OPINIONS
& DÉBATS

N°25 - Juillet 2022

Assurance : discrimination, biais et équité

Arthur Charpentier

INVESTIR L'AVENIR

LABEX Louis Bachelier

Les articles publiés dans la série “Opinions & Débats” offrent aux spécialistes, aux universitaires et aux décideurs économiques un accès aux travaux de recherche les plus récents. Ils abordent les principales questions d’actualité économique et financière et fournissent des recommandations en termes de politiques publiques.

Opinions & Débats N° 25 - Juillet 2022

Publication de l'Institut Louis Bachelier
Palais Brongniart - 28 place de la Bourse 75002 Paris ◆ Tél. : 01 73 01 93 40 ◆ www.institutlouisbachelier.org
DIRECTEUR DE LA PUBLICATION : Jean-Michel Beacco ◆ **COORDINATION ÉDITORIALE** : Ryadh Benlahrech
CONTACT : ryadh.benlahrech@institutlouisbachelier.org

Les derniers numéros parus dans la collection « Opinions & Débats » :

Cyber-assurance : enjeux, modélisations et leviers de mutualisation
Caroline Hillairet et Olivier Lopez

Scénarios et modèles économie-climat : une grille de lecture pour la finance durable
Jean-Charles Hourcade, Frédéric Ghersi, Romain Grandjean, Julien Lefèvre, Peter Tankov et Stéphane Voisin

Comment favoriser l'accès à l'emploi des jeunes peu qualifiés ?
Pierre Cahuc et Jérémy Hervelin

Robo-Advising : Moins d'IA et plus de XAI ?
Milo Bianchi et Marie Brière

L'évaluation économique des engagements en assurance vie : écueils, bonnes pratiques et préconisations pour une mise en œuvre pertinente
Kamal Armel et Frédéric Planchet

La responsabilité des actionnaires doit-elle toujours être limitée ?
Guillaume Vuillemey

L'intégralité des numéros de la collection « Opinions & Débats » est accessible sur le site web de l'Institut Louis Bachelier : https://www.institutlouisbachelier.org/opinions-debats-louis-bachelier/

SOMMAIRE

EDITO

Jean-Michel Beacco
Délégué général
de l'Institut Louis Bachelier

Le développement de l'apprentissage machine (machine learning) au cours des dix dernières années s'est avéré utile dans de nombreux domaines afin de faciliter l'aide à la décision, particulièrement dans un contexte où les données sont abondantes et disponibles, mais difficilement manipulables par les humains.

Dans le secteur financier, les algorithmes sont couramment utilisés par les opérateurs à haute fréquence, les gestionnaires d'actifs ou les hedge funds pour tenter de prédire l'évolution des marchés financiers.

Le secteur assurantiel ne fait pas exception à la règle. Les assureurs recourent de plus en plus à une segmentation fine de leurs assurés ou futurs clients pour les catégoriser dans des sous-groupes homogènes en termes de risques et ainsi individualiser les tarifs de leurs contrats en fonction des risques encourus. Cependant, le recours massif aux algorithmes et à des outils fonctionnant avec de l'Intelligence Artificielle (IA) par les actuaires pour segmenter les assurés remet en cause le principe même sur lequel se base l'assurance, à savoir la mutualisation des risques entre tous les assurés.

Dans ce contexte, où le numérique est davantage exploité, plusieurs problématiques se posent, dont : comment le business model du secteur doit-il évoluer si l'individualisation se massifie au détriment de la mutualisation ? Comment les assureurs peuvent-ils effectuer de la segmentation sans appliquer de critères discriminatoires ? Quels sont les biais à éviter dans la construction d'algorithmes ? Quid des critères d'équité, une notion à la fois abstraite, mais très ancrée dans notre société ?

Dans ce nouveau numéro de la collection Opinions & Débats, Arthur Charpentier, un chercheur spécialisé dans les questions liées au secteur assuranciel et aux données massives, a réalisé un travail complet pour tenter de répondre aux enjeux posés par les notions de discrimination, de biais et d'équité dans les assurances. Outre les débats très intéressants posés par ces sujets, Arthur a effectué une revue exhaustive de la littérature académique existante, tout en y apportant des démonstrations et explications mathématiques.

Bonne lecture !

BIOGRAPHIE

Arthur Charpentier, PhD en mathématiques appliquées (KU Leuven), MSc en statistique et économie (ENSAE, Paris), est professeur à l'Université du Québec à Montréal, actuaire agrégé, membre de l'Association Actuarielle Internationale, et Louis Bachelier Fellow. Il a publié avec Michel Denuit les deux tomes de *Mathématiques de l'Assurance Non-Vie*, (Economica Eds.) en 2004 et 2005, et a édité *Computational Actuarial Science with R* (CRC Press Eds.) en 2015. Il publie à la rentrée 2022 un *Manuel d'Assurance* (Presses Universitaires de France), écrit avec Gilles Bénéplanc et Patrick Thourot.

http://freakonometrics.github.io/

charpentier.arthur@uqam.ca

L'auteur tenait à remercier, par ordre alphabétique, Avner Ben-Har, Laurence Barry, Philippe Besse, Rodolphe Bigot, Émilie Biland-Curinier, Baptiste Coulmont, Olivier l'Haridon, Jean Michel Loubes, Elisabeth Vallet et Bertrand Villeuneuve, pour les retours et les discussions qu'ils ont pu avoir.

Dans ce rapport, conformément à l'usage dans la littérature, le genre sera souvent utilisé comme variable "protégée" pour illustrer différentes notions en lien avec les discriminations, et sera traité comme une variable binaire (prenant deux valeurs "homme" et "femme"). Il ne s'agit aucunement d'une prise de position contre le concept de "non-binarité" (ou "*genderqueer*" en anglais) utilisé en sciences sociales. Une discussion plus poussée sur ce point sera proposée dans la section 2.2.2.

Le titre de ce rapport, *Assurance : biais, discrimination et équité* fait écho à Bertail *et al.* 2019.

Résumé

Les **données massives** et les performances obtenues par les algorithmes d'**apprentissage automatique** ont chamboulé l'assurance et l'actuariat. Les questions soulevées par ces nouveaux outils dans d'autres contextes (que ce soit la justice prédictive (ou justice "actuarielle" comme l'appelle Harcourt 2008) ou les débats sur les *fake news*, en passant par les véhicules autonomes et la médecine prédictive) poussent les actuaires au doute, et à la méfiance. Kranzberg 1986 affirmait que "*technology is neither good nor bad; nor is it neutral*", mettant en avant que, même sans mauvaises intentions, les algorithmes d'apprentissage pouvaient être injustes. Et corriger ces possibles injustices n'est pas simple. Pour Nielsen 2020 "*technology does not necessarily self-regulate, via either market or social pressures*" (la main invisible des marchés ou de la pression sociale ne suffira peut être pas). C'est dans ce contexte que nous allons revenir ici sur les problématiques de biais, de discrimination et d'équité, des modèles prédictifs utilisés en assurance. Ces changements, tant sur les données que sur les modèles, que l'on observe depuis une petite dizaine d'années, avaient déjà questionné l'existence même de l'assurance. Pour Löffler *et al.* 2016 "*this leads to demutualization and a focus on predicting and managing individual risks rather than communities*", l'**individualisation** de plus en plus grande des primes force à s'interroger sur l'avenir de la **mutualisation**, et de la **solidarité** entre assurés. Les problèmes de discrimination sont alors à envisager dans ce contexte de perte de solidarité. Car paradoxalement, la discrimination n'a de sens qu'en voyant l'individu en tant que membre d'un groupe caractérisé par un trait partagé (les femmes, les personnes d'origine étrangères, les personnes âgées, etc).

Ce principe de mutualisation des risques se traduit par le fait que l'**assurance** est "*the contribution of the many to the misfortune of the few*". De par l'inversion du cycle de production, l'assureur vend au souscripteur une promesse d'indemnisation, dans le futur, d'un risque aléatoire, en échange du paiement d'une contribution "juste" ou "équitable", a priori proportionnelle au risque de l'assuré (Thiery et Van Schoubroeck 2006 parleront d'**équité actuarielle**). Le vrai facteur de risque sous-jacent étant une information non observable lors de la signature du contrat, l'assureur va construire des **algorithmes prédictifs**, à partir d'information disponible, pour prédire la fréquence de sinistre, le coût des sinistres, mais aussi la probabilité de frauder, ou la probabilité de souscrire une garantie supplémentaire par exemple. En ne voyant plus un groupe d'assurés comme une mutualité parfaitement homogène, les actuaires ont utilisé des algorithmes de plus en plus fins pour créer des sous-groupes davantage homogènes. Avec le développement des techniques d'**apprentissage machine**, l'idée de personnalisation, d'individualisation (très présente dans la communauté informatique depuis plusieurs années, comme le soulignaient Adomavicius et Tuzhilin 2005 avec des "profils" individualisés) fait son chemin, et pousse les assureurs à démutualiser de plus en plus. "*At the core of insurance business lies discrimination between risky and non-risky insureds*" avait affirmé Avraham 2017. Aussi, d'un côté, l'opération d'assurance relève de

la technique et a fondamentalement une dimension collective, reposant sur la mutualisation des risques au sein de groupes de risques homogènes. Les systèmes de classification des assurances reposent sur l'hypothèse que les individus répondent aux caractéristiques moyennes (**stéréotypées** d'une certaine manière) d'un groupe auquel ils appartiennent. C'est la discrimination au sens statistique (mise en œuvre par des outils statistiques puis économétriques). De plus, le contrat d'assurance relève du droit, et a une dimension individuelle. En ce sens, un individu ne peut être traité différemment en raison de son appartenance à tel ou tel groupe, en particulier à un groupe auquel il n'a pas choisi d'appartenir, sinon c'est de la discrimination, au sens légal du terme. Et dans le contexte de données de plus en plus massives, et d'algorithmes prédictifs de plus en plus complexes (pour ne pas utiliser le terme de "**boîte noire**"), il est devenu de plus en plus difficile de garantir que les assureurs demandent une contribution "juste" aux souscripteurs de polices d'assurance.

Réfléchir aux égalités de traitement des assurés revient à s'interroger sur la possibilité même de souscrire un contrat, en vue d'une couverture, mais aussi à l'idée de demander une prime non prohibitive, et non dissuasive. Car contrairement à ce que nous apprennent les mathématiques financières (et l'hypothèse de marchés complets, Froot *et al.* 1995), il n'existe pas en assurance de **loi du prix unique**, le prix d'un risque étant vu au travers d'une mutualité d'assurés, et d'un modèle de tarification. De plus, les souscripteurs n'achètent pas "une assurance", mais une **garantie** de couverture contre certains risques. Si certaines garanties sont souscrites majoritairement par certaines populations, et pas par d'autres, la différence de prix ne correspond pas forcément à une discrimination, *stricto sensu*. C'est dans ce contexte que nous allons discuter des biais, des discriminations et de l'équité en assurance.

Les **données**, de plus en plus massives, posent de nombreux défis. Tout d'abord, la réglementation cherche à protéger des informations dites **sensibles** ou **protégées**, interdisant parfois de collecter et de stocker certaines variables. Le principal danger est qu'il devient alors difficile d'assurer qu'un modèle ne discrimine pas suivant un critère, si ce critère n'est pas observé. Poser un voile d'ignorance sur certaines caractéristiques ne suffit pas pour imposer l'équité d'un modèle, et ne sert qu'à masquer un potentiel problème (ou comme l'affirmaient Kearns et Roth 2019, "*machine learning won't give you anything like gender neutrality 'for free' that you didn't explicitely ask for*"). Un autre défi est celui des innombrables **biais** des données collectées à travers toutes sortes de sources (questionnaires, objets connectés, données obtenues via différentes sources, etc). Parmi ces derniers, on peut mentionner les biais de variable manquante, les biais de définition ou d'interprétation, les biais de mesure, les biais de survie, les biais de rétroaction, etc. Ces "*dark data*" (pour reprendre le terme utilisé par Hand 2020) forcent à s'interroger sur la pertinence d'une classification des risques, certaines discriminations étant parfois perçues sur la base d'informations biaisées, ou mal interprétées. Si le genre du conducteur principal a longtemps été utilisé par les assureurs, on peut s'interroger sur sa signification dans un couple (hétérosexuel) partageant une voiture. On retrouve ici la difficulté de la **définition** des variables, bien connue par les statisticiens. On reviendra ainsi sur le **paradoxe de Simpson** et l'inférence écologique (en anglais, on parle d'*ecological fallacy*) où l'absence de certaines variables peut donner une interprétation fausse, fallacieuse, sur le sens d'une potentielle discrimination. Et dans le contexte d'assurance, les données télématiques, et les mécanismes incitatifs de type "*gamification*" posent des questions sur les **biais de rétroaction**, les assureurs ayant la possibilité d'influencer directement les comportements de tel ou tel assurés, sur la base de données arrivant en temps réel. On retrouve ici une forme de **biais de sélection**, ce dernier signifiant simplement que les données historiques ont été collectées sur des personnes qui ont choisi de souscrire un contrat et qui ont

été acceptées par un assureur au préalable (potentiellement sur la base d'un précédent modèle). Tout comme l'analyse de la fraude ne peut pas se faire de la même manière, si les enquêtes en lien avec la fraude sont menées de manière aléatoire ou si elles reposent sur un modèle préalable de détection de fraude. On retrouve les débats classiques entre les données d'expériences (souvent **randomisées**, pour reprendre le terme anglais) et les données administratives ou observationnelles.

On l'a déjà mentionné, une notion centrale sera celle de **discrimination**, terme particulièrement ambigu, puisque les actuaires utiliseront la version statistique du terme (on peut penser à l'analyse discriminante introduite par Ronald Fisher), alors que les juristes y voient un traitement inégal et défavorable appliqué à certaines personnes en raison de certains critères. Même s'il existe des différences culturelles, entre les pays, on retrouvera souvent un certain nombre de caractéristiques protégées (par la **morale**, ou par la **loi**) comme le genre ou le sexe de la personne, la race ou l'origine nationale ou ethnique, le handicap et toute information génétique, etc. Ces critères sont parfois présentés comme des **clubs** dans lesquels on tombe à la naissance, pour reprendre l'expression de Macnicol 2006 (qui font aussi écho au concept de "voile d'ignorance" et de "loterie génétique"). D'autres critères comme l'âge sont plus complexes car un assuré traversera tous les âges au cours de sa vie : s'il y a une "discrimination" contre les jeunes, l'assuré en souffrira à 20 ans alors qu'il est dans le groupe défavorisé, avant de passer progressivement dans le groupe privilégié (sans évoquer une possible solidarité inter-générationnelle). Enfin, des critères relèvent davantage de choix, plus ou moins conscients. Une première difficulté est que de nombreuses discriminations ne sont pas intentionnelles. Pire encore, contrairement à ce qui peut exister dans la littérature traditionnelle sur les discriminations (où des proxys sont potentiellement utilisés à la place d'une variable sensible, comme le *redlining* où les quartiers d'une ville sont un proxy d'une information éthique et raciale), en assurance, certaines variables sensibles (comme le genre) ont longtemps été utilisées comme proxy d'informations difficilement accessibles (comme des informations comportementales en matière de conduite automobile). Une autre difficulté repose sur un problème classique en grande dimension, et sur la multicolinéarité des variables prédictives. Ceci peut donner lieu à une **discrimination par proxy** (parfois appelée **discrimination statistique** ou **discrimination indirecte** dans les directives européennes en lien avec la discrimination), qui consiste à utiliser une variable très corrélée à la variable protégée. L'utilisation intensive de proxys (non détectées) dans le développement de modèles a soulevé des inquiétudes quant à l'équité. Et l'enrichissement de données rajoute de plus en plus de variables pouvant être vues comme engendrant une discrimination indirecte.

