $\mathcal{W}_k(\mathcal{B}(p_1), \mathcal{B}(p_2)) = (p_2 - p_1)^{1/k}$,

$$\mathcal{W}_2(\mathcal{B}(p_1), \mathcal{B}(p_2)) = \sqrt{p_2 - p_1} \text{ and } \mathcal{W}_1(\mathcal{B}(p_1), \mathcal{B}(p_2)) = p_2 - p_1.$$

# "Distances" between distributions (or measures)

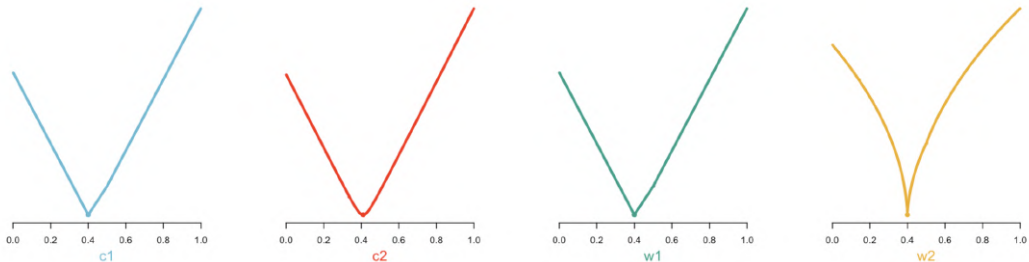
- $\mu$: multinomial distribution on $\{0, 1, 10\}$, with $\boldsymbol{p} = (.5, .1, .4)$
- $\nu_\theta$: binomial type distribution on $\{0, 10\}$, with $\boldsymbol{q}_\theta = (1 - \theta, \theta)$
- Let $\theta^\star = \text{argmin}\{d(p, q_\theta)\}$ or $\theta^\star = \text{argmin}\{d(p\|q_\theta)\}$




- with $d_{\mathrm{KL}}(p\|q_\theta)$, $d_{\mathrm{JS}}(p, q_\theta)$, $d_{\mathrm{H}}(p, q_\theta)$ and $d_{\mathrm{H}_{\chi^2}}(p\|q_\theta)$

# "Distances" between distributions (or measures)

- $\mu$: multinomial distribution on $\{0, 1, 10\}$, with $\boldsymbol{p} = (.5, .1, .4)$
- $\nu_\theta$: binomial type distribution on $\{0, 10\}$, with $\boldsymbol{q}_\theta = (1 - \theta, \theta)$
- Let $\theta^\star = \text{argmin}\{d(p, q_\theta)\}$




- with $C_1(p, q_\theta)$, $C_2(p, q_\theta)$, $\mathcal{W}_1(p, q_\theta)$ and $\mathcal{W}_2(p, q_\theta)$.

# "Distances" between distributions (or measures)

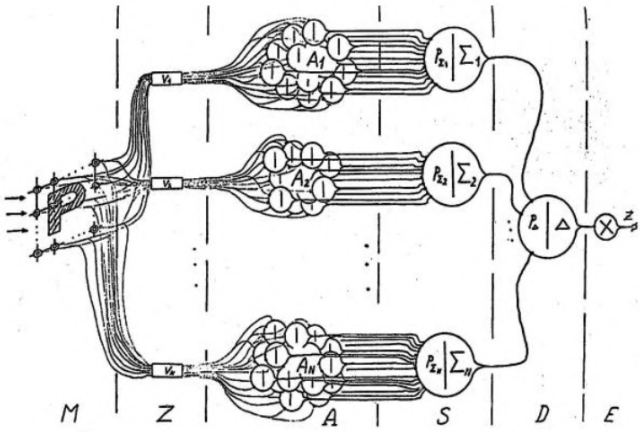
$\mathcal{W}_1$ is an IPM where $\mathcal{F}$ the set of 1-Lipschitz functions, Kantorovich and Rubinstein (1958), i.e., if $p$ and $q$ have bounded support,

$$\mathcal{W}_1(p, q) = \sup_{f \in \mathcal{F}} \left\{ \int_{-\infty}^{+\infty} f(x) \, \mathrm{d}(p - q)(x) \right\},$$

$\mathcal{F}$ being the class of 1-Lipschitz functions

Gretton et al. (2012) introduced Maximum Mean Discrepancy (MMD) as a distance between two measures (using RKHS representations).

# Transfer learning in Machine Learning Literature



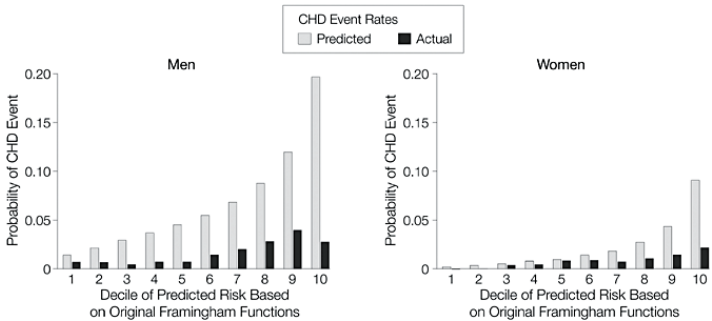

Source: Bozinovski and Fulgosi (1976), The influence of pattern similarity and transfer learning