La dernière notion que nous décrirons est la notion d'**équité** d'un modèle prédictif. Après un rapide survol des concepts de **justice**, nous présenterons les mesures classiques d'équité qu'il est possible d'utiliser pour quantifier l'ampleur d'une possible discrimination. Si on formalise rapidement, on dispose d'un triplet $(y, \boldsymbol{x}, p)$, où y est une variable d'intérêt (nombre de sinistres, coût annuel, nombre de visites chez le médecin, etc), $\boldsymbol{x}$ un ensemble de variables explicatives admissibles, utilisées pour prédire y, et p une variable sensible, ou protégée (supposée unique, ici). Construire un modèle prédictif $\widehat{y} = m(\boldsymbol{x})$ en utilisant seulement les variables $\boldsymbol{x}$ et pas p ne suffit pas à garantir que le modèle ne puisse pas discriminer suivant p, tout simplement car p peut être très corrélée à certaines caractéristiques $\boldsymbol{x}$ (on retrouve l'idée de proxy). Barocas, Hardt *et al.* 2019 notent que les grands principes associés à l'équité se traduisent (1) par une indépendance entre $\widehat{y}$ et p, autrement dit la prédiction n'a rien à voir avec le groupe de p (2) par une notion de séparation : $\widehat{y}$ est indépendante de p étant donné y, et (3) une notion de suffisance : y est indépendante de p étant donné $\widehat{y}$. Ces principes vont se traduire par différentes notions d'équité de groupe, les plus populaires étant la **parité démographique** et la notion d'**égalité des chances**.
Ces notions (dites de groupe), très populaires et largement utilisées (par exemple sur le marché

du travail, aux États-Unis) sont à distinguer des approches individuelles qui émergent dans la littérature scientifique, inspirées des techniques d'inférence causale et visant à chercher à un **contrefactuel** afin de répondre à la question *que se serait-il passé si l'assuré avait la caractéristique* $p = 1$ *au lieu de* $p = 0$? (si on suppose que la variable protégée est binaire, $p \in \{0, 1\}$). C'est une relation **causale** entre la variable sensible p et la variable de risque y, qui peut légitimer une discrimination statique, comme le suggérait la Commission Européenne, qui proposait d'autoriser "*des différences proportionnelles dans les primes et les prestations des particuliers lorsque l'utilisation du sexe est un* **facteur déterminant** *dans l'évaluation du risque, sur la base de données actuarielles et statistiques pertinentes et précises*". Néanmoins, la présence de proxys pose de nombreux défis, car l'approche contrefactuelle usuelle (consistant à changer la variable protégée p seulement, *ceteris paribus*) n'a pas de sens en grande dimension, en présence de proxys fortement corrélés à la variable sensible : une intervention (conceptuelle et fictive) sur la variable sensible p doit avoir un impact sur une ou plusieurs variables prédictives $\boldsymbol{x}$, et donc sur la prévision.

D'autres concepts seront aussi évoqués ici, sans pour autant faire l'objet de chapitres spécifiques, comme la **responsabilité**. En effet, si un algorithme reproduit ce qu'il observe dans les données, peut-il être jugé responsable de reproduire les biais sociaux? Sous un angle épistémologique, on demandait historiquement aux modèles de bien "décrire le réel" (ou disons le réel tel qu'il apparaît dans les données, on parlera d'**accuracy** en apprentissage statistique), c'est-à-dire "ce qui est", alors qu'en introduisant une dimension morale et éthique, on demande que le modèle soit en accord avec ce qui "devrait être", suivant une norme éthique (la fameuse opposition "*is–ought*" de Hume 1739), ou entre la "normalité" statistique opposée à la norme morale. L'autre souci est que pour quantifier l'équité, il convient d'avoir accès à ces données personnelles, privées et sensibles, ce qui renvoie aux discussions sur la **vie privée** (ou *privacy*) et la **conformité** (ou *compliance*). Finalement, on le verra tout au long du rapport, ces discussions autour de la discrimination, des biais et de l'équité sont très proches de celles portant sur l'**interprétation** des modèles prédictifs et de la notion d'**explicabilité**. Cet aspect **narratif** de la construction de modèle est important, en particulier lorsque l'on cherche à créer des **graphs causaux dirigés** afin de comprendre les liens entre la variable protégée p, les possibles variables prédictives $\boldsymbol{x}$ et la variable d'intérêt y. Mais en grande dimension, cet exercice devient vite impossible. En affirmant que "*all models are wrong but some models are useful*", Georges Box insistait sur l'aspect narratif de la modélisation, sur l'interprétation qui en découle. Une compréhension fine des données et des modèles est aujourd'hui indispensable, l'époque des calculs froids et objectifs (ou supposés objectifs) des actuaires semblant révolue.

.

Chapitre 1

Introduction et motivations

1.1 Motivations

Après plusieurs mois d'enquêtes, Angwin *et al.* 2016 revenaient sur l'outil *Correctional Offender Management Profiling for Alternative Sanctions*, ou COMPAS, largement utilisé comme outil d'aide à la décision aux État-Unis pour évaluer le risque de récidive d'un criminel, dans le cadre d'une série intitulée *Machine Bias* (et sous-titrée *Investigating Algorithmic Injustice*). Depuis, de nombreux ouvrages et articles sont revenus sur les points soulevés par cet article, à savoir le pouvoir grandissant pris par ces outils prédictifs d'aide à la décision, leur opacité qui ne cesse de croître, les discriminations qu'elles reproduisent (ou amplifient), les données "biaisées" utilisées pour entraîner ou calibrer ces algorithmes, et le sentiment d'injustice engendré. Les données massives et l'apprentissage machine ont été l'occasion de redécouvrir un sujet abordé par les juristes, les économistes, les philosophes et les statisticiens depuis au moins cinquante ans. Nous allons revenir ici sur ces réflexions, les éclairer d'un regard nouveau, en mettant l'emphase sur l'assurance, et en étudiant les réponses possibles. Les juristes, en particulier, ont beaucoup discuté ces modèles prédictifs, cette "justice actuarielle" comme l'appellent R. G. Thomas 2007, Harcourt 2011, Gautron et Dubourg 2015 ou Rothschild-Elyassi *et al.* 2018.

L'idée des biais et de la discrimination algorithmique n'est pas nouvelle, comme en témoigne Pedreshi *et al.* 2008, par exemple. Mais depuis, les exemples ne cessent de se multiplier : "*A.I. Bias Caused 80 % Of Black Mortgage Applicants To Be Denied*" (Hale 2021), ou "*How the use of A.I. runs the risk of re-creating the insurance industry's inequities of the previous century*" (Ito 2021). Poursuivant l'analyse de David 2015, McKinsey 2017 annonçait que l'intelligence artificielle allait bouleverser le monde du travail (y compris dans les secteurs de l'assurance et de la banque, Mundubeltz-Gendron 2019) en particulier pour remplacer un travail (humain) répétitif peu reluisant [1]. Ces remplacements soulèvent des questions, et forcent le marché et le régulateur à la prudence : pour Reijns *et al.* 2021, "*the Dutch insurance industry makes it a mandate*", dans un article portant sur *ethical artificial intelligence*, et en France, Défenseur des droits 2020 rappelait que "*les biais algorithmiques doivent pouvoir être identifiés puis corrigés*" car "*la non-discrimination*

1. Même si cela semble exagéré, car au contraire, ce sont souvent des humains qui effectuent les tâches répétitives pour aider des robots : "*dans la plupart des cas, la tâche est répétitive et machinale. L'un des travailleurs a expliqué qu'il avait un jour dû écouter des enregistrements pour trouver ceux contenant le nom de la chanteuse Taylor Swift afin d'enseigner à l'algorithme qu'il s'agit d'une personne*" comme le rapportait Radio Canada en avril 2019.

n'est pas une option, mais renvoie à un cadre juridique". Bergstrom et West 2021 notaient, avec un brin d'ironie, qu'il y avait du monde pour rédiger une déclaration des droits des robots, ou à la conception de moyens de protéger l'humanité contre les machines super-intelligentes, de type Terminator, mais qu'entrer dans les détails de l'audit algorithmique est souvent perçu comme ennuyeux, mais indispensable : "*to address the problems that AI is creating now, we need to understand the data and algorithms we are already using for more mundane purposes*".

Masquer un problème le résout rarement, même si c'est la solution qui a longtemps été préconisée pour lutter contre les discriminations. Comme l'ont montré Budd *et al.* 2021, revenant sur une expérience d'Amazon, supprimer les noms des CV, pour éliminer la discrimination entre les hommes et les femmes, ne fonctionne pas, puisqu'en cachant le nom du candidat (ou de la candidate), l'algorithme a continué de choisir préférentiellement des hommes plutôt que des femmes. Pourquoi ? Tout simplement parce qu'Amazon a entraîné l'algorithme à partir de ses CV existants, avec une sur-représentation des hommes, et certains éléments d'un CV (outre le nom) peuvent révéler le genre d'une personne, par exemple un diplôme d'une université pour femmes, l'appartenance à une organisation professionnelle féminine, ou un passe-temps où les sexes sont représentés de manière disproportionnée. Des proxys plus ou moins corrélés à la variable "protégée" peuvent entretenir une forme de discrimination.

Dans ce document, nous allons revenir sur ces questionnements, en nous limitant aux modèles actuariels dans un contexte assurantiel, et presque exclusivement, la tarification de contrats d'assurance. Seligman 1983 posait la question simple suivante : "*if young women have fewer automobile accidents than young men — which they do — why shouldn't the women get a better rate ? If the industry's experience shows — as it does — that women spend more time in the hospital why shouldn't women pay more for*". Ce type de question sera le point de départ des réflexions que nous aurons ici. Paraphrasant Georges Clémenceau qui affirmait (en 1887) que "*la guerre, c'est une chose trop grave pour la confier à des militaires*", Wortham 1985 affirmait que la segmentation en assurance [2] était une tâche trop importante pour être laissée aux actuaires. Vingt-cinq années plus tard, on peut se demander s'il n'est pas encore plus grave de la laisser à des algorithmes, et de clarifier le rôle des actuaires dans ces débats. Dans cette introduction, nous allons commencer par revenir sur la segmentation en assurance, et les fondements de la tarification actuarielle des contrats d'assurance. Nous reviendrons ensuite sur les différents termes mentionnés dans le titre, à savoir la notion de biais, de discrimination et d'équité (ou en anglais, de "*fairness*"), tout en proposant une typologie des modèles prédictifs et des données (en particulier des données dites "sensibles", pouvant être en lien avec une possible discrimination).

1.2 Fondements de la tarification actuarielle

L'activité d'assurance "*se caractérise par un cycle inversé de production : en contrepartie d'une prime dont le montant est connu à la souscription du contrat, l'assureur s'engage à couvrir un risque dont il ignore la date de réalisation et le montant* ", selon la définition de Tendil 1998. Et pour se faire, l'assureur va rassembler les risques au sein d'une mutualité. "*Le secret universel de l'assurance est [donc] le regroupement d'un grand nombre de contrats d'assurance au sein d'une mutualité, pour que se réalisent des compensations entre les risques sinistrés et ceux pour lesquels l'assureur aura perçu des primes sans avoir dû régler des prestations*", comme l'affirmait

2. On parlera, en anglais, de "*insurance classification*".

Petauton 1998. Pour reprendre la formulation de Chaufton 1886, l'assurance est la "*compensation des effets du hasard par la mutualité organisée suivant les lois de la statistique*".

En allant un peu plus loin, l'assurance n'élimine pas le risque, elle le déplace, et ce transfert se fait en accord avec une philosophie sociale choisie par l'assureur. En raisonnant au sens large, en plus des "mutuelles" et des "compagnies d'assurance", on peut inclure les "assurances publiques". Ces assurances publiques, comme le rappelle Ewald 1986, visent à transférer le risque des individus vers un groupe social plus large, en "socialisant", ou en redistribuant le risque de manière "plus équitable" au sein de la population. A fortiori, des individus à faible risque paient des primes d'assurance à un taux plus élevés que leur profil de risque ne le suggérerait, même si cela semble "inefficace" sous un angle économique. L'assurance sociale, organisée selon des principes de solidarité, où l'accès et la couverture sont indépendants du statut de risque, et parfois de la capacité à payer ("*ability to pay*" comme le note Mittra 2007, même si dans de nombreux cas, la prime est proportionnelle au revenu de l'assuré) est généralement fournie par des entités publiques plutôt que privées. "*For certain social goods, such as healthcare, long-term care and perhaps even basic life assurance attached to a mortgage, it may simply be inappropriate to provide such products through a mutuality-based model that inevitably excludes some individuals*" car "*primary social goods, because they are defined as something everybody has an inalienable right or entitlement to, cannot be distributed through a system that excludes individuals based on their risk status or ability to pay*". Les mutuelles sont souvent vues comme un intermédiaire entre ces assurances publiques, et les compagnies d'assurance à but lucratif.

Et comme le souligne Lasry 2015, "*l'assurance est depuis longtemps confrontée à un dilemme : d'une part, mieux connaître un risque, c'est mieux le tarifer; mieux connaître les facteurs de risque peut aussi permettre d'encourager la prévention; d'autre part, la mutualisation, qui est le fondement de l'assurance, ne peut subsister dans la plupart des cas que dans une situation de relative ignorance (voire d'une obligation légale d'ignorance)*". Les actuaires vont alors chercher à classifier ou segmenter les risques, tout reposant sur l'idée de mutualisation. Nous reviendrons sur le formalisme mathématique de ce dilemme. De Pril et Dhaene 1996 (repris en français dans Corlier 1998) soulignaient que la segmentation est "*une technique que l'assureur utilise pour différencier la prime et éventuellement aussi la couverture, en fonction d'un certain nombre de caractéristiques spécifiques du risque à assurer [ci-après appelés critères de segmentation] et ce aux fins de parvenir à une meilleure concordance entre le coût estimé du sinistre et les frais qu'une personne déterminée met à charge de la collectivité des preneurs d'assurance et la prime que cette personne doit payer pour la couverture offerte*".

La souscription est le terme utilisé pour décrire le processus de décision par lequel les assureurs déterminent s'ils vont offrir, ou refuser, une police d'assurance à un individu sur la base des informations disponibles (et le montant demandé). Gandy 2016 affirme que "droit de souscription" ("*right to underwrite*") est fondamentalement un droit à la discrimination. En France, "sélection des risques" a un sens légal assez particulier, comme le rappel l'encadré suivant.

Assurance & sélection des risques, par Rodolphe Bigot [3]

L'opération d'assurance et la mutualisation sous-jacente reposent sur ce qu'on appelle la sélection des risques. Mises à part la plupart des assurances collectives qui consistent en une sorte de mutualisation dans la mutualisation, l'assureur refuse ou accepte ainsi chaque candidat à l'assurance à entrer dans la mutualisation constituée du groupe des assurés. Cette sélection des risques « *cantonne la mutualisation aux assurés acceptés par l'assureur qui est encore considéré, dans le droit du contrat d'assurance, comme celui qui accepte le contrat qui lui est proposé par le candidat à l'assurance* » (Monnet 2017, p. 13 et suivantes). Rappelons à ce titre que l'économie de l'opération d'assurance exige de laisser à l'entreprise d'assurance une large liberté d'accepter ou de refuser le risque qui lui est proposé. Preuves de cette liberté en ce qui concerne l'assurance de personnes, les dispositions de l'article 225-3 du Code pénal qui excluent du champ de la répression pénale de la discrimination en matière de fourniture de biens et de services prévue à l'article 225-2 les « *discriminations fondées sur l'état de santé, lorsqu'elles consistent en des opérations ayant pour objet la prévention et la couverture du risque décès, des risques portant atteinte à l'intégrité physique de la personne ou des risques d'incapacité de travail ou d'invalidité* ». Toutefois, n'admettre aucune limite à cette liberté de l'entreprise d'assurance conduirait à chasser des considérations sociales importantes et à exclure non seulement de l'assurance mais encore des biens et des services qui y sont liés (tel l'emprunt, donc l'accession à la propriété) les personnes les plus exposées (Bigot et Cayol 2020, p. 540). La question du droit à l'assurance se pose ici (Pichard 2006). À cet effet, « *accéder à l'assurance s'entend non seulement de la possibilité même de souscrire un contrat en vue d'une couverture, mais peut-être, également, à un coût économique raisonnable, non prohibitif partant non dissuasif. Dans des sociétés où le besoin de sécurité voire le confort est un leitmotiv, le questionnement a tout son sel* » (Noguéro 2010, p. 633)

1.2.1 Cas d'une population homogène

Avant de parler de segmentation, plaçons-nous dans le cas d'un portefeuille homogène (les assurés font face à la même probabilité d'occurrence d'un sinistre, la même étendue des préjudices, la même somme assurée, etc.) En assurance dommage, l'assureur s'engage (en échange du versement d'une prime dont le montant est décidé lors de la signature du contrat) à couvrir les sinistres qui surviendront dans l'année. Si Y dénote la charge annuelle d'un assuré pris au hasard, la "prime pure" est alors définie [4] par $\mathbb{E}[Y]$. Si on considère le risque de perdre 100 avec une probabilité p (et rien avec probabilité $1-p$), la prime pure de ce risque (les économistes parleraient de loterie) est alors $100p$. La prime d'un contrat d'assurance est alors proportionnelle à la probabilité d'avoir un sinistre. Aussi par la suite, nos exemples se limiteront souvent à l'estimation de cette probabilité. Dans un contrat d'assurance de personne, la période de couverture est plus longue, et il convient d'escompter les payements futurs [5]. Par exemple, considérons un contrat d'assurance décès qui verserait 100 lors du décès d'une personne (à ses ayant-droits) : si T est la durée de vie résiduelle (aléatoire) d'un ou d'une assurée, à la souscription du contrat, alors la prime pure correspond à la valeur actuelle probable des flux futurs, soit

$$a = \mathbb{E}[\frac{100}{(1+r)^T}] = \sum_{t=0}^{\infty} \frac{100}{(1+r)^t} \cdot \mathbb{P}[T=t],$$

3. Maître de conférences en droit privé, UFR de Droit, Le Mans Université, membre du Thémis-UM et Ceprisca.

4. Denuit et Charpentier 2004 revient sur le formalisme mathématique qui autorise cette écriture, en particulier l'espérance est calculée suivant une probabilité $\mathbb{P}$ correspondant à la probabilité "historique" (il n'y a pas de loi du prix unique en assurance, contrairement à l'approche classique en finance de marché, au sens de Froot *et al.* 1995). L'assurance, comme l'économétrie, vit dans un monde probabiliste, peut être contrairement à de nombreux algorithmes d'apprentissage machine, qui peuvent fonctionner sans modèle probabiliste, comme le note Charpentier, Flachaire *et al.* 2018.

5. Sans escompte, le décès étant (à horizon temporel infini) certain, la prime pure serait exactement le montant du capital versé aux ayant-droits.

pour un taux d'escompte r. Mais cette hypothèse de risques homogènes est clairement trop forte dans de nombreuses applications en assurance, par exemple dans le cas de l'assurance décès, car la loi de T devrait dépendre (par exemple) de l'âge de l'assuré à la souscription du contrat.

1.2.2 La crainte de l'aléa moral et de l'antisélection

Tous les actuaires ont été bercé par la fable des "*lemons*" de Akerlof 1970. Certains acheteurs d'assurance sont vus comme des pêches ("*peaches*") à faible risque, alors que d'autres sont des citrons ("*lemons*") à haut risque. Dans certains cas, les acheteurs d'assurance savent (dans une certaine mesure) s'ils sont des citrons ou des pêches. Si la compagnie d'assurance pouvait faire la différence entre les citrons et les pêches, elle devrait demander aux pêches une prime en lien avec le risque des pêches et aux citrons une prime en lien avec le risque des citrons, conformément à un concept d'équité actuarielle, comme le rappelle Baker 2011. Mais si les actuaires ne sont pas en mesure de faire la différence entre les citrons et les pêches, alors ils devront demander le même prix pour un contrat d'assurance. La principale différence entre le marché décrit par Akerlof 1970 (il s'agissait dans la fable initiale d'un marché de l'occasion de véhicules automobiles) et un marché d'assurance est que l'asymétrie d'information était initialement (dans l'exemple des véhicules automobiles) en faveur du vendeur d'un bien. En assurance, la situation est souvent plus complexe. En assurance automobile, Dalziel et Job 1997 soulignaient le biais d'optimisme de la plupart des conducteurs qui pensent tous être des "bons risques". On retrouvera le même biais dans de nombreux autres exemples, comme le mentionnent Royal et Walls 2019, en écartant toutefois l'assurance santé, où l'assuré peut effectivement avoir davantage d'information que l'assureur.

Pour reprendre la description faite par Chassagnon 1996, supposons qu'un assureur couvre "*un grand nombre d'agents hétérogènes dans leurs probabilités de subir un dommage. Il propose un prix unique qui reflète la probabilité moyenne de perte de l'agent représentatif de cette économie, et il devient inintéressant pour les agents dont la probabilité de subir un accident est faible de s'assurer. Il s'opère donc un phénomène de sélection par les prix et on dit qu'elle est adverse parce que ce sont les mauvais agents qui demeurent*". Pour se prémunir contre ce phénomène d'anti-sélection, la sélection des risques et la segmentation des primes sont nécessaires. "*L'anti-sélection disparaît lorsque l'analyse des risques devient suffisamment performante pour que les marchés soient segmentés efficacement*, affirmait Picard 2003, la difficulté qu'ont les économètres à mettre en évidence de réelles situations d'anti-sélection sur le marché de l'assurance automobile ne reflète-t-elle pas l'appréciation sans cesse plus précise des risques souscrits par les assureurs ?

1.2.3 Les primes et les garanties

Comparer des assurés est toujours un exercice délicat car, non seulement ils ont potentiellement des risques différents, mais les assurés peuvent aussi avoir des préférences différentes (et donc peuvent choisir des contrats différents). Tout d'abord, il importe de distinguer les garanties. En assurance automobile, la garantie "responsabilité civile" correspond à la composante obligatoire, couvrant exclusivement les dommages que la voiture assurée pourrait causer à un tiers. Mais certains assurés pourraient avoir envie (ou besoin) de davantage de protection. Parmi les autres garanties classiques, on retrouvera la garantie "dommages tous risques" prenant en charge tous les dommages subis par le véhicule (quelles que soient les circonstances de l'accident ou la responsabilité du conducteur), la garantie "dommages collision" qui rembourse les dégâts causés au véhicule en cas de collision avec un tiers, ou la garantie "incendie et vol" qui indemnise le propriétaire du véhicule lorsque celui-ci est endommagé ou détruit par le feu, ou lorsqu'il est volé.

Certains assureurs proposent aussi une garantie “panne mécanique”, qui permet l’indemnisation par l’assurance des frais de réparation liés à une panne, ou une garantie “contenu du véhicule” qui propose une indemnisation en cas de dégradation ou de disparition d’objets présents à l’intérieur du véhicule assuré. Il peut également exister une garantie “assistance”, qui assure des prestations en cas de panne, comme le dépannage, le remorquage, le rapatriement, etc. Une autre source de différence possible est l’indemnité qui peut varier en fonction du choix du niveau de franchise. Pour rappel, la franchise est la somme qui reste à la charge du souscripteur après l’indemnisation d’un sinistre par l’assureur. La franchise absolue (ou fixe) est la franchise la plus courante de l’assurance auto : dans un contrat avec une franchise de 150€, si les frais de réparation s’élèvent à 250€, l’assurance prendra en charge 100€ et les 150€ restants seront à la charge du souscripteur. De nombreux assureurs proposent aujourd’hui des “franchises kilométriques”, définissant un périmètre autour du lieu de stationnement habituel du véhicule : à l’intérieur de ce périmètre, la garantie assistance ne fonctionnera pas. En revanche, si une panne survient en dehors de ce périmètre, la garantie assistance pourra être invoquée.

On peut schématiser cela sur la Figure 1.1 : (1) un souscripteur (caractérisé par $(\boldsymbol{x}, p)$) peut choisir, entre plusieurs garanties ou plusieurs niveaux de franchises (et les choix peuvent être dépendants de la variable protégée p), (2) les primes pures (fonction de la probabilité d’avoir un accident, du coût moyen) peuvent dépendre de la variable protégée p, et finalement (3) les primes commerciales peuvent aussi être fonction de la variable protégée p.

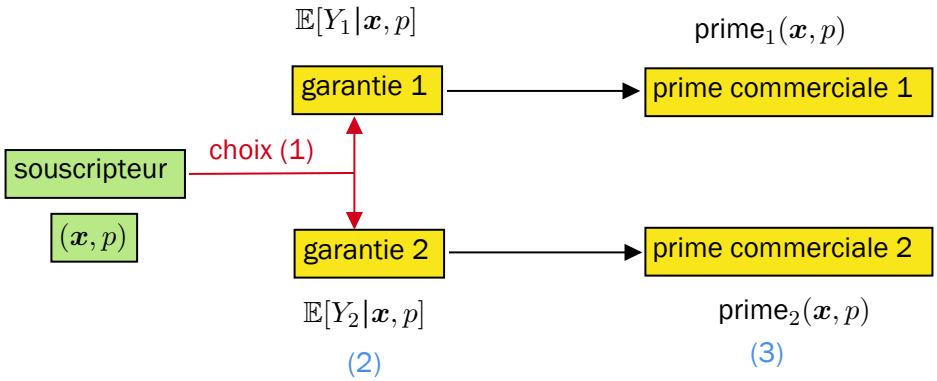


Figure 1.1 – Schéma présentant les différents moments dans l’achat d’une police d’assurance, en lien avec les caractéristiques de l’assuré.

Aussi, il est difficile de comparer la prime d’assurance automobile payée par des personnes différentes. Sur la Figure 1.2 on peut voir que le choix de la garantie en assurance automobile (étape (1) de la Figure 1.1) dépend fortement de l’âge, les jeunes conducteurs optant massivement pour la garantie obligatoire (un tiers des conducteurs entre 20 et 25 ans), les conducteurs plus âgés prenant davantage l’assurance “tous risques” (90 % des conducteurs entre 70 et 80 ans). Choisir des garanties différentes se traduit forcément sur la facture, les personnes plus âgées pouvant avoir une police d’assurance plus chère simplement parce qu’ils demandent davantage de couverture.

1.2.4 Cas d’une population hétérogène

Pour revenir à notre exemple introductif, si T_x est l’âge au décès (aléatoire) de l’assuré d’âge x à la souscription du contrat (ou dit autrement, $T_x - x$ est la durée de vie résiduelle), alors la prime pure

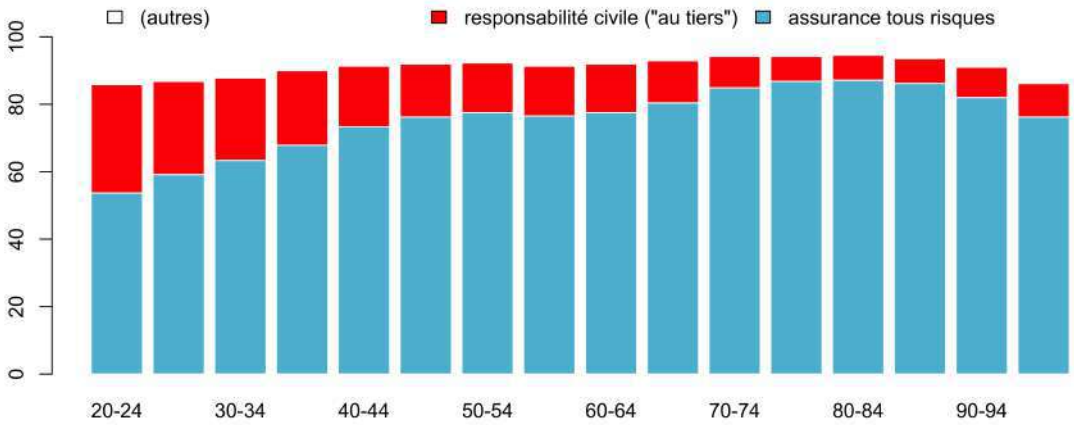


Figure 1.2 – Couvertures retenues par des souscripteurs en assurance automobile en fonction de l'âge, avec la couverture obligatoire de base, ■ "assurance au tiers" et la couverture la plus large ■ dite "tous risques" (source : compagnie d'assurance en France).

correspond à la valeur actuelle probable des flux futurs, soit

$$a_x = \mathbb{E}[\frac{100}{(1+r)^{T_x - x}}] = \sum_{t=0}^{\infty} \frac{100}{(1+r)^t} \cdot \mathbb{P}[T_x = x+t],$$

pour un taux d'escompte r, ce qui s'écrit, avec une terminologie plus statistique

$$a_x = \sum_{t=0}^{\infty} \frac{100}{(1+r)^t} \cdot \frac{L_{x+t-1} - L_{x+t}}{L_{x+t-1}},$$

où L_t est le nombre de personnes en vie au bout de t années dans une cohorte que l'on suivrait, de telle sorte que $L_{x+t-1} - L_{x+t}$ est le nombre de personnes en vie au bout de $x+t-1$ années mais pas de $x+t$ années (et donc décédées dans leur t-ième année). C'est De Witt 1671 qui, le premier, a proposé cette prime pour une assurance décès, où discriminer suivant l'âge semble légitime.

Mais on peut aller plus loin, car $\mathbb{P}[T_x = t]$, la probabilité que l'assuré d'âge x à la souscription décède dans t années, pourrait dépendre aussi de son genre, de son historique de santé, et probablement d'autres variables que l'assureur pourrait connaître. Et dans ce cas, il convient de calculer des probabilités conditionnelles, $\mathbb{P}[T_x = t|\text{femme}]$ ou $\mathbb{P}[T_x = t|\text{homme fumeur}]$.

1.2.5 Tables de mortalité et assurance-vie

Aussi étonnant que cela puisse paraître, Pradier 2011 notait qu'avant la fin du XVIII$^{\text{ème}}$siècle, en France, le prix des rentes viagères ne dépendait pratiquement jamais du sexe du souscripteur. Pourtant, les premières tables de mortalités séparées, entre les hommes et les femmes, constituées dès 1740 par Nicolas Struyck (publiées en annexes d'un article de géographie, Struyck 1740) montraient que les femmes vivaient généralement plus longtemps que les hommes (Tableau 1.1).

Struyck 1740 (publié en français dans Struyck 1912) montre qu'à 20 ans, l'espérance de vie (résiduelle) est de 30 ans $3/4$ pour les hommes et 35 ans $1/2$ pour les femmes. Il propose aussi des tables de rentes viagères par sexe. Pour une femme de 50 ans, une rente viagère valait 969 florins, contre 809 florins pour un homme du même âge. Cette différence substantielle semblait légitimer une différenciation des primes.

hommes						femmes					
x	L_x	${}_5p_x$	x	L_x	${}_5p_x$	x	L_x	${}_5p_x$	x	L_x	${}_5p_x$
0	1000	29.0 %	45	371	16.6 %	0	1000	28.9 %	45	423	11.8 %
5	710	5.6 %	50	313	19.2 %	5	711	5.2 %	50	373	14.7 %
10	670	4.2 %	55	253	22.9 %	10	674	3.3 %	55	318	18.2 %
15	642	5.5 %	60	195	27.2 %	15	652	4.3 %	60	260	21.2 %
20	607	6.6 %	65	142	31.7 %	20	624	5.8 %	65	205	26.8 %
25	567	7.9 %	70	97	37.1 %	25	588	6.8 %	70	150	33.3 %
30	522	9.2 %	75	61	45.9 %	30	548	7.3 %	75	100	45.0 %
35	474	10.5 %	80	33	51.5 %	35	508	7.9 %	80	55	56.4 %
40	424	12.5 %	85	16		40	468	9.6 %	85	24	

Table 1.1 – Extrait des tables Hommes et Femmes en 1720 (source : Struyck 1912, page 231), pour des pseudo-cohortes de 1000 personnes ($L_0 = 1000$). Ici, 424 hommes (L_x) et 468 femmes (sur 1000 naissances respectives) avaient alors atteint 40 ans ($x = 40$). Et parmi les personnes qui ont atteint 40 ans, 12.5 % des hommes et 9.6 % des femmes décéderont dans les 5 ans qui suivent (${}_5p_x = \mathbb{P}[T \leq x + 5 | T > x]$).