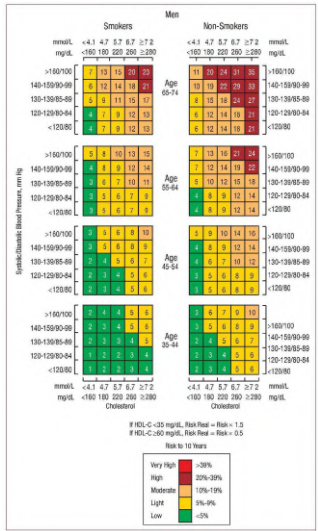
# Transfer learning in Machine Learning Literature

- Framingham coronary heart disease (CHD) risk score, Wilson et al. (1987, 1998); D'Agostino et al. (2001)

6 risk factors: age, BP, smoking, diabetes, total cholesterol (TC), and high-density lipoprotein cholesterol (HDL-C)

Framingham (U.S.) participants are of European descent
what if we use it on Chinese people ?, Liu et al. (2004)

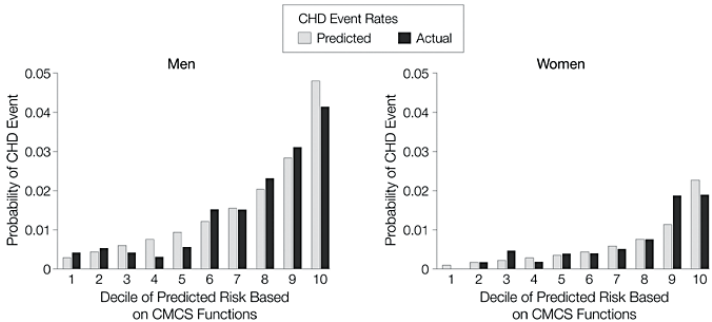


# Transfer learning in Machine Learning Literature

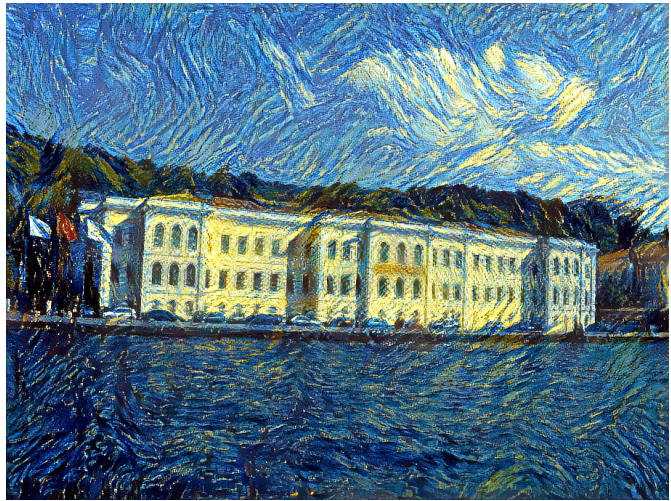
- Framingham coronary heart disease (CHD) risk score, Liu et al. (2004)

Refitted on Chinese population,
Chinese Multi-provincial Cohort Study (CMCS)



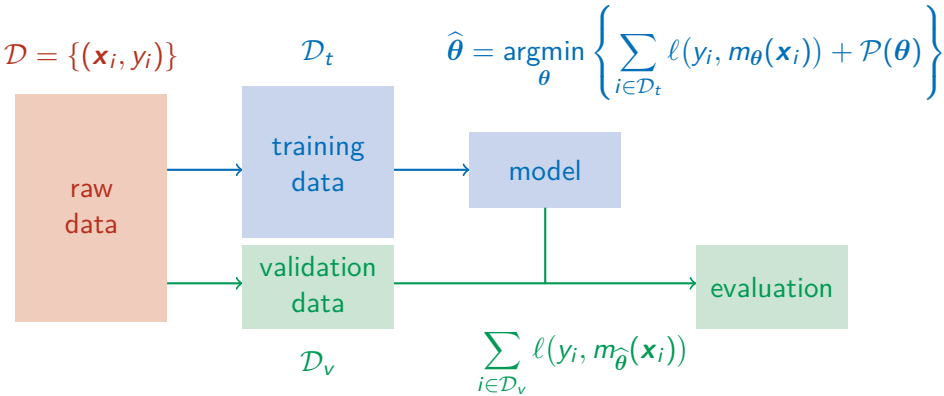

| Risk Factors | CMCS β | Framingham β |
|---|---|---|
| Age | 0.07 | 0.05 |
| Age squared | NA | NA |
| Blood pressure | | |
| Optimal | −0.51 | 0.09 |
| Normal | | |
| High normal | 0.21 | 0.42 |
| Stage 1 hypertension | 0.33 | 0.66 |
| Stage 2-4 hypertension | 0.77 | 0.90 |
| TC, mg/dL | | |
| <160 | −0.51 | −0.38 |
| 160-199 | | |
| 200-239 | 0.07 | 0.57 |
| 240-279 | 0.32 | 0.74 |
| ≥280 | 0.52 | 0.83 |
| HDL-C, mg/dL | | |
| <35 | −0.25 | 0.61 |
| 35-44 | 0.01 | 0.37 |
| 45-49 | | |
| 50-59 | −0.07 | 0.00 |
| ≥60 | −0.40 | −0.46 |
| Diabetes | 0.09 | 0.53 |
| Smoking | 0.62 | 0.73 |