Selon Pradier 2012, il a fallu attendre que la caisse des veuves du duché de Calenberg fasse faillite, en 1779, pour que l'âge et le sexe des souscripteurs soient utilisés conjointement pour calculer le prix des rentes. En France, en 1984, les autorités réglementaires des marchés d'assurance avaient décidé de retenir des tables réglementaires établies pour la population générale par l'INSEE, basées sur la population observée sur 4 années, à savoir les tables PM 73-77 pour les hommes et PF 73-77 pour les femmes, renommées tables TD et TV 73-77, respectivement (avec un prolongement analytique au-delà de 99 ans). Si le premier facteur de la mortalité est l'âge, le genre est aussi un facteur important, comme le montre la Table 1.2. Depuis plus d'un siècle, la mortalité des hommes est plus élevée que celle des femmes, en France.

En pratique néanmoins, les tarifications actuarielles des contrats d'assurance-vie ont continué à être établies sans tenir compte du sexe des assurés. En fait, si les deux tables ont été utilisées, c'est parce que la table masculine était la table réglementaire pour les assurances *en cas de décès* (PM devient alors TD, pour *table de décès*), et la table féminine devint la table pour les assurances en cas de vie (PF devient alors TV, pour *table de vie*). En 1993, les tables TD et TV 88-90 ont remplacé les deux tables précédentes, avec le même principe, à savoir l'utilisation d'une table construite sur une population masculine pour les assurances en cas de décès, et une table construite sur un population féminine pour les assurances en cas de vie. D'un point de vue prudentiel, la table des femmes modélise une population ayant, en moyenne, une mortalité moindre, et donc vivant plus longtemps.

En 2005, les tables TH et TF 00-02 ont été retenues comme tables réglementaires, avec toujours

TD 73-77		TV 73-77		TD 88-90		TV 88-90		INED Hommes		INED Femmes	
0	100000	0	100000	0	100000	0	100000	0	100000	0	100000
10	97961	10	98447	10	98835	10	99129	10	99486	10	99578
20	97105	20	98055	20	98277	20	98869	20	99281	20	99471
30	95559	30	97439	30	96759	30	98371	30	98656	30	99247
40	93516	40	96419	40	94746	40	97534	40	97661	40	98810
50	88380	50	94056	50	90778	50	95752	50	95497	50	97645
60	77772	60	89106	60	81884	60	92050	60	90104	60	94777
70	57981	70	78659	70	65649	70	84440	70	78947	70	89145
80	28364	80	52974	80	39041	80	65043	80	59879	80	77161
90	4986	90	14743	90	9389	90	24739	90	25123	90	44236
100	103	100	531	100	263	100	1479	100	1412	100	4874
110	0	110	0	110	0	110	2				

Table 1.2 – Extrait des tables TD et TV 73-77 à gauche, TD et TV 88-90 au centre, et INED 2017-2019 à droite.

des tables construites sur des populations différentes, à savoir respectivement des hommes et des femmes. Mais cette fois, le terme hommes (H) et femmes (F) est maintenu dans le nom, car la réglementation autorisait la possibilité d'une tarification différente pour les hommes et les femmes. Un arrêt de la Cour de Justice de l'Union Européenne, en date du 1er mars 2011, a toutefois rendu impossible une tarification différenciée en fonction du genre (à compter du 21 décembre 2012), au motif qu'elles constitueraient des discriminations. À titre de comparaison, les tables récentes de l'INED [6] sont également mentionnées dans la Table 1.2, à droite.

Au-delà du sexe, toutes sortes de "variables discriminantes" ont été étudiées, afin de construire, par exemple, des tables de mortalité fonction du fait que la personne est fumeuse, ou pas, comme dans Benjamin et Michaelson 1988 (Table 1.3). En effet, depuis Hoffman 1931, ou Johnston 1945, les actuaires avaient observé que l'exposition au tabac, et le fait de fumer, avait un impact important sur la santé des assurés. Comme l'écrivait Miller et Gerstein 1983, "*it is clear that smoking is an important cause of mortality*".

Il existe aussi des tables de mortalité (ou des calculs d'espérance de vie résiduelle) par niveau d'indice de masse corporelle (IMC, ou BMI *Body Mass Index*, introduit par Adolphe Quetelet au milieu du XIX^{ème}siècle), telles que calculées par Steensma *et al.* 2013 au Canada. Un indice "normal" désigne les personnes ayant un indice compris entre 18.5 et 25 kg/m^2 ; le "surpoids" correspond à un indice entre 25 et 30 kg/m^2 ; l'obésité de niveau I un indice entre 30 et 35 kg/m^2, et l'obésité de niveau II un indice dépassant 35 kg/m^2. La Table 1.4 reprend quelques éléments. Ces ordres de grandeurs sont comparables avec K. R. Fontaine *et al.* 2003 parmi les études pionnières, Finkelstein *et al.* 2010, ou plus récemment Stenholm *et al.* 2017

6. https://www.ined.fr/fr/tout-savoir-population/chiffres/france/mortalite-cause-deces/table-mortalite/

	Hommes			Femmes	
	non-fumeur	fumeur		non-fumeur	fumeur
25	48.4	42.8	25	52.8	49.8
35	38.7	33.3	35	43.0	40.1
45	29.2	24.2	45	33.5	31.0
55	20.3	16.5	55	24.5	22.6
65	12.8	10.4	65	16.2	15.1

Table 1.3 – Espérance de vie résiduelle (en années), en fonction de l'âge (entre 25 et 65 ans) pour les fumeurs et non-fumeurs (source : Benjamin et Michaelson 1988, pour des données datant de 1970-1975 aux États-Unis).

	Hommes				Femmes			
	normal	surpoids	obèse I	obèse II	normal	surpoids	obèse I	obèse II
20	57.2	61.0	59.1	53.5	62.8	66.5	64.6	59.3
30	47.6	51.4	49.4	44.1	53.0	56.7	54.8	49.5
40	38.1	41.7	39.9	34.7	43.3	46.9	45.0	39.9
50	28.9	32.4	30.6	25.8	33.8	37.3	35.5	30.6
60	20.4	23.6	21.9	17.6	24.9	28.1	26.4	21.9
70	13.2	15.8	14.4	10.9	16.8	19.7	18.2	14.3

Table 1.4 – Espérance de vie résiduelle (en années), en fonction de l'âge (entre 20 et 70 ans) en fonction du niveau d'IMC (source : Steensma *et al.* 2013).

1.2.6 Des variables aux groupes

Les exemples précédents permettent surtout de voir qu'il existe une multitude de critères qui peuvent être utilisés pour créer des classes tarifaires. Comme dans les modèles de régression standard, on se contente ici de chercher des variables significativement corrélées avec la variable d'intérêt. Bailey et Simon 1960 proposaient ainsi de tenir compte des informations suivantes, en assurance automobile : l'usage (loisir - "*pleasure*" ou professionnel - "*business*"), l'âge (moins de 25 ans, ou pas), le genre, et le statut marital (marié ou non-marié). Plus précisément, cinq classes de risques sont considérées, avec des surcharges tarifaires par rapport à la première classe (qui sert ici de référence)

- "*pleasure, no male operator under 25*" (référence)
- "*pleasure, non-principal male operator under 25*", $+65$ %
- "*business use*", $+65$ %
- "*married owner or principal operator under 25*", $+65$ %
- "*unmarried owner or principal operator under 25*", $+140$ %

Dans les années 60, les classes tarifaires ressemblaient à celles qui seraient obtenues par des arbres de classification (ou de régression) comme ceux introduits par Breiman 1984. Mais en utilisant des algorithmes plus avancés, Davenport 2006 note que quand un actuaire crée des

classes de risques et des groupes tarifaires, et dans la plupart des cas, ces “groupes” ne sont pas conscients d’eux-mêmes, ils ne sont pas réfléchis (à la rigueur, l’actuaire essayera de les décrire en regardant les moyennes des différentes variables). Ces groupes, ou classes de risques, sont construits sur la base des données disponibles, et existent principalement en tant que produit de modèles actuariels. Et comme le souligne Gandy 2016, il n’existe pas de base matérielle permettant aux membres du groupe d’identifier d’autres personnes se trouvant dans “leur groupe”. Ces groupes de risques, développés à un instant donné, crée une connivence passagère entre les assurés, qui changeront probablement de groupe en déménageant, en changeant de voiture, voire simplement en vieillissant.

1.2.7 Classification et proxys

Pour comprendre l’impact de la classification des risques sur les calculs de primes pures, formellement, la formule des probabilités totales garantie (Feller 1957 ou S. M. Ross 2014) que

$$\mathbb{P}(A) = \sum_{i\in I} \mathbb{P}(A \cap B_i) = \sum_{i\in I} \mathbb{P}(A|B_i) \cdot \mathbb{P}(B_i),$$

si $(B_i)_{i\in I}$ est un un ensemble exhaustif (fini ou dénombrable) d’évènements. Une conséquence immédiate est la formule des espérances totales, pour une variable aléatoire Y,

$$\mathbb{E}(Y) = \sum_{i\in I} \mathbb{E}(Y|B_i) \cdot \mathbb{P}(B_i).$$

Cette formule peut s’écrire simplement dans le cas où deux ensembles, deux sous-groupes, sont considérés, par exemple liés au genre de l’individu,

$$\mathbb{E}(Y) = \mathbb{E}(Y|\text{femme}) \cdot \mathbb{P}(\text{femme}) + \mathbb{E}(Y|\text{homme}) \cdot \mathbb{P}(\text{homme}).$$

Si Y désigne la durée de vie à la naissance d’un individu, la traduction littérale de l’expression précédente est que l’espérance de vie à la naissance d’un individu pris au hasard est une moyenne pondérée des espérances de vie à la naissance des femmes et des hommes, les pondérations étant les proportions d’hommes et de femmes dans la population. Et comme $\mathbb{E}(Y)$ est une moyenne des deux,

$$\min\{\mathbb{E}(Y|\text{femme}), \mathbb{E}(Y|\text{homme})\} \leq \mathbb{E}(Y) \leq \max\{\mathbb{E}(Y|\text{femme}), \mathbb{E}(Y|\text{homme})\},$$

dit autrement, traiter la population comme homogène, alors qu’elle ne l’est pas, revient à subventionner un groupe par l’autre, ce qui est appelé une inéquité actuarielle (“*actuarialy unfair*”, telle que discuté par Landes 2015, Frezal et Barry 2019 ou Heras *et al.* 2020). L’inéquité sera d’autant plus grande que l’écart entre les deux espérances conditionnelles est importante (on parlera formellement de “variance des espérances conditionnelles”, ou “variance inter-groupes”).

Le concept d’équité actuarielle, tel qu’il apparaît chez Borch 1962, ou “*actuarial fairness*”, repose sur une correspondance parfaite entre la valeur totale des primes collectées et le montant total des réclamations légitimes faites par les assurés. Et comme il est impossible pour l’assureur de savoir ce que seront réellement les demandes d’indemnisation futures, on dit qu’il est actuariellement juste de fixer le niveau des primes sur la base des antécédents de demandes d’indemnisation des personnes appartenant à la même classe de risque (supposée). C’est sur cette base que la discrimination est vue comme “équitable” en termes de distribution. Autrement, la redistribution

serait vue “injuste”, avec une solidarité forcée, du groupe de faible risque vers le groupe à haut risque. Cette “équité” a été mise à mal dans les années 80, quand des assureurs privés ont limité l'accès à l'assurance pour les personnes atteintes du SIDA, ou risquant de le développer, comme le rappelle Daniels 1990. Feiring 2009 va plus loin dans le contexte de l'information génétique, “*since the individual has no choice in selecting her genotype or the expression of it, it is unfair to hold her accountable for the consequences of the genes she inherits—as it is unfair to hold her accountable for the consequences of any distribution of factors that are a result of a natural lottery*”. À la fin des années 70, Boonekamp et Donaldson 1979, Kimball 1979 ou encore Maynard 1979, l'idée est que la proportionnalité entre la prime et le risque encouru garantirait l'équité entre assurés a commencé à se traduire à l'aide d'espérance conditionnelle (conditionnelle aux facteurs de risque retenus).

Un point important à mentionner ici est qu'on s'intéresse à une forme d'équité *ex-ante*. En effet, *ex-post*, une fois les sinistres survenus, la distribution des pertes est très concentrée, une proportion très faible du portefeuille représentant une part importante des pertes. En assurance santé, en France, 5 % de la population concentrant 50 % des dépenses (Com-Ruelle et Dumesnil 1999). En assurance dommage, moins de 10 % de la population est sinistrée chaque année, autrement dit, 10 % de la population représente 100 % des dépenses. Mais *a priori*, on ne sait pas qui sera sinistré. Le but de l'actuaire sera de distinguer ceux qui ont 5 % de chances[7] d'être sinistrés de ceux qui ont 15 % de chances de l'être (et qui devraient payer 3 fois plus que les autres, selon le principe d'équité actuarielle). Formellement, cette hétérogénéité sera modélisée par un facteur latent Θ. Si y désigne la survenance (ou pas) d'un accident, y est vue comme la réalisation d'une variable aléatoire Y qui suit une loi de Bernoulli, $\mathcal{B}(\Theta)$, où Θ est une variable latente non-observable (comme dans Gouriéroux 1999b ou Gouriéroux et Jasiak 2011). Si y désigne le nombre d'accidents survenus pendant l'année, Y suit une loi de Poisson, $\mathcal{P}(\Theta)$ (ou un modèle binomial-négative, à inflation de zéros, etc, comme dans M. Denuit *et al.* 2007). Si y désigne le coût annuel, Y suit une loi Tweedie, ou plus généralement une loi dite “Poisson-composée”, que l'on notera $\mathcal{L}(\Theta, \varphi)$, où $\mathcal{L}$ désigne une loi de moyenne Θ, et où φ est un paramètre de dispersion. Le but de la segmentation est de constituer des classes (B_i) de manière optimale, c'est-à-dire en veillant à ce qu'une classe ne subventionne pas l'autre, à partir de caractéristiques observables, notées $\boldsymbol{x} = \{x_1, x_2, \cdots, x_k\}$. Crocker et Snow 2013 parle de “*categorization based on immutable characteristics*”. Pour Gouriéroux 1999a, c'est la “partition statique” utilisée pour constituer des sous-groupes de risques homogènes (“*in a given class, the individual risks are independent, with identical distributions*”). C'est ce que fait un arbre de classification ou de régression, les (B_i) étaient les feuilles de l'arbre, avec les notations probabilistes précédentes. Si y désigne la survenance d'un accident, ou la charge annuelle (aléatoire), l'actuaire va chercher à approcher $\mathbb{E}[Y|\boldsymbol{X}]$, à partir de données d'apprentissage. Dans une approche économétrique, si y désigne la survenance (ou pas) d'un accident, et si $\boldsymbol{x}$ désigne l'ensemble des caractéristiques observables de l'assuré, $Y|\boldsymbol{X} = \boldsymbol{x}$ suit une loi de Bernoulli, $\mathcal{B}(p_{\boldsymbol{x}})$, par exemple

$$p_{\boldsymbol{x}} = \frac{\exp\big(\boldsymbol{x}^\top \boldsymbol{\beta}\big)}{1+\exp\big(\boldsymbol{x}^\top \boldsymbol{\beta}\big)} \text{ ou } p_{\boldsymbol{x}} = \Phi(\boldsymbol{x}^\top \boldsymbol{\beta}),$$

pour une régression logistique ou probit, respectivement[8]. Si Y désigne le nombre d'accidents survenus pendant l'année, $Y|\boldsymbol{X} = \boldsymbol{x}$ suit une loi de Poisson, $\mathcal{P}(\lambda_{\boldsymbol{x}})$, avec classiquement $\lambda_{\boldsymbol{x}} =$

7. On utilisera le terme probabiliste de “chances”, même si ici, il s'agirait plutôt d'une malchance.
8. Φ est ici la fonction de répartition de la loi normale centrée et réduite, $\mathcal{N}(0,1)$.