# Transfer learning in Machine Learning Literature



Source: https://sefiks.com/2018/07/20/artistic-style-transfer-with-deep-learning/, "learning style"

# Quick Overview of "Transfer Learning"




where formally, $\mathcal{D}$ is a a collection of i.i.d. observations from $(\boldsymbol{X}_i, Y_i) \sim \mathbb{P}$ (or $\mathcal{D} \sim \mathbb{P}$) and we suppose that both $\mathcal{D}_t \sim \mathbb{P}$ and $\mathcal{D}_v \sim \mathbb{P}$ (concept of "generalization")

# Quick Overview of "Transfer Learning"

More precisely, let us distinguish

- the feature domain, $\mathcal{D}_x = (\mathcal{X}, \mathbb{P})$, with $\mathcal{X} \subset \mathbb{R}^k$ and $\mathbb{P}$ is a distribution on $\mathcal{X}$, we have a **source** (or training) and a **target** (or validation), and $\mathcal{D}_s$ and $\mathcal{D}_t$
- the task, $(\mathcal{Y}, m(\cdot))$, with $\mathcal{Y} \subset \mathbb{R}$ is a label space and $m$ is our predictive model, $m : \mathcal{X} \to \mathcal{Y}$

Transfer learning is about improving the target predictive model $m_t(\cdot)$ by using $\mathcal{D}_s$ and $m_s(\cdot)$, even if we know that $\mathcal{D}_s \neq \mathcal{D}_t$ and $m_s(\cdot) \neq m_t(\cdot)$.

- $\mathcal{X}_s \neq \mathcal{X}_t$: heterogeneous transfer learning (different language)
- $\mathbb{P}_s \neq \mathbb{P}_t$: domain adaptation (different topic)
- $\mathcal{Y}_s \neq \mathcal{Y}_t$: possibly, $\{0, 1\}$ vs. $\{0, N, 1\}$
- $\mathbb{P}_s(y|\boldsymbol{x}) \neq \mathbb{P}_t(y|\boldsymbol{x})$ (and therefore $m_s(\cdot) \neq m_t(\cdot)$)

# Transfer learning and domain adaptation ($\mathbb{P}_s \neq \mathbb{P}_t$)

"Traditional machine learning is characterized by training data and testing data having the same input feature space and the same data distribution. When there is a difference in data distribution between the training data and test data, the results of a predictive learner can be degraded," Furht et al. (2016)

- **notations**

Consider some training (source) sample $\mathcal{D}_s = \{(\boldsymbol{x}_{s,i}, y_{s,i})\}$ and some test (target) sample $\mathcal{D}_t = \{(\boldsymbol{x}_{t,i})\}$, both being i.i.d., with distributions $\mathbb{P}_s$ and $\mathbb{P}_t$.
In a regression problem, $y = m(\boldsymbol{x}) + \varepsilon$, i.e. $m(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}}[Y|\boldsymbol{E} = \boldsymbol{x}]$
Consider a parametric model, $m(\boldsymbol{x}|\boldsymbol{\theta})$), for some $\boldsymbol{\theta} \in \Theta$.
Classical empirical risk minimization (ERM) leads to

$$\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_{s,i}, m(\boldsymbol{x}_{s,i}|\boldsymbol{\theta})) \right\}$$

# Transfer learning and domain adaptation ($\mathbb{P}_s \neq \mathbb{P}_t$)

If $\mathbb{P}_s = \mathbb{P}_t$, $\widehat{\boldsymbol{\theta}}$ is said to be consistent Shimodaira (2000). Otherwise...
Importance weighted empirical risk minimization (IWERM) is

$$\tilde{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{P}_s(\boldsymbol{x}_{s,i})}{\mathbb{P}_t(\boldsymbol{x}_{s,i})} \right.$$

$\ell(y_{s,i}, m(\boldsymbol{x}_{s,i}|\boldsymbol{\theta}))) which is now consistent.$
One can define adaptative importance weighted empirical risk minimization (AIWERM)

$$\tilde{\boldsymbol{\theta}}_\gamma \in \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left( \frac{\mathbb{P}_s(\boldsymbol{x}_{s,i})}{\mathbb{P}_t(\boldsymbol{x}_{s,i})} \right)^{\gamma} \ell(y_{s,i}, m(\boldsymbol{x}_{s,i}|\boldsymbol{\theta}))) \right\},$$

$\gamma \in [0, 1]$ is the flattening parameter,

$$\begin{cases} \gamma = 0, \text{ ordinary ERM} \\ \gamma = 1, \text{ IWERM} \end{cases}$$

# Transfer learning and domain adaptation ($\mathbb{P}_s \neq \mathbb{P}_t$)

One could consider regularlized importance weighted empirical risk minimization (RIWERM)

$$\tilde{\boldsymbol{\theta}}_\lambda \in \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{P}_s(\boldsymbol{x}_{s,i})}{\mathbb{P}_t(\boldsymbol{x}_{s,i})} \ell(y_{s,i}, m(\boldsymbol{x}_{s,i}|\boldsymbol{\theta})) + \lambda\mathcal{P}(\boldsymbol{\theta}) \right\},$$

for some penalty function $\mathcal{P}(\boldsymbol{\theta})$ (classically $\|\boldsymbol{\theta}\|_{\ell_1}$ (lasso) or $\|\boldsymbol{\theta}\|_{\ell_2}$ (ridge) types of penalty), and $\lambda \geq 0$.

# Transfer learning and domain adaptation ($\mathbb{P}_s \neq \mathbb{P}_t$)

- Application in a regression context

Polynomial regression model,

$$\mathbb{P}_{x,\boldsymbol{\theta}} \sim \mathcal{N}(P_\beta(x), \sigma^2) \text{ and } \boldsymbol{\theta} = (\beta, \sigma^2), \text{ for some polynomial } P_\beta$$

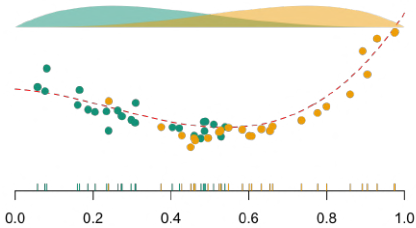
i.e., $y = \beta_0 + \beta_1 x + \cdots + \beta_k x^k + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Suppose that the "true" distribution is

$$\mathbb{Q}_x \sim \mathcal{N}(Q(x), 1)$$

e.g., $Q(x) = -(2x - 1/2) + (2x - 1/2)^3$
Suppose also that

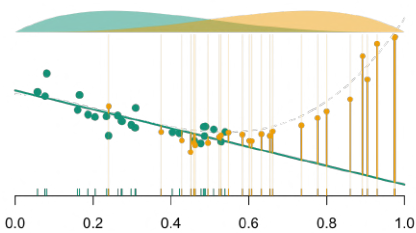
$$\begin{cases} \text{source}: & \pi_s \sim \mathcal{B}(a_s, b_s) \\ \text{target}: & \pi_t \sim \mathcal{B}(a_t, b_t) \end{cases}$$

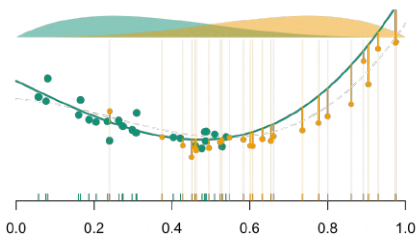


# Transfer learning and domain adaptation ($\mathbb{P}_s \neq \mathbb{P}_t$)

Linear model (mis-specified) and cubic model (well-specified)

$$\max_{\boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x}) = \max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y_i - P_{\beta(x_i)})^2$$

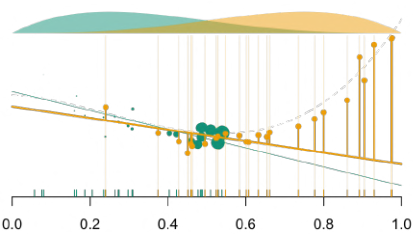


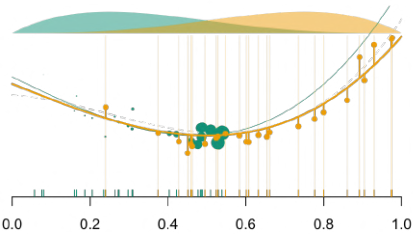



# Transfer learning and domain adaptation ($\mathbb{P}_s \neq \mathbb{P}_t$)

Linear model (mis-specified) and cubic model (well-specified)

$$\max_{\boldsymbol{\theta}} \log \mathcal{L}_\omega(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{x}) = \max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \omega(x_i) \log p(y|\boldsymbol{x},\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \frac{x_i^{a_t}(1-x_i)^{b_t}}{x_i^{a_s}(1-x_i)^{s_t}} (y_i - P_{\beta(x_i)})^2$$



# Transfer learning and domain adaptation ($\mathbb{P}_s \neq \mathbb{P}_t$)

Another example would be Pielke and Landsea (1998); Chavas et al. (2013); Weinkle et al. (2018); Martinez (2020),

"Both population and wealth have increased dramatically over the last several decades and act to enhance the recent hurricane damages preferentially over those occurring previously. More appropriate trends in the United States hurricane damages can be calculated when a normalization of the damages are done to take into account inflation and changes in coastal population and wealth," Pielke and Landsea (1998)

# Transfer learning and domain adaptation ($\mathbb{P}_s \neq \mathbb{P}_t$)

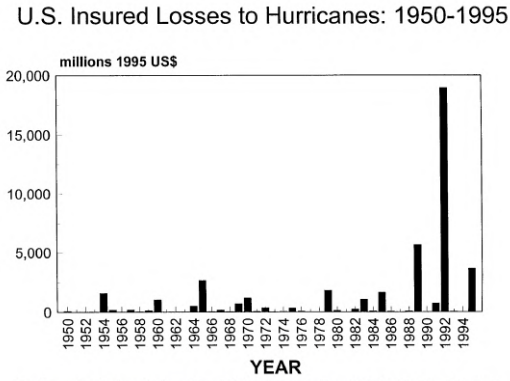



FIG. 2. Time series of hurricane-related insured losses in the United States (in millions of 1995 U.S. dollars) from 1950 to 1995 (data provided courtesy of Property Claims Services, Inc.).