$e^{\boldsymbol{x}^\top \boldsymbol{\beta}}$. Si Y désigne le coût annuel, $Y|\boldsymbol{X}=\boldsymbol{x}$ suit une loi Tweedie, ou plus généralement une loi Poisson composée, $\mathcal{L}(\mu_{\boldsymbol{x}},\varphi)$, où $\mathcal{L}$ désigne une loi de moyenne μ, avec $\mu_{\boldsymbol{x}}=\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}]$ (pour plus de détails, Charpentier et Denuit 2004 et 2005).

Pour reprendre l'analyse de Denuit et Charpentier 2004, si on suppose les risques homogènes, la prime pure sera $\mathbb{E}[Y]$, et on a le tableau de partage des risques suivant :

	Souscripteur	Assureur
Perte	$\mathbb{E}[Y]$	$Y-\mathbb{E}[Y]$
Perte moyenne	$\mathbb{E}[Y]$	0
Variance	0	$\text{Var}[Y]$

À l'autre extrémité, si le facteur de risque latent Θ était observable, la prime pure demandée serait $\mathbb{E}[Y|\Theta]$, et on aurait le partage suivant :

	Souscripteur	Assureur
Perte	$\mathbb{E}[Y\|\Theta]$	$Y-\mathbb{E}[Y\|\Theta]$
Perte moyenne	$\mathbb{E}[Y]$	0
Variance	$\text{Var}[\mathbb{E}[Y\|\Theta]]$	$\text{Var}[Y-\mathbb{E}[Y\|\Theta]]$

Si l'assureur est, en moyenne, à l'équilibre financier, notons que $\text{Var}[Y-\mathbb{E}[Y|\Theta]]$ peut s'écrire $\mathbb{E}[\text{Var}[Y|\Theta]]$, de telle sorte que

$$\text{Var}[Y]=\underbrace{\mathbb{E}[\text{Var}[Y|\Theta]]}_{\rightarrow \text{ assureur}}+\underbrace{\text{Var}[\mathbb{E}[Y|\Theta]]}_{\rightarrow \text{ souscripteur}}.$$

En utilisant finalement des caractéristiques observables, $\boldsymbol{x}=\{x_1,x_2,\cdots,x_k\}$, on aurait la décomposition suivante :

	Souscripteur	Assureur
Perte	$\mathbb{E}[Y\|\boldsymbol{X}]$	$Y-\mathbb{E}[Y\|\boldsymbol{X}]$
Perte moyenne	$\mathbb{E}[Y]$	0
Variance	$\text{Var}[\mathbb{E}[Y\|\boldsymbol{X}]]$	$\mathbb{E}[\text{Var}[Y\|\boldsymbol{X}]]$

Avec là encore, en moyenne, un équilibre financier de l'assureur, et ici

$$\begin{aligned}\mathbb{E}[\text{Var}[Y|\boldsymbol{X}]] &= \mathbb{E}[\mathbb{E}[\text{Var}[Y|\Theta]|\boldsymbol{X}]]+\mathbb{E}[\text{Var}[\mathbb{E}[Y|\Theta]|\boldsymbol{X}]]\\ &= \underbrace{\mathbb{E}[\text{Var}[Y|\Theta]]}_{\text{tarification parfaite}}+\underbrace{\mathbb{E}\{\text{Var}[\mathbb{E}[Y|\Theta]|\boldsymbol{X}]\}}_{\text{mauvaise classification}}.\end{aligned}$$

Cette "mauvaise classification" (à droite) est appelée "*subsidierende solidariteit*" par De Pril et Dhaene 1996, ou "solidarité subsidiante", par opposition à la "*kanssolidariteit*" ou "solidarité aléatoire" (à gauche[9]). Pour Corlier 1998, la segmentation "*diminue la solidarité des risques appartenant à des segments différents*". Et pour Löffler *et al.* 2016, citant un rapport de McKinsey

9. Ce terme de "solidarité aléatoire" est fondamental, et contrairement à ce qu'affirme Zuboff 2019 ("*as certainty replaces uncertainty*"), l'incertitude ne disparaîtra jamais, en tous cas tant il s'agit de prédire ce qui peut survenir pendant une année à la date de souscription.

consacré à l'avenir de l'industrie de l'assurance, mentionnaient que les données massives conduiront inévitablement à la démutualisation et à une concentration accrue sur la prédiction. Le lien avec la solidarité est discuté dans Gollier 2002 qui rappelle que "*la solidarité, c'est fondamentalement réaliser des transferts au profits de personnes défavorisées, en provenance de personnes plus favorisées*". Mais une version très limitée de la solidarité est prise en compte dans le contexte de l'assurance : "*la solidarité en assurance, c'est décider de ne pas segmenter le marché du risque correspondant sur une base des caractéristiques observables des risques des individus*", comme en assurance santé ou en assurance chômage. Notons que si historiquement, les variables x étaient discrétisées pour construire des "classes tarifaires", il est aujourd'hui classique de considérer les variables continues en tant que telles, voire de les transformer, tout en conservant une relative régularité. Comme le note Bouk 2015, rapportant les propos d'Emory McClintock (mathématicien ayant travaillé en tant qu'actuaire pendant plus de trente ans, Fiske 1917), "*McClintock insisted that actuaries smoothed because smoothing was a "mathematical and ethical" good. McClintock defended his practices to Hughes on "moral" grounds*".

Encore une fois, la difficulté de la tarification est que ce facteur de risque sous-jacent Θ n'est pas observable. Ne pas le capturer engendrerait une inéquité, car cela reviendrait à faire subventionner indûment les personnes les "plus risquées" (susceptibles de présenter des demandes d'indemnisation plus coûteuses) par les "moins risquées". Baker et Simon 2002 allaient plus loin, affirmant que la raison pour laquelle certaines personnes sont classées comme étant "à faible risque" et d'autres comme étant "à haut risque" n'est pas pertinente. En parlant d'automatiser la responsabilité ("*automating accountability*"), Baker et Simon 2002 affirmaient qu'il était important de rendre les gens responsables du risque qu'ils apportent à la mutualité, en particulier les assurés les plus risqués, afin que les assurés les moins risqués puissent "se sentir moralement à l'aise" ("*feel morally comfortable*" comme le disait Stone 1993). Le danger est qu'ainsi, l'allocation des contributions de chacun dans la mutualité serait le résultat d'un calcul actuariel "impartial", comme l'affirmait Stone 1993. Porter 2020 disait que ce processus était "*a way of making decisions without seeming to decide*". Nous reviendrons sur ce point lorsque nous parleront des exclusions, et de l'interprétabilité des modèles. L'assureur va alors utiliser des "*proxys*" pour capturer cette hétérogénéité, comme nous venons de le voir. Un "*proxy*" (on pourrait parler de "variable de substitution") est une variable qui n'est pas significative en soi, mais qui remplace une variable utile mais non observable, ou non mesurable, selon Upton et Cook 2014.

L'essentiel de notre discussion portera sur la discrimination tarifaire, et plus précisément le tarif "technique". Comme évoqué en introduction, du point de vue du souscripteur, ce n'est pas la grandeur la plus pertinente. En effet, à la prime actuarielle (la prime pure évoquée auparavant) s'ajoute une partie commerciale, un agent d'assurance pouvant décider d'offrir un rabais à un assuré ou à un autre, tenant compte d'une aversion pour le risque différente, ou d'une élasticité aux prix plus ou moins grande. Mais une question importante sous-jacente reste "le service offert est-il le même?". Ingold et Soper 2106 reviennent sur l'exemple d'Amazon qui n'offre pas les mêmes services à tous ses clients, en particulier les offres de livraison le jour même ("*same-day-delivery*"), offertes dans certains quartiers, choisis par un algorithme qui a finalement renforcé les préjugés raciaux (en ne proposant jamais de livraison le jour même dans les quartiers composés principalement de groupes minoritaires). Une lecture naïve des prix sur Amazon serait faussée par ce biais important dans les données, dont il convient de tenir compte. Comme le rappellent Calders et Žliobaitė 2013, "*unbiased computational processes can lead to discriminative decision procedures*". En assurance, on pourrait imaginer qu'un gestionnaire de sinistres n'offre pas la même indemnité à des personnes ayant des profils différents, certaines personnes ayant moins tendance à

contester que d'autres. Mieux comprendre l'articulation entre les différents concepts est important.

1.2.8 Interprétabilité et explicabilité

Une grande partie du travail de l'actuaire est de motiver, d'expliquer, une classification. Pasquale 2015b, Castelvecchi 2016 ou Kitchin 2017 ont souligné que les algorithmes d'apprentissage machine sont caractérisés par leur opacité et leur "incompréhensibilité", parfois appelé propriétés de "boîte noire". Et il est indispensable de les expliquer, de raconter une histoire. Pour Rubinstein 2012, les modèles sont des "*fables*" : "*in economic theory, as in Harry Potter, the Emperor's New Clothes or the tales of King Solomon, we amuse ourselves in imaginary worlds. Economic theory spins tales and calls them models. An economic model is also somewhere between fantasy and reality* $(\cdots)$ *the word model sounds more scientific than the word fable or tale, but I think we are talking about the same thing*". De la même la manière, l'actuaire devra raconter son modèle, avant de convaincre la souscription et les agents d'assurance de l'adopter. Mais cette narration est forcément imprécise.

On peut entendre que l'âge doit intervenir dans la prédiction de la fréquence de sinistres en assurance automobile, et effectivement, comme on le voit sur la Figure 1.3, la prédiction ne sera pas la même à 18 ans, 25 ans ou 55 ans. Assez naturellement, on peut rendre légitime une surprime pour des jeunes conducteurs, à cause d'une faible expérience de conduite, associée à des réflexes non-acquis. Mais cette histoire ne nous dit pas l'ordre de grandeur de cette surcharge qui semblerait légitime. Si on va plus loin, le choix du modèle est loin d'être neutre sur la prédiction : pour un assuré de 22 ans, des modèles relativement simples proposent une surprime de 27 %, 73 %, 82 % ou 110 % (par rapport à la prime moyenne sur l'ensemble de la population). Si la discrimination par l'âge pourrait sembler avoir du sens, à quel point peut-on s'autoriser ici à discriminer ?

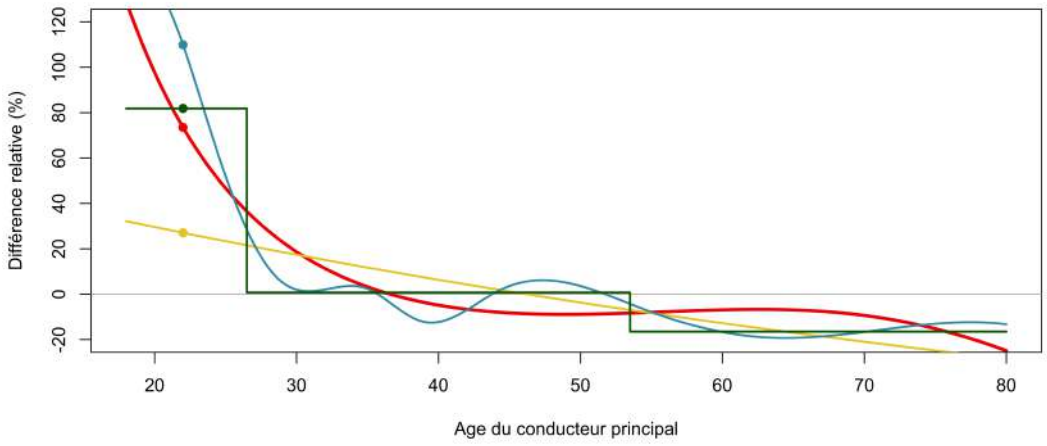


Figure 1.3 – Évolution de la fréquence de sinistres en assurance automobile, en fonction de l'âge du conducteur principal, relativement à la fréquence annuelle globale, avec une régression de Poisson en jaune, une régression lissée en rouge, une régression lissée avec une fenêtre *trop* petite en bleu, et avec un arbre de régression en vert. Les points sont les prédictions pour un conducteur de 22 ans (source : Charpentier 2014).

1.2.9 Supprimer la discrimination

Lindholm *et al.* 2021 présentent un exemple permettant de comprendre les difficultés pour supprimer une possible discrimination. Dans la Table 1.5, on dispose de deux statistiques de sinistralité basées sur deux variables, une variable protégée (notée p), le genre, et une variable autorisée (notée x) indiquant si la personne est fumeuse, ou pas.

nombre n	femmes	hommes	total
fumeur	32	4	36
non-fumeur	28	48	76
total	60	52	112

exposition e	femmes	hommes	total
fumeur	133	24	157
non-fumeur	131	301	432
total	264	325	589

Table 1.5 – Nombre de sinistres (à gauche) et exposition (à droite), en fonction du genre (variable protégée p) et du fait de fumer, ou pas (x). (source : Lindholm *et al.* 2021)

La fréquence annuelle de sinistre Y, par année, se déduit des tableaux de contingence de la Table 1.5, soit ici

$$\mathbb{E}[Y] = \frac{n}{e} = \frac{112}{589} \approx 19.0\,\%,$$

et si on segmente suivant le tabagisme

$$\mathbb{E}[Y|X = \text{fumeur}] = \frac{36}{157} \approx 22.9\,\% \text{ et } \mathbb{E}[Y|X = \text{non-fumeur}] = \frac{76}{432} \approx 17.6\,\%.$$

Notons que l'on retrouve la formule des espérances totales évoquée auparavant,

$$\mathbb{E}[Y] = \mathbb{E}[Y|X = \text{fumeur}] \cdot \mathbb{P}[X = \text{fumeur}] + \mathbb{E}[Y|X = \text{non-fumeur}] \cdot \mathbb{P}[X = \text{non-fumeur}].$$

On le voit sur cet exemple, le modèle tarifaire basé sur le tabagisme (et pas le genre) n'est pas indépendant du genre pour autant. En effet, si la prime est proportionnelle à $\mathbb{E}[Y|X = x]$, les primes moyennes des hommes et des femmes sont proportionnelles à

$$\begin{cases} \text{femmes} & : \dfrac{133}{264} \cdot \mathbb{E}[Y|X = \text{fumeur}] + \dfrac{131}{264} \cdot \mathbb{E}[Y|X = \text{non-fumeur}] \approx 20.3\,\% \\ \text{hommes} & : \dfrac{24}{325} \cdot \mathbb{E}[Y|X = \text{fumeur}] + \dfrac{301}{325} \cdot \mathbb{E}[Y|X = \text{non-fumeur}] \approx 18.0\,\%. \end{cases}$$

Autrement dit, la prime des hommes est, en moyenne, plus élevée que celle des femmes. Lindholm *et al.* 2021 proposent une méthode relativement simple pour proposer une prime dite "*discrimination-free*", en proposant d'utiliser

$$\text{prime}(x) = \sum_p \mathbb{E}[Y|X = x, P = p] \cdot \mathbb{P}[P = p],$$

soit ici

$$\begin{cases} \text{prime(fumeur)} = \dfrac{32}{133} \cdot \dfrac{264}{584} + \dfrac{4}{24} \cdot \dfrac{325}{584} \approx 20.0\,\% \quad (< \mathbb{E}[Y|X = \text{fumeur}] \approx 22.9\,\%) \\ \text{prime(non-fumeur)} = \dfrac{28}{131} \cdot \dfrac{264}{584} + \dfrac{48}{301} \cdot \dfrac{325}{584} \approx 18.4\,\% \quad (> \mathbb{E}[Y|X = \text{non-fumeur}] \approx 17.6\,\%) \end{cases}$$

Cette technique est à rapprocher des notions de "*partial dependence plot*", introduites par Friedman 2001, et utilisées pour interpréter et expliquer des modèles "boite noire", en apprentissage automatique. Notons que dans ce cas, on va globalement sous-tarifer

$$\begin{cases} \text{total :} & : \frac{157}{589} \cdot \text{prime(fumeur)} + \frac{432}{589} \cdot \text{prime(non-fumeur)} \approx 18.8\,\% \quad (< \mathbb{E}[Y] \approx 19.0\%) \\ \text{femmes} & : \frac{133}{264} \cdot \text{prime(fumeur)} + \frac{131}{264} \cdot \text{prime(non-fumeur)} \approx 19.2\,\% \\ \text{hommes} & : \frac{24}{325} \cdot \text{prime(fumeur)} + \frac{301}{325} \cdot \text{prime(non-fumeur)} \approx 18.4\,\%. \end{cases}$$

Cette méthode se rapproche des approches de projection orthogonales, proposées par Frisch-Waugh-Lovel (Frisch et Waugh 1933, et Lovell 1963), qui consiste à rendre x et y orthogonales à p (simplement en prenant les résidus de la régression linéaire de y et x sur p). Nous discuterons davantage les techniques permettant de supprimer les discriminations dans la section 4.