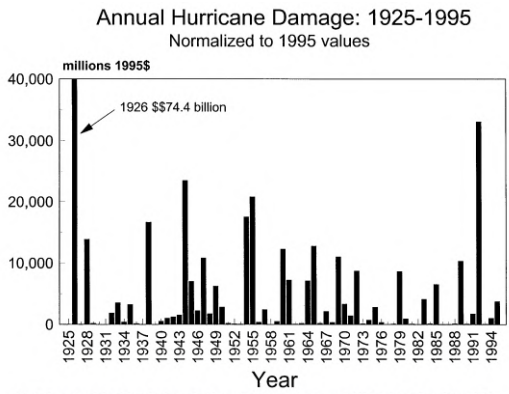



FIG. 4. Time series of United States hurricane-related losses (direct damages in millions of 1995 U.S. dollars) from 1925 to 1995 in normalized 1995 damage amounts (utilizing inflation, coastal county population changes, and changes in wealth).

(Source: Pielke and Landsea (1998))

# Transfer learning and heterogeneity ($\mathcal{X}_s \neq \mathcal{X}_t$)

Actuarial science is usually based on tabular data

Observation is $(\boldsymbol{x}_i, y_i)$

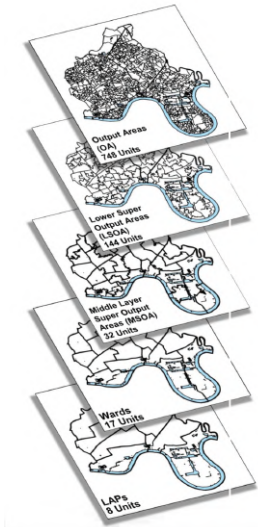
In many applications, use of "demographic information"

> An ecological fallacy is a formal fallacy in the interpretation of statistical data that occurs when inferences about the nature of individuals are deduced from inferences about the group to which those individuals belong W

As in Goodman (1953, 1959).

See also Holt et al. (1996), Sedgwick (2015)

(Source: https://www.towerhamlets.gov.uk/Documents/Borough_stat...)



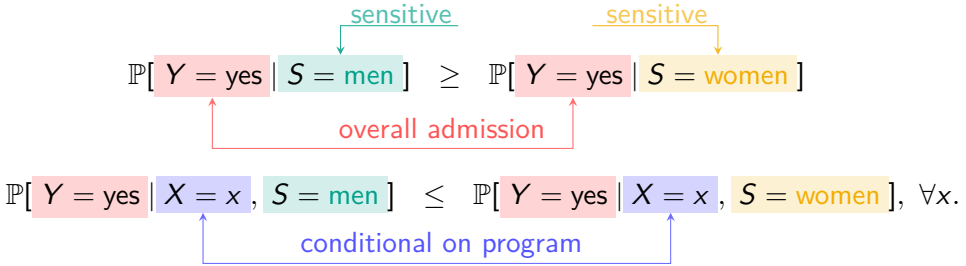
# Transfer learning and heterogeneity ($\mathcal{X}_s \neq \mathcal{X}_t$)

> Simpson's paradox is a phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined W

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

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

# Transfer learning and heterogeneity ($\mathcal{X}_s \neq \mathcal{X}_t$)




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

# Transfer learning and heterogeneity ($\mathcal{X}_s \neq \mathcal{X}_t$)

Consider the following mortality rates in two hospitals (fake data)

| | Total | Healthy | Pre-condition | Proportions |
|---|---|---|---|---|
| Hospital A | $800/1000 = 80\%$ | $590/600 \sim \mathbf{98}\%$ | $210/400 \sim \mathbf{53}\%$ | 60%-40% |
| Hospital B | $900/1000 = \mathbf{90}\%$ | $870/900 \sim 97\%$ | $30/100 \sim 30\%$ | 90%-10% |

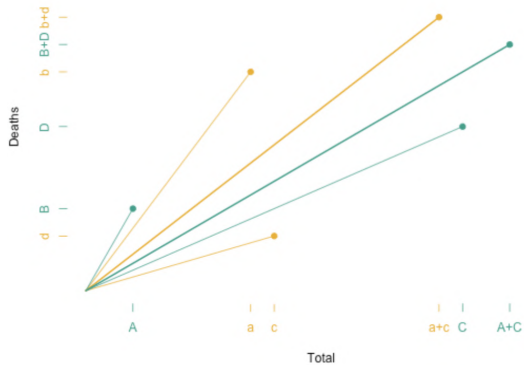
There is no mathematical "paradox", *per se*.

We could have

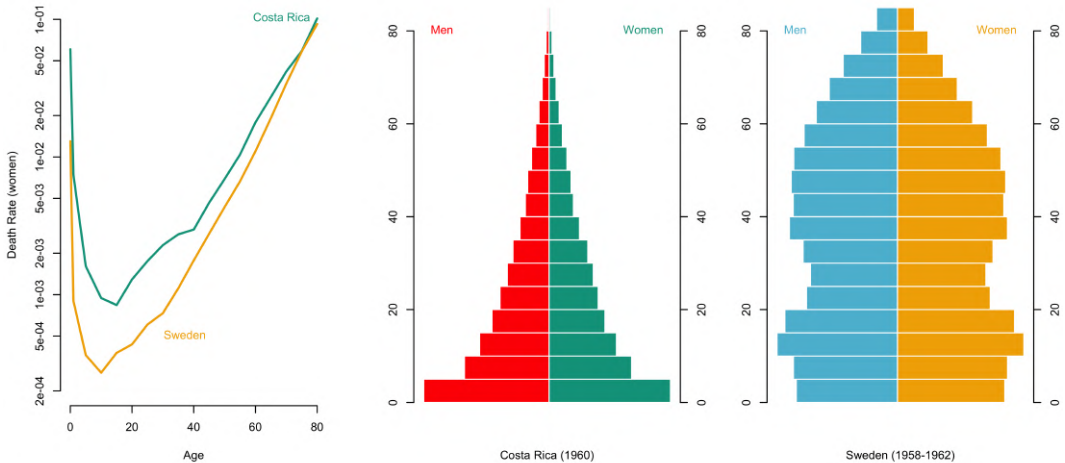
$$\frac{A}{B} \geq \frac{a}{b} \text{ and } \frac{C}{D} \geq \frac{c}{d}$$

and at the same time

$$\frac{A+C}{B+D} \leq \frac{a+c}{b+d}$$



# Transfer learning and heterogeneity ($\mathcal{X}_s \neq \mathcal{X}_t$)




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

# Transfer learning and heterogeneity ($\mathcal{X}_s \neq \mathcal{X}_t$)

September 27, 2023, the Colorado Division of Insurance exposed a new proposed regulation entitled Concerning Quantitative Testing of External Consumer Data and Information Sources, Algorithms, and Predictive Models Used for Life Insurance Underwriting for Unfairly Discriminatory Outcomes

– Section 5 (Estimating Race and Ethnicity) –

Insurers shall estimate the race or ethnicity of all proposed insureds that have applied for coverage on or after the insurer's initial adoption of the use of ECDIS, or algorithms and predictive models that use ECDIS, including a third party acting on behalf of the insurer that used ECDIS, or algorithms and predictive models that used ECDIS, in the underwriting decision-making process, by utilizing: BIFSG and the insureds' or proposed insureds' name and geolocation (...)

Bayesian Improved First Name Surname Geocoding, or "BIFSG"
External Consumer Data and Information Source, or "ECDIS"

# Transfert learning and actuarial science, wrap-up

- Finance

"Past performance is no guarantee of future returns," Brain (2010)

- Climate models

"The common investment advice that '*past performance is no guarantee of future returns*' and to '*own a portfolio*' appears also to be relevant to climate projections," Reifen and Toumi (2009)

- Use of proxies

"In practice, we often have limited data on the true predictive task of interest, and must instead rely on more abundant data on a closely-related proxy predictive task (...) hospitals often rely on medical risk scores trained on a different patient population (proxy) rather than their own patient population (true cohort of interest) to assign interventions. Yet, not accounting for the bias in the proxy can lead to sub-optimal decisions." Bastani (2021)

# References


Adjaye-Gbewonyo, D., Bednarczyk, R. A., Davis, R. L., and Omer, S. B. (2014). Using the bayesian improved surname geocoding method (bisg) to create a working classification of race and ethnicity in a diverse managed care population: a validation study. *Health services research*, 49(1):268–283.

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

Astesan, E. (1938). Les réserves techniques des sociétés d'assurances contre les accidents d'automobiles. *Collection d'études sur le droit des assurances*.

Bastani, H. (2021). Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984.

Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213.

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.

Bousquet, O., Boucheron, S., and Lugosi, G. (2003). Introduction to statistical learning theory. In *Summer school on machine learning*, pages 169–207. Springer.

# References


Bozinovski, S. and Fulgosi, A. (1976). The influence of pattern similarity and transfer learning upon training of a base perceptron b2. In *Proceedings of Symposium Informatica*, volume 3, pages 121–126.

Brain, J. (2010). "past performance is not necessarily indicative of future results"—the proven-in-use argument and the retrospective application of modern standards. In *5th IET International Conference on System Safety 2010*, pages 1–4. IET.

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

Bénéplanc, G., Charpentier, A., and Thourot, P. (2022). *Manuel d'assurance*. Presses Universitaires de France.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734.

Charpentier, A. (2014). *Computational actuarial science with R*. CRC press.

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

Charpentier, A. and Geoffard, P. (2024). To sue or not to sue. *in progress*.

Charpentier, A., James, M., and Ali, H. (2022). Predicting Drought and Subsidence Risks in France. *Natural Hazards and Earth System Sciences*.