Si on modifie maintenant cet exemple, imagions deux scénarios (tout aussi fictifs que le précédant), où x serait une variable soit difficile à observer, soit observée potentiellement *ex-post*,

- y est toujours un nombre d'accidents en assurance automobile, mais au lieu de $x \in \{\text{fumeur}, \text{non-fumeur}\}$, on ait $x \in \{\text{conduite dangereuse}, \text{conduite non-dangereuse}\}$ ou $x \in \{\text{non respect du code de la route}, \text{respect du code de la route}\}$ (qui sont non-observables *ex-ante*, lors de la signature du contrat d'assurance, mais qui pourraient l'être après, sur la base de boîtiers connectés). Ne serait-il pas pertinent de demander des primes différentes sur la base du genre car cette variable est un proxy d'une variable tarifaire pertinente ?
- y est ici un nombre de jours d'hospitalisation, et la variable x prend ses valeurs dans $\{\text{prédisposition à certaines maladies}, \text{non-prédisposition}\}$ (qui pourraient être observées sur la base de séquençage génomique, ou d'antécédents médicaux). Serait-ce (moralement) acceptable d'utiliser le genre de la personne pour proposer des tarifs différents sur la base d'un risque sous-jacent qui serait fondamentalement différent ?

C'est autour de ces interrogations que nous allons revenir ici, mais avant, il est probablement important de définir certaines notions qui seront discutées par la suite.

1.3 Données, Modèles, biais, discrimination et équité

1.3.1 Données, données personnelles et données sensibles

Comme le disait Debet 2007, "*pour lutter contre les discriminations, encore faut-il pouvoir les identifier ; pour les identifier, il paraît naturel de procéder à l'observation statistique des différences, de la diversité*". Aussi, il va falloir collecter des **données** (ou **data** en anglais, du latin *datum* [10]). Et bon nombre d'informations collectées sont parfois jugées "sensibles", ou "protégées".

Une **donnée à caractère personnel**, ou "donnée personnelle", correspond en droit français à toute information relative à une personne [11] physique identifiée ou qui peut être identifiée, directement ou indirectement. La définition d'une donnée personnelle est précisée à l'article 4 du RGPD (Règlement général sur la protection des données). Ces informations peuvent être un identifiant (un nom, un numéro d'identification, des données de localisation, par exemple) ou alors un

10. Un traité de géométrie plane d'Euclide s'intitule δεδομένα, "données".
11. Elles concernent les personnes physiques et vivantes.

ou des éléments spécifiques propres à l'identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale de la personne. Parmi la liste (non exhaustive) donnée par la CNIL, il peut y avoir le nom, prénom, numéro de téléphone, plaque d'immatriculation, numéro de sécurité sociale, adresse postale, mail, un enregistrement vocal, une photographie, etc.

Les **données sensibles** regroupent les croyances religieuses, l'orientation sexuelle, l'engagement syndical, l'appartenance ethnique, la situation médicale, les condamnations et infractions pénales, les données biométriques, les informations génétiques ou encore les activités sexuelles. Selon le RGPD, en 2016, "*le traitement des données à caractère personnel qui révèle l'origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique sont interdits*". Ces informations seront considérées comme "sensibles". En Europe, la "Convention 108" (ou *convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel*) de 2018 précise davantage les contours. Nous reviendrons sur les variables sensibles, dans le contexte de l'assurance dans le Chapitre 2.

Une autre précaution qu'il convient d'avoir en tête est liée à la distinction entre ce "*qui révèle*" ou "*qui est susceptible de révéler*", comme le dit Debet 2007. Certaines informations sont déclarées par les personnes, et d'autres sont des **données inférées**. Par exemple il serait possible de demander le "sexe à la naissance", mais dans certains cas, une variable est construite à partir d'une question sur le titres de civilité (où "madame" ou "monsieur" sont proposés). Mais il peut s'agir de modèles plus complexes, pour inférer une information. On peut imaginer que "*être enceinte*" puisse être une information sensible dans de nombreuses situations. Cette information existe dans certaines bases de données d'organismes de santé (ou de remboursements de soins). Mais comme l'avait montré Duhigg 2012 (dans "*how companies learn your secrets* "), il existe des organismes qui tentent d'inférer cette information, à partir d'achats. C'est la fameuse histoire de cet homme, dans les environs de Minneapolis, qui avait été surpris que des bons de réductions pour divers produits destinés aux jeunes mamans soient adressés à sa fille. Dans cette histoire, l'inférence par le modèle avait été correcte. Récemment, Lovejoy 2021 rappelle qu'en juin 2020, LinkedIn avait eu une brèche massive (exposant les données de 700 millions d'utilisateurs), avec une base de données d'enregistrements comprenant des numéros de téléphone, des adresses physiques, des données de géolocalisation et… des salaires induits ("*inferred salaries*"). Les données télématiques permettent d'inférer beaucoup d'information, avec toutefois une incertitude plus ou moins grande. Comme l'évoquait C.-S. Bigot et Charpentier 2019, en observant qu'une personne stationne presque tous les vendredis matins à proximité d'une mosquée, on pourrait affirmer qu'il y a de grandes chances qu'elle soit musulmane (sur la base d'enquêtes sur les pratiques des musulmans). Mais il est possible que cette inférence soit complètement fausse, et que cette personne va, en réalité au club de gym, en face de la mosquée, et en plus, elle y est assidue.

La mode du **big data**, des **données massives**, a été l'occasion d'évoquer leur volume important, leur valeur, mais aussi leur variété (et toutes sortes de mots commençant par la lettre 'v'). Si pour les actuaires, les données ont souvent été des données tabulaires, des matrices de chiffres, depuis quelques années, la variété des formes des données s'est imposée. On va retrouver naturellement du texte, à commencer par un nom, un prénom, une adresse (qui pourra être convertie en coordonnées spatiales), des noms de médicaments, des conversations téléphoniques avec un souscripteur ou une gestionnaire de sinistres, ou pour des entreprises, des contrats dont

les clauses ont été numérisées, etc. On peut avoir des images, comme une photo de l'automobile après un accrochage ou d'un toit de maison après un incendie, des images médicales (rayons X, de l'imagerie par résonance magnétique), une image satellite d'un champ pour un contrat d'assurance récolte, ou d'un village suite à une inondation, etc. On va finalement avoir aussi des informations associées à des objets connectés, des données obtenues à l'aide de boîtiers embarqués dans une flotte automobile, un détecteur de fuite d'eau ou des appareils destinés à la surveillance et au contrôle des cheminées. Mais bien souvent, des **scores**, ou des résumés statistiques, sont constitués, à partir de ces données brutes (auxquelles l'assureur n'a pas souvent accès). Ça sera le nombre de kilomètres parcourus une semaine donnée par un souscripteur d'un contrat d'assurance automobile, ou un score d'accélération. On le verra, bien souvent, ces données bien plus riches que les variables tabulaires avec des champs prédéfinis (certes, avec parfois des difficultés de définition, comme l'expliquait Desrosières 2016) peuvent apporter des informations sensibles qui pourraient être exploitées par un algorithme "boite noire", possiblement à l'insu de l'actuaire.

1.3.2 Modèle prédictif, algorithmes et "intelligence artificielle"

Pour Ekeland 1995, la modélisation, c'est "*la construction (intellectuelle) d'un modèle mathématique c'est-à-dire d'un réseau d'équations censé décrire la réalité*". Et bien souvent, un **modèle** c'est aussi (surtout) une simplification de cette réalité. Un modèle trop complexe n'est pas un bon modèle. C'est l'idée de sur-apprentissage (ou "*overfit*") que l'on retrouve en statistique, ou le concept de parcimonie, parfois appelé "rasoir d'Okham" (comme sur la Figure 1.4) classique en économétrie [12]. Comme l'avait dit Milanković 1920, "*pour pouvoir traduire en langage mathématique les phénomènes de la nature, il est toujours nécessaire d'admettre des simplifications et de simplifier certaines influences et irrégularités*". Le modèle est une simplification du monde, ou, comme l'avait dit Korzybski 1958 dans un contexte de géographie, "*a map is not the territory it represents, but, if correct, it has a similar structure to the territory, which accounts for its usefulness*". La carte n'est pas le territoire : la carte correspond à la représentation que l'on se fait du monde, alors le territoire est le monde tel qu'il est réellement. On pensera naturellement à Borges 1946 (ou le pastiche d'Umberto Ecco *de l'impossibilité de construire la carte 1:1 de l'Empire*, dans Eco 1992), "*En aquel Imperio, el Arte de la Cartografía logró tal Perfección que el mapa de una sola Provincia ocupaba toda una Ciudad, y el mapa del Imperio, toda una Provincia. Con el tiempo, estos Mapas Desmesurados no satisficieron y los Colegios de Cartógrafos levantaron un Mapa del Imperio, que tenía el tamaño del Imperio y coincidía puntualmente con él.*".

À la notion de modèle se substituera de plus en plus le terme d'**algorithme**, voire d'**intelligence artificielle**, ou **IA** (en particulier dans la presse, France Info 2019 ou Le Monde 2021a). Pour Zafar *et al.* 2017, par "algorithme", il convient d'entendre des modèles prédictifs (règles de décision) calibrés à partir de données historiques grâce à l'exploration de données. Pour comprendre la différence, Cardon 2019 donne un exemple pour expliquer ce qu'est l'apprentissage machine. Il est assez simple d'écrire un programme qui convertit une température en degrés Celsius en une température donnée en degrés Fahrenheit. Pour cela, il existe une règle simple : soustraire 32 de la température en Celcius et multiplier le résultat par 5/9 (ou diviser par 1.8). Une approche en apprentissage machine (ou en intelligence artificielle) propose une solution toute différente : au lieu de coder la règle dans la machine (ce que les informaticiens pourraient appeler "*Good Old Fashioned Artificial Intelligence*", comme Haugeland 1989), on lui donne seulement des exemples de correspondances entre des températures en degrés Celsius et en degrés Fahrenheit. On entre

12. *pluralitas non est ponenda sine necessitate* (Guillaume d'Ockham, XIVe siècle), discuté par Bera 2001

les données dans une base d'apprentissage, et l'algorithme va apprendre "lui-même" une règle de conversion, en cherchant, parmi un exemple de fonctions candidates la plus proche des données. On peut alors tomber sur un exemple comme sur la Figure 1.4.

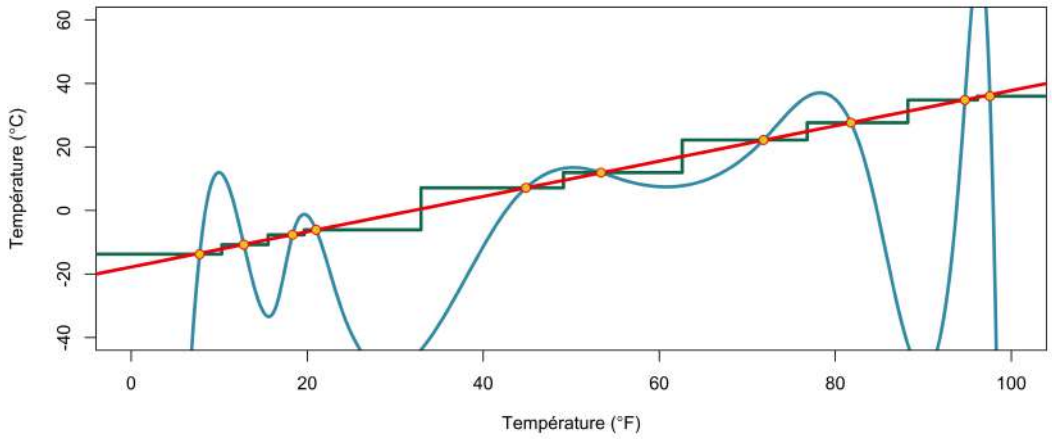


Figure 1.4 – Un modèle simple (linéaire), un modèle constant par morceau, ou un modèle complexe (non-linéaire mais continu), à partir d'une dizaine d'observations (x_i, y_i) où x est une température en degrés Fahrenheit et y la température en degrés Celsius, au même endroit.

Notons que la "complexité" de certains algorithmes, ou leur "opacité" (qui donnera la qualification de "**boîte noire**"), n'a pas à voir avec l'algorithme d'optimisation utilisé (en apprentissage profond, la rétroprogation n'est jamais qu'un mécanisme itératif permettant d'optimiser un objectif clairement décrit). C'est surtout que le modèle obtenu peut sembler complexe, impénétrable, pour tenir compte de possibles interactions entre les variables prédictives par exemple. Pour être complet, on devra toutefois distinguer les algorithmes supervisés classiques d'apprentissage machine et les techniques d'apprentissage par renforcement. Ce dernier cas décrit les méthodes d'apprentissage séquentiel, où l'algorithme apprend en expérimentant, comme le décrivent É. Charpentier et Remlinger 2020. On retrouvera ces algorithmes en conduite automatique par exemple, ou si on souhaitait modéliser proprement les liens entre les données, le modèle construit, les nouvelles données connectées, la mise à jour du modèle, etc. Mais nous n'insisteront pas davantage ici sur cette classe de modèles.

Si nous allons surtout parler ici de modèles de tarification en assurance, c'est-à-dire de modèle supervisés où la variable d'intérêt y sera la survenance d'un sinistre dans l'année à venir, le nombre de sinistres, ou la charge totale, il convient de garder en mémoire que les données en entrées ($\boldsymbol{x} \in \mathcal{X}$) peuvent elles-mêmes être les prédictions d'un modèle. Par exemple $\boldsymbol{x}_1$ pourrait désigner un score d'accélération observé l'année précédente (calculé par un prestataire externe qui a eu accès aux données télématiques brutes), $\boldsymbol{x}_2$ pourrait être la distance à la caserne de pompiers la plus proche (extrapolée à partir de logiciel de calculs de distance à partir d'une adresse), $\boldsymbol{x}_3$ pourrait être une prédiction du nombre de kilomètres parcourus, etc. Et dans la base d'apprentissage, les "observations passées" y_i pourraient aussi être des prévisions, surtout si on souhaite garder des sinistres récents, encore ouverts, mais dont la gestionnaire de sinistres peut donner une estimation, sur la base d'experts humains, mais aussi d'algorithmes "boites noires". On peut penser à ces applications qui donnent l'estimation du coût d'un sinistre matériel automobile sur la base de photo

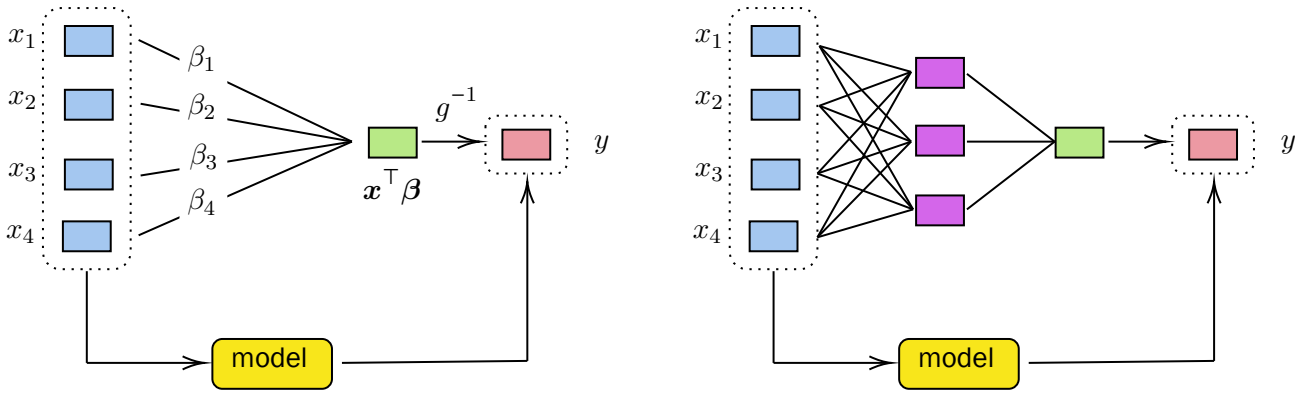


Figure 1.5 – Modèle linéaire généralisé (au gauche) et réseau de neurones (à droite), partant de mêmes variables prédictives $\boldsymbol{x} = (x_1, \cdots, x_k)$ et avec la même variable cible y. La couche intermédiaire (des réseaux de neurones) peut être vue comme la constitution de micro-modèles qui sont ensuite agrégés.

du véhicule envoyée par l'assuré, ou l'utilisation de barèmes d'indemnisation pour des sinistres non encore clôturés. En fait, si on compare un modèle linéaire généralisé (largement utilisé par les actuaires) et un réseau de neurones plus ou moins profond, on peut voir les neurones intermédiaires comme les sorties de précédents modèles (mais non-explicitement définis comme telles), comme sur la Figure 1.5, puisque

$$m_{\text{glm}}(\boldsymbol{x}) = g^{-1}(\boldsymbol{x}^\top\boldsymbol{\beta}) \text{ et } m_{\text{nnet}}(\boldsymbol{x}) = \sum_{j=1}^{3} \omega_j \, h_j(\boldsymbol{x}^\top\boldsymbol{\beta}_j).$$

1.3.3 Biais d'estimateur, de modèle et de données

Pour un statisticien le terme de **biais** a un sens bien précis, et correspond à une propriété mathématique et théorique d'un estimateur. Pourtant, dans la littérature scientifique, le terme de biais (ou "*bias*" en anglais) peut être utilisé dans un autre sens. Puhl et Brownell 2001, parlant d'obésité, de discrimination et de biais, écrivent "*employment bias was much greater for obese candidates than for average-weight applicants; the bias was more apparent for women than for men*". O'Neil 2016, Eubanks 2018, Hand 2020 utilisent aussi abondamment le terme, sans pour autant le définir clairement. L'adjectif "*biaisé*" se trouve associé à un modèle ou une décision, Castro 2019 parle ainsi de "*machine bias*".

Le sens retenu ici sera en lien avec la notion d'**échantillon biaisé** (ou "*sample bias*"), désignant le cas où l'ensemble d'individus échantillonné dans une population, censé la représenter, n'a pas été sélectionné correctement. On parle alors de biais car les indicateurs mesurés sur l'échantillon ne correspondent pas à ceux de l'ensemble de la population. Autrement dit, l'échantillon a été sélectionné ici de façon biaisée. Un exemple classique est le sondage du Literary Digest en 1936, en vue de prédire le vainqueur de l'élection présidentielle aux États-Unis (opposant alors Franklin Delano Roosevelt à Alf Landon). Comme le soulignent Cuddeback *et al.* 2004, si des groupes entiers de la population sont exclus d'un échantillon, aucun ajustement statistique ne peut produire des

estimations qui soient représentatives de l'ensemble de la population. Mais si certains groupes sont seulement sous-représentés, et que le degré de sous-représentation peut être quantifié, alors les poids de l'échantillon peuvent corriger le biais. Cependant, le succès de la correction est limité par le modèle de sélection choisi. Si certaines variables sont manquantes, les méthodes utilisées pour corriger le biais pourraient être inexactes. Nous reviendrons dans le Chapitre 3 plus spécifiquement sur les problèmes des données, et de biais.

1.3.4 Discriminations, versions juridique et statistique

La **discrimination** est (selon un dictionnaire de la langue française) "*l'action de séparer, de distinguer deux ou plusieurs êtres ou choses à partir de certains critères ou caractères distinctifs*". Kroll *et al.* 2017 rappellent que "*the word "discrimination" carries a very different meaning in engineering conversations than it does in public policy. Among computer scientists, the word is a value-neutral synonym for differentiation or classification : a computer scientist might ask, for example, how well a facial recognition algorithm successfully discriminates between human faces and inanimate objects. But, for policymakers, "discrimination" is most often a term of art for invidious, unacceptable distinctions among people—distinctions that either are, or reasonably might be, morally or legally prohibited*". Le terme discrimination peut alors être utilisé à la fois dans un sens purement descriptif (dans le sens de faire des distinctions) ou dans un sens normatif, qui implique que la différence de traitement de certains groupes est moralement mauvaise, comme le montrent Alexander 1992, ou plus récemment Loi et Christen 2021. Dans le domaine de l'assurance, comme le montrent Meyer et Rothstein 2004, les différences de traitement se traduisent souvent uniquement par des différences de primes (et c'est ce que nous évoquerons ici). Une discrimination pourrait être décrite comme "actuariellement équitable", caractérisant ainsi une "discrimination juste" dès lors que la différence de prime reflète une différence de risque, alors que la "discrimination injuste" se réfère au traitement inégal d'individus présentant le même niveau de risque, par exemple en proposant une indemnité plus faible à une personne qui ne se fait pas représenter par un avocat (comme dans Frees 2009).

Pour Al Ramiah *et al.* 2010, les **préjugés** désignent une attitude négative injustifiable envers un groupe et ses membres individuels. Les **stéréotypes** sont des croyances sur les attributs personnels d'un groupe de personnes, et peuvent être généralisés à l'excès, inexacts, et résister au changement en présence de nouvelles informations. Au sens juridique, la discrimination désigne un comportement négatif injustifiable à l'égard d'un groupe ou de ses membres (uniquement sur la base de l'appartenance à ce groupe [13]). Correll *et al.* 2010 donnent une définition très utile de la discrimination, à savoir un comportement dirigé vers les membres d'une catégorie qui a des conséquences sur leurs résultats et qui est dirigé vers eux non pas en raison d'un mérite ou d'une réciprocité particulière, mais simplement parce qu'ils sont membres de cette catégorie, "*behaviour directed towards category members that is consequential for their outcomes and that is directed towards them not because of any particular deservingness or reciprocity, but simply because they happen to be members of that category*". La notion de "mérite" est centrale dans l'expression et l'expérience de la discrimination (nous en parlerons en évoquant l'éthique par la suite). Il ne s'agit pas d'un critère défini objectivement, mais d'un critère qui trouve ses racines dans les inégalités et les normes sociétales historiques et actuelles.

Pour les économistes, l'étude de la discrimination est un ancien problème, décrit dans Charles et

13. Ou parfois simplement sur la supposition d'appartenance à un groupe, comme pour les préférences sexuelles, ou la discrimination contre "les étrangers", pour les traits non-visibles.

Guryan 2011, dont les fondements théoriques ont été posés par Becker 1957 (on peut trouver des réflexions plus anciennes dans Edgeworth 1922, par exemple). On peut également mentionner Phelps 1972, qui avait tenter de comprendre les origines de la discrimination, et qui avait affirmé (dans le contexte de la discrimination raciale) "*what has been called racism – similar remarks apply to sexism – can be hypothesized to be the consequence of scientific management in the impersonal pursuit of maximum profit, not racial hostility or intolerance*". Cette idée sera la base de la "discrimination statistique", où la question centrale était de relier la discrimination avec un comportement rationnel, et donc une notion d'efficacité. Bohren *et al.* 2019 rappellent que la discrimination statistique est parfois qualifiée de **discrimination efficace**, dans la mesure où elle constitue la réponse optimale à un problème d'extraction de signaux. Phelps 1972 et Aigner et Cain 1977 ont posé les bases de cette discrimination statistique. Par exemple, si la probabilité d'être "délinquant" et d'appartenir à un groupe ayant une caractérique visible plus élevée, en moyenne, que pour les autres groupes, alors un policier aura plus de chance de contrôler un "délinquant" s'il contrôle un membre de ce groupe (ce qui sera toujours valorisé s'il s'agit de "faire du chiffre"). Il y a donc, au niveau du groupe, une "raison statistique" qui va s'opposer avec le principe individuel de non-discrimination. Aussi, au nom de l'efficacité de la procédure, Gary Becker défendait le "profilage racial", comme le rappellent Lang et Spitzer 2020. On retrouve cet argumentation dans la lutte contre le terrorisme, car, comme le dit Becker 2005 "*if young Moslem Middle Eastern males were in fact much more likely to commit terrorism against U.S. than were other groups, putting them through tighter security clearance would reduce current airport terrorism*", autrement dit, le "profilage racial" est "efficace", même si "*such profiling is 'unfair' to the many young male Moslems who are not terrorists, and to the many minority shoppers who are honest*" [14]. Et il propose une méthode pour "tester" l'efficacité "*some profiling by governments and the private sector has been due to prejudice against various groups, not as a way of achieving efficiency. So it is crucial to be able to distinguish whether a profiling is efficient from whether it is evidence of discrimination. This distinction can be made in the terrorist field by keeping records on the fractions of young Moslem males and others who were searched and found with weapons or other evidence of intent to commit a terrorist act*". Un autre exemple classique de "discrimination statistique", justifiée par les économistes de part leur efficacité économique, est celui de la discrimination à l'embauche des jeunes femmes (qui pourrait tomber enceinte, et interrompre (temporairement) leur travail). Dans ce dernier cas, il n'est pas nécessaire d'avoir des statistiques car effectivement, seule une femme peut tomber enceinte [15]. Pour Gary Becker, cette "raison statistique" est, et doit être, l'unique critère de décision utilisé. On le voit, cette recherche de l'efficacité pose de nombreuses questions morales, et éthiques.

Sous l'angle juridique, Cornu 2016, présente la discrimination comme une "*différenciation contraire au principe de l'égalité civile consistant à rompre celle-ci au détriment de certaines personnes physiques en raison de leur appartenance raciale ou confessionnelle, plus généralement de critères sur lesquels la loi interdit de fonder des distinctions juridiques*" (tout en admettant "*plus rarement, dans un sens neutre, synonyme de distinction (non nécessairement odieuse)*" qui rappelle la vision du statisticien). L'égalité, que l'on retrouve énoncée dans la Déclaration Universelle des Droits de l'Homme, était initialement vu dans un sens "vertical", imposant une contrainte sur le comportement de l'État face aux citoyens. La loi va imposer des contraintes "horizontales", dans le droit privé, que ce soit sur le marché du travail (lors des entretiens d'embauche en particulier)

14. Gary Becker va plus loin en proposant "*That could be made up in part by compensating groups who are forced to go through more careful airport screening through putting them in shorter security lines, or in other ways. Similarly, innocent shoppers who are stopped and searched could be compensated for their embarrassment and time*".

15. De nombreux pays proposent maintenant des longs congés parentaux, qui peuvent être pris par n'importe quel parent, peu importe son sexe, ce qui remet en cause l'efficacité économique de ce profilage.

ou immobilier (pour la location de logements). L'article 23 de la Charte des droits fondamentaux de l'Union européenne impose le principe général selon lequel "*l'égalité entre les femmes et les hommes doit être assurée dans tous les domaines, y compris en matière d'emploi, de travail, et de rémunération*". Ce droit ne peut pas être invoqué par "les femmes", mais de manière individuel, par chacun (par exemple dans un litige avec un assureur). Ce "droit à l'égalité" (de traitement) appartient à une personne, en tant qu'individu, et non pas dans sa capacité en tant que membre d'un groupe sexuel (par exemple). Le droit dit qu'un individu ne peut pas être traité différemment en raison de son appartenance à un tel groupe, en particulier à un groupe auquel il n'a pas choisi d'appartenir. Cette vision individualiste du droit s'oppose fortement à la conception mutualiste et collective de l'assurance, les assureurs visant une forme d'égalité au sein de groupe, sur des moyennes, et non pas au niveau individuel.

Sous un angle plus moral, selon le principe du choix mentionné par Lippert-Rasmussen 2007, les gens ne devraient pas être soumis à un traitement désavantageux en raison de quelque chose qui ne reflète pas leurs propres choix. Cela peut expliquer pourquoi la discrimination fondée sur le sexe, la race, l'origine ethnique ou la génétique est largement perçue comme moralement problématique, comme le défendent Daniels 2004, Palmer 2007 ou Avraham *et al.* 2014. Toutefois, le principe de choix n'est pas violé, si les personnes ont imposé des coûts supplémentaires en conséquence de leur choix. Et que penser de la discrimination sur la base de croyances religieuses ? Comme on le voit, le terme "discrimination" semble englober toutes sortes de réalité. En particulier, pour reprendre la typologie de Thomsen 2017 et Khaitan 2017, on peut distinguer entre une **discrimination directe** et une **discrimination indirecte**, distinction sur laquelle nous reviendrons en détails dans le Chapitre 2, mais en un mot, une discrimination indirecte ("*proxy discrimination*" ou "*statistical discrimination*") signifie qu'au lieu de discriminer suivant une variable protégée p (ce qui ne serait pas autorisé par la loi), une variable x_j, très corrélée à p, est utilisée.

Dans le Chapitre 4, la variable protégée prendra des valeurs $\{0, 1\}$, ou $\{\bullet, \bullet\}$, avec une interchangeabilité entre les deux, par soucis de simplicité, et parce que l'objectif est une forme de **parité**. Nous mettrons souvent de côté le fait que les membres d'un des deux groupes est défavorisé, au moment où la décision est prise. Notons finalement que ces notions de favorisé/défavorisé ou dominant/dominé est vu suivant un critère non-explicite, comme le rappellent Deschamps et Personnaz 1979. Concernant le genre, les hommes sont vus comme favorisés et les femmes défavorisées, même si en assurance, les femmes vivent davantage et on souvent des risques plus faibles que les hommes, *ceteris paribus*. Nous ne reviendrons pas ici sur les difficultés à raisonner avec une dichotomie binaire favorisé/défavorisé ou dominant/dominé lorsque la variable protégée est non-binaire, voire continue, comme l'âge [16].

1.3.5 Justice et équité

Les humains ont un sens inné de l'équité et de la justice, des études montrant que même des enfants de trois ans ont démontré leur capacité à prendre en compte le mérite lors du partage des récompenses, comme le montraient Kanngiesser et Warneken 2012, ainsi que les chimpanzés et les primates, Brosnan 2006, et de nombreuses autres espèces animales. Et compte tenu du fait que ce trait est largement inné, il est difficile de définir ce qui est "**juste**", même si de nombreux scientifiques ont tenté de définir des notions de partage "justes", comme le rappellent Brams *et al.* 1996. "*Le juste* est, selon Ricœur 1991 *entre le légal et le bon*", où dans un premier sens,

16. L'âgisme, tel que défini par Butler 1969, repose sur l'idée que le groupe défavorisé est constitué de personnes âgées, même si dans certains contextes, les jeunes sont vus comme le groupe défavorisé.