# References


Chavas, D., Yonekura, E., Karamperidou, C., Cavanaugh, N., and Serafin, K. (2013). Us hurricanes and economic damage: Extreme value perspective. *Natural Hazards Review*, 14(4):237–246.

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.

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

D'Agostino, R. B., Grundy, S., Sullivan, L. M., Wilson, P., Group, C. R. P., et al. (2001). Validation of the framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *Journal of the American Medical Association*, 286(2):180–187.

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

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.

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

# References


Denuit, M. and Charpentier, A. (2005). *Mathématiques de l'assurance non-vie: Tome II Tarification et provisionnement*. Economica.

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

Denuit, M. and Robert, C. (2007). *Actuariat des assurances de personnes: modélisation, tarification et provisionnement*. Economica.

Du, T., Xiong, L., Xu, C.-Y., Gippel, C. J., Guo, S., and Liu, P. (2015). Return period and risk analysis of nonstationary low-flow series under climate change. *Journal of Hydrology*, 527:234–250.

Fong, C., Hazlett, C., and Imai, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177.

Frazier, R. (2021). California's ban on climate-informed models for wildlife insurance premiums. *Ecology L. Currents*, 48:24.

Freedman, D. A. and Berk, R. A. (2008). Weighting regressions by propensity scores. *Evaluation review*, 32(4):392–409.

Fuller, W. E. (1914). Flood flows. *Transactions of the American Society of Civil Engineers*, 77(1):564–617.

# References


Furht, B., Villanustre, F., Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). Transfer learning techniques. In *Big data technologies and applications*, pages 53–99. Springer.

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.

Goodman, L. A. (1953). Ecological regressions and behavior of individuals. *American sociological review*, 18(6).

Goodman, L. A. (1959). Some alternatives to ecological correlation. *American Journal of Sociology*, 64(6):610–625.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

Gumbel, E. J. (1941a). Probability-interpretation of the observed return-periods of floods. *Eos, Transactions American Geophysical Union*, 22(3):836–850.

Gumbel, E. J. (1941b). The return period of flood flows. *The annals of mathematical statistics*, 12(2):163–190.

Gumbel, E. J. (1958). *Statistics of extremes*. Columbia university press.

# References


Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hazen, A. (1930). *Flood flows: a study of frequencies and magnitudes*. Wiley.

Heckman, J. (1974). Shadow prices, market wages, and labor supply. *Econometrica: journal of the econometric society*, pages 679–694.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4*, pages 475–492. NBER.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161.

Heinze-Deml, C. and Meinshausen, N. (2021). Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2):303–348.

Herbster, M. and Warmuth, M. K. (1998). Tracking the best expert. *Machine learning*, 32:151–178.

Hjorth, U. (1982). Model selection and forward validation. *Scandinavian Journal of Statistics*, pages 95–105.

Hjorth, U. (1994). *Computer intensive statistical methods: Validation, model selection, and bootstrap*. Chapman and Hall/CRC.

# References


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

Holt, D., Steel, D. G., Tranmer, M., and Wrigley, N. (1996). Aggregation and ecological effects in geographically based data. *Geographical analysis*, 28(3):244–261.

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

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

Hume, D. (1748). *An Enquiry concerning Human Understanding*. Cambridge University Press.

Imai, K., Olivella, S., and Rosenman, E. T. (2022). Addressing census data problems in race imputation via fully bayesian improved surname geocoding and name supplements. *Science Advances*, 8(49):eadc9824.

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

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning*. Springer.

Jia, L., DelSole, T., and Tippett, M. K. (2014). Can optimal projection improve dynamical model forecasts? *Journal of Climate*, 27(7):2643–2655.

# References


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.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Karimi, H., Khan, M. F. A., Liu, H., Derr, T., and Liu, H. (2022). Enhancing individual fairness through propensity score matching. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

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.

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

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

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

# References


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.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting us mortality. *Journal of the American statistical association*, 87(419):659–671.

Li, F. and Li, F. (2019). Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics*, 13:2389–2415.

Li, N. and Lee, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, 42:575–594.

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

Liu, J., Hong, Y., D'Agostino Sr, R. B., Wu, Z., Wang, W., Sun, J., Wilson, P. W., Kannel, W. B., and Zhao, D. (2004). Predictive value for the chinese population of the framingham chd risk assessment tool compared with the chinese multi-provincial cohort study. *Journal of the American Medical Association*, 291(21):2591–2599.

Lorenz, E. N. (1996). Predictability: A problem partly solved. In Palmer, T. and Hagedorn, R., editors, *Predictability of Weather and Climate*, volume 1. Cambridge University Press.

# References


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

Machado, A. F., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024b). Probabilistic scores of classifiers, calibration is not enough. *in progress*.

Mack, T. (1991). A simple parametric model for rating automobile insurance or estimating ibnr claims reserves. *ASTIN Bulletin: The Journal of the IAA*, 21(1):93–109.

Martinez, A. B. (2020). Improving normalized hurricane damages. *Nature Sustainability*, 3(7):517–518.

Mill, J. S. (1848). *Principles of Political Economy*. Baldwin, Cradock, and Joy.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Monteleoni, C. and Jaakkola, T. (2003). Online learning of non-stationary sequences. *Advances in Neural Information Processing Systems*, 16.