"juste" renvoie à la **légalité** (et à la justice des hommes, traduite dans un ensemble de lois et de règlements), et dans un second sens, "juste" renvoie à un concept **éthique** ou **moral** (et à une idée de justice naturelle). Selon un dictionnaire, l'équité "*consiste à attribuer à chacun ce qui lui est dû par référence aux principes de la justice naturelle*". Et être "juste" pose des questions en lien avec l'éthique et la morale (nous ne ferons pas ici de différence entre l'éthique et la morale). Nous évoquerons ici l'éthique des modèles, ou, comme évoquée par certains auteurs (Mittelstadt *et al.* 2016 ou Tsamados *et al.* 2021), l'éthique des algorithmes. Une nuance existe vis-à-vis de l'éthique de l'intelligence artificielle, qui porte sur nos comportements ou nos choix (en tant qu'être humains) en lien avec les voitures autonomes, par exemple, et qui tentera de répondre à des questions telles que "*une technologie doit-elle est adoptée si elle est plus efficace*?". L'éthique des algorithmes questionne, elle, les choix faits "par la machine" (même si bien souvent, cela reflète des choix - ou des objectifs - imposés par la personne qui a programmé l'algorithme).

Programmer un algorithme de façon morale doit se faire suivant un certain nombre de normes. Deux types de normes sont généralement considérées par les philosophes. Les premières sont liées aux conventions, c'est-à-dire les règles du jeu (aux échecs ou au Go), ou le code de la route (pour les voitures autonomes). Les secondes sont les normes morales, qui doivent être respectées par tout le monde, et visent l'intérêt général. Ces normes se doivent d'être universelles, et donc ne favoriser aucun individu, ou aucun groupe d'individus. Cette universalité est essentielle pour Singer 2011 qui demande à ne pas juger une situation avec sa propre perspective, ou celle d'un groupe auquel on appartient, mais de prendre un point de vue neutre et "équitable". Formellement, de manière assez classique, en éthique normative, on opposera le **conséquentialisme** au **déontologisme**.

L'analyse éthique des discriminations se rattache au concept d'"égalité des chances", qui veut que le statut social des individus dépende uniquement du service qu'ils peuvent apporter à la société. Comme l'affirme la seconde phrase de l'article 1 de la Déclaration des droits de l'homme de 1789 "*Les distinctions sociales ne peuvent être fondées que sur l'utilité commune*", ou comme le disait Rawls 1999, "*offhand it is not clear what is meant, but we might say that those with similar abilities and skills should have similar life chances. More specifically, assuming that there is a distribution of natural assets, those who are at the same level of talent and ability, and have the same willingness to use them, should have the same prospects of success regardless of their initial place in the social system, that is, irrespective of the income class into which they are born*". Dans l'approche déontologique, inspirée par Emmanuel Kant, on va oublier les utilités de chacun, et simplement imposer des normes, et des devoirs. Ici, peu importe les conséquences (pour l'ensemble de la communauté), il y a des choses qui ne se font pas. On distinguera classiquement les **égalitariens** et **proportionnalistes**. Pour aller plus loin, Roemer 1996 et 1998 proposent une approche philosophique, Fleurbaey 1996 et Moulin 2004 une vision économique. Et dans un contexte plus informatique, Leben 2020 revient sur les principes normatifs pour évaluer l'équité d'un modèle.

Tous les cours d'éthique présentent des expériences de pensée, comme le dilemme du tramway. Dans la version initiale de Foot 1967, un tramway sans frein est sur le point d'écraser cinq personnes, et on a la possibilité d'actionner un aiguillage qui ferait dévier le tramway vers une autre voie où se trouve une personne qui sera alors condamnée. Que fait-on? Ou que devrait-on faire? Thomson 1976 a proposé une variante, avec une passerelle, où on a la possibilité de pousser une personne un peu corpulente, qui en s'écrasant sur la voie mourrait mais arrêterait le tramway. Cette dernière version est souvent plus dérangeante car l'action est alors indirecte, et on commence par commettre un meurtre pour sauver d'autres personnes. Certains auteurs ont

utilisé cette expérience de pensée pour distinguer l'explication (sur des bases scientifiques, et sur la base d'arguments causaux) de la justification (sur la bases de préceptes moraux). Cette expérience du tramway a été reprise dans l'expérience de psychologie morale, appelée la *machine morale* présentée dans Awad *et al.* 2018 [17]. Dans ce "jeu", on se retrouvait virtuellement au volant d'une voiture, et des choix étaient proposés : 'écrasez vous une personne ou cinq personnes ?', 'écrasez vous une personne âgée ou un enfant ?', 'écrasez vous un homme ou une femme ?'. Bonnefon 2019 revient sur l'expérimentation, et la série de dilemmes moraux, pour lesquels ils ont obtenus plus de 40 millions de réponses, venues de 130 pays. Si, naturellement, le nombre était une variable importante (on préfère tuer moins de monde), l'âge était aussi très important (priorité aux jeunes), et des arguments légalistes semblaient ressortir (on préfère tuer des piétons qui traversent en dehors des passages dédiés). Ces questionnements sont importants pour les voitures autonomes, comme le mentionnaient Thornton *et al.* 2016.

Pour un philosophe, la question "*ce modèle est-il juste pour le groupe* x *?*" sera toujours suivie de "*équitable selon quel principe normatif ?*". La mesure des effets globaux sur toutes les personnes affectées par le modèle (et pas seulement les droits de quelques-uns) conduira à intégrer des mesures d'équité dans un calcul général des coûts et avantages sociaux. Si nous choisissons une approche, d'autres en pâtiront. Mais c'est la nature des choix moraux, et la seule façon responsable d'atténuer les titres négatifs est de développer une réponse cohérente à ces dilemmes, plutôt que de les ignorer. Parler d'éthique des modèles pose des questions philosophiques dont on ne peut s'affranchir, car, comme nous l'avons dit, un modèle vise à représenter la réalité, *ce qui est*. Or lutter contre les discriminations, ou invoquer des notions d'équité, c'est parler de *ce qui doit être*. On retrouve la fameuse opposition "*is – ought*" de Hume 1739. Car quand on parle de "norme", il est important de ne pas confondre le descriptif et le normatif, la statistique (qui nous dit comment les choses sont) et l'éthique (qui nous dit comment les choses devraient être). La loi statistique relève de ce qui *est* parce qu'on l'a observé ainsi (par exemple '*les hommes sont plus grands que les chiens*'). La loi humaine (divine, ou judiciaire) relève de ce qui est *est* parce qu'on l'a décrété, et donc *doit être* ('*les hommes sont libres et égaux*' ou '*l'homme est bon*'). On peut voir la "norme" comme une régularité de cas, observée à l'aide de fréquences (ou de moyennes, comme on l'a évoqué dans la première partie), par exemple, sur la taille des individus, la durée du sommeil, autrement dit des données qui constituent la description d'individus. Les données anthropométriques ont ainsi permis de définir par exemple une taille moyenne des individus dans une population donnée, en fonction de leur âge ; par rapport à cette taille moyenne, un écart de 20% en plus ou en moins détermine le gigantisme ou le nanisme. Si l'on pense aux accidents de la route, il peut être considéré comme "anormal" d'avoir un accident de la route une année donnée, à un niveau individuel (micro), car la majorité des conducteurs n'ont pas d'accident. Néanmoins, du point de vue de l'assureur (macro), la norme est que 10 % des conducteurs aient un accident. Il serait donc anormal que personne n'ait d'accident. C'est l'argument que l'on retrouve dans Durkheim 1897. De l'acte singulier qu'est le suicide, s'il est considéré du point de vue de l'individu qui le commet, Durkheim tente de le voir comme un acte social, relevant alors d'une réelle régularité, au sein d'une société donnée. Dès lors, le suicide devient, selon Durkheim, un phénomène "normal". Les statistiques permettent alors de quantifier la tendance au suicide dans une société donnée, dès lors que l'on observe non plus l'irrégularité qui apparaît dans la singularité d'une histoire individuelle, mais une "normalité sociale" du suicide. L'anormalité est définie comme "*contraire à l'ordre habituel des choses*" (on pourrait y voir une notion empirique, statistique), ou "*contraire à l'ordre juste des choses*" (cette notion de *juste* appelle probablement à une définition normative),

17. Projet https://www.moralmachine.net/

mais aussi *non conforme au modèle*. Définir une norme n'est déjà pas simple si on ne s'intéresse qu'à l'aspect descriptif, empirique, comme peuvent le faire les actuaires lorsqu'ils construisent un modèle, mais si on intègre en plus une dimension de justice et d'éthique, on se doute que la complexité va augmenter. Nous reviendrons dans le Chapitre 4 sur les propriétés (mathématiques) qu'un modèle "juste", ou "équitable" devrait vérifier. Car si on demande à ce qu'un modèle vérifie des critères non nécessairement observés dans les données, il est indispensable d'intégrer une contrainte spécifique dans l'algorithme d'apprentissage du modèle, avec une pénalisation en lien avec une mesure de l'équité (tout comme il existe des mesures de complexité du modèle).

Chapitre 2

Discrimination et segmentation

La classification des risques, la segmentation, et donc une certaine forme de discrimination, sont le cœur de l'assurance. Comme le rappelle Bigot et Cayol 2020, l'assurance a deux visages : "*il convient en effet de distinguer deux choses lorsque l'on parle d'assurance. La première, l'opération d'assurance, relève de la technique et a une dimension collective, la seconde, le contrat d'assurance, relève du droit et a une dimension individuelle*". La base juridique de l'assurance est le contrat par lequel une partie (le souscripteur, ou l'assuré) se fait promettre par une autre partie (l'assureur) une prestation en cas de réalisation d'un risque, en contrepartie du paiement d'une prime (ou cotisation) dont le montant est décidé à la signature du contrat. Pour rendre la mutualisation des risques possibles, l'assureur va segmenter les risques afin de proposer la prime la plus "juste" possible. Pour reprendre les termes de Bigot et Cayol 2020, "*la sélection des risques désigne l'opération par laquelle l'assureur exclut les candidats à l'assurance qu'il ne souhaite pas ou ne peut pas couvrir, et répartit en groupes homogènes les assurés présentant des risques similaires, afin d'obtenir des résultats conformes à ses prévisions, fondées sur des données statistiques. En l'état du droit positif, le principe est celui de la licéité de la sélection. Notamment, l'assureur n'encourt pas le reproche de discrimination*". La segmentation est alors justifiée comme un impératif technique. Et comme nous l'avons mentionné dans l'introduction, elle est aussi justifiée économiquement (à cause d'une possible anti-sélection) mais aussi (probablement) moralement.

Nous allons revenir ici en détails sur certains exemples de discriminations, ou de pratiques qui ont été perçues ou jugées comme discriminatoires. En formalisant un peu, il existe des attributs protégés $\boldsymbol{p}$, que nous allons présenter ici, que nous considérerons comme des variables catégorielles, comme (historiquement) le genre ou origine ethnique, mais qui peuvent aussi être continues, comme l'âge. Pour simplifier, supposons qu'il existe une seule variable protégée p (nous reviendrons rapidement sur la discrimination multi-sources à la fin de cette partie). Il existe une variable d'intérêt y, qui sera en lien avec la sinistralité pour la tarification, mais on peut imaginer la fraude, ou un défaut de remboursement de crédit. Et il existe des variables $\boldsymbol{x} = (x_1, \cdots, x_k)$ que l'on espère très corrélées au facteur latent non observé θ, et dont certaines peuvent être aussi très corrélées à p, que nous appellerons des "*proxys*". Dans la partie 4, nous proposerons des définitions de ce que "discriminer" veut dire, formellement (ou disons "mathématiquement"), mais auparavant, nous présenterons des exemples de ces variables p et x_j, tout en présentant les approches juridiques et économiques.

2.1 Quelles variables tarifaires ?

Les actuaires en charge de construire un modèle tarifaire ont pour mission de trouver les variables prédictives $\boldsymbol{x} = (x_1, \cdots, x_k)$ qui permettront de mieux prédire le risque, comme discuté dans l'introduction. Mais au-delà de l'aspect statistique, il existe des régulations. Le Tableau 2.1 présente quelques variables tarifaires classiquement utilisées en assurance automobile [18], autorisées (légalement) par État et Province [19], aux États-Unis d'Amérique et au Canada [20]. Cette hétérogénéité permet de mettre en avant le caractère culturel de la discrimination.

	CA	HI	GA	NC	NY	MA	PA	FL	TX	AL	ON	NB	NL	QC
Genre	x	x	•	x	•	x	x	•	•	•	•	x	x	•
Âge	x	x	•	x*	•	x	•	•	•	•*	•	x	x	•
Expérience de conduite	•	x	•	•	•	•	•	•	•	•	•	•	•	•
Antécédents de crédit	x	x	•	•	•	x	•*	•	•	x*	x	•*	x	•
Éducation	x	x	x	x	x	x	•	•	•	•	•	•	•	•
Profession	x	x	x	•	x	x	•	•	•	•	•	•	•	•
Situation d'emploi	x	x	x	•	x	x	•	•	•	•	•	•	•	•
Situation de famille	•	x	•	•	•	x	•	•	•	•	•	•	•	•
Situation résidentielle	x	x	•	•	•	x	•	•	•	x	x	•	•	•
Adresse/code postal	•	•	•	•	•	•	•	•	•	x	x	•	•	•
Antécédents d'assurance	•	•	•	•	•	•	•	•	•	•	•	•	•	•

Table 2.1 – Un facteur est considéré comme "autorisé" (•) lorsqu'il n'existe pas de lois ou de politiques réglementaires dans l'État ou la Province qui interdisent aux assureurs d'utiliser ce facteur. Dans le cas contraire, il sera "interdit" (x). * En Caroline du Nord, l'âge est autorisé uniquement lorsqu'il s'agit d'accorder un rabais aux conducteurs âgés de 55 ans et plus. En Pensylvanie, la cote de crédit peut être utilisée pour les nouvelles affaires et pour réduire les taux au renouvellement, mais pas pour augmenter les taux au renouvellement. En Alberta, la cote de crédit et l'ancienneté du permis de conduire ne peuvent être utilisées pour la couverture obligatoire (mais peuvent l'être sur des couvertures optionnelles). Au Labrador, l'âge ne peut être utilisé avant 55 ans, et pour les plus de 55 ans, seul un rabais est autorisé (comme en Caroline du Nord).

La discrimination actuarielle, que Gandy 2016 appelle aussi "*discrimination based on actuarial evidence*", semble autorisée dans de nombreuses juridictions. En Europe, l'article 5-2 de la directive 2004/113, sur le fondement de la Charte des droits fondamentaux, prévoyait que "*les États membres peuvent décider (⋯) d'autoriser des différences proportionnelles dans les primes et les*

18. Pour les États-Unis, https://www.thezebra.com/resources/research/car-insurance-rating-factors-by-state/ et et au Canada, Bureau d'Assurance du Canada 2021.

19. Pout les ÉTats-Unis, CA : Californie, HI : Hawaii, GA : Georgia, NC : Caroline du nord, NY : New York, MA : Massachusetts, PA : Pennsylvanie, FL : Floride, TX : Texas, et pour le Canada, AL : Alberta, ON : Ontario, NB : Nouveau-Brunswick, NL : Terre-Neuve-et-Labrador, QC : Québec.

20. Contrairement à la plupart des pays européens, qui ont une culture juridique romano-civiliste (où la principale source du droit se trouve dans les codes juridiques), les Provinces canadiennes et les États des États-Unis d'Amérique ont un système de "*common law*" (où les règles sont principalement édictées par les tribunaux au fur et à mesure des décisions individuelles). Le Québec utilise un droit mixte. Il existe dans la plupart des États et Provinces des documents listant les "*Prohibited Rating Variables*", comme Automobile Insurance Rate Board 2022 en Alberta.

prestations des particuliers lorsque l'utilisation du sexe est un facteur déterminant dans l'évaluation du risque, sur la base de données actuarielles et statistiques pertinentes et précises", comme le rappelle Laulom 2012. En France, l'article L. 111-7 du Code des Assurances affirme que "*le ministre chargé de l'économie peut autoriser par arrêté des différences de primes et de prestations fondées sur la prise en compte du sexe et proportionnées