Monteleoni, C., Schmidt, G. A., Saroha, S., and Asplund, E. (2011). Tracking climate models. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(4):372–392.

Moodie, E. E. and Stephens, D. A. (2022). Causal inference: Critical developments, past and future. *Canadian Journal of Statistics*, 50(4):1299–1320.

# References


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.

Murray, W. and Sinnreich, R. H. (2006). *The past as prologue: The importance of history to the military profession*. Cambridge University Press.

Neddermeyer, J. C. (2009). Computationally efficient nonparametric importance sampling. *Journal of the American Statistical Association*, 104(486):788–802.

Olsen, J. R., Lambert, J. H., and Haimes, Y. Y. (1998). Risk of extreme events under nonstationary conditions. *Risk Analysis*, 18(4):497–510.

Owen, A. B. (2013). *Monte Carlo theory, methods and examples*. Stanford Lectures Notes.

Pakdaman Naeini, M., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1):2901–2907.

Parthipan, R. and Wischik, D. J. (2022). Don't waste data: Transfer learning to leverage all data for machine-learnt climate model emulation. *arXiv*, 2210.04001.

# References


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

Pielke, R. A. and Landsea, C. W. (1998). Normalized hurricane damages in the united states: 1925–95. *Weather and forecasting*, 13(3):621–631.

Pitacco, E., Denuit, M., Haberman, S., and Olivieri, A. (2009). *Modelling longevity dynamics for pensions and annuity business*. Oxford University Press.

Quarg, G. and Mack, T. (2004). Munich chain ladder. *Blätter der DGVFM*, 26(4):597–630.

Randall, D. A., Wood, R. A., Bony, S., Colman, R., Fichefet, T., Fyfe, J., Kattsov, V., Pitman, A., Shukla, J., Srinivasan, J., et al. (2007). Climate models and their evaluation. In *Climate change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the IPCC (FAR)*, pages 589–662. Cambridge University Press.

Read, L. K. and Vogel, R. M. (2015). Reliability, return periods, and risk under nonstationarity. *Water Resources Research*, 51(8):6381–6398.

Reichler, T. and Kim, J. (2008). How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society*, 89(3):303–312.

Reifen, C. and Toumi, R. (2009). Climate projections: Past performance no guarantee of future skill? *Geophysical Research Letters*, 36(13).

# References


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

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology: methodological issues*, pages 297–331. Springer.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. (2018). Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, pages 159–183.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

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.

Salas, J. D. and Obeysekera, J. (2014). Revisiting the concepts of return period and risk for nonstationary hydrologic extreme events. *Journal of hydrologic engineering*, 19(3):554–568.

# References


Sartin, E. B., Metzger, K. B., Pfeiffer, M. R., Myers, R. K., and Curry, A. E. (2021). Facilitating research on racial and ethnic disparities and inequities in transportation: Application and evaluation of the bayesian improved surname geocoding (bisg) algorithm. *Traffic injury prevention*, 22(sup1):S32–S37.

Satten, G. A. and Datta, S. (2001). The kaplan–meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, 55(3):207–210.

Schmidt, G. (2024). Climate models can't explain 2023's huge heat anomaly — we could be in uncharted territory. *Nature*, 627:467.

Sedgwick, P. (2015). Understanding the ecological fallacy. *British Medical Journal*, 351.

Shakespeare, W. (1610). *The Tempest*. Oxford Paperbacks.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

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

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

Sorbero, M. E., Euller, R., Kofner, A., and Elliott, M. N. (2022). *Imputation of race and ethnicity in health insurance marketplace enrollment data, 2015-2022 open enrollment periods*. RAND.

# References


Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133.

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.

Task Committee on Hydrology Handbook of Management Group D of ASCE (1996). Hydrology handbook. American Society of Civil Engineers.

Taylor, G. C. (1977). Separation of inflation and other effects from the distribution of non-life insurance claim delays. *ASTIN Bulletin: The Journal of the IAA*, 9(1-2):219–230.

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

Twain, M. and Warner, C. D. (1874). *The Gilded Age: a tale of today*. Penguin.

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 der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge university press.

Vapnik, V. N. and Chervonenkis, A. Y. (1971). On uniform convergence of the frequencies of events to their probabilities. *Teoriya Veroyatnostei i ee Primeneniya*, 16(2):264–279.

# References


Verrall, R. (1996). Claims reserving and generalised additive models. *Insurance: Mathematics and Economics*, 19(1):31–43.

Voicu, I. (2018). Using first name information to improve race and ethnicity classification. *Statistics and Public Policy*, 5(1):1–13.

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

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

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.

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

Weinkle, J., Landsea, C., Collins, D., Musulin, R., Crompton, R. P., Klotzbach, P. J., and Pielke Jr, R. (2018). Normalized hurricane damage in the continental united states 1900–2017. *Nature sustainability*, 1(12):808–813.

Wilson, P. W., Castelli, W. P., and Kannel, W. B. (1987). Coronary risk prediction in adults (the framingham heart study). *The American journal of cardiology*, 59(14):G91–G94.

Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., and Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847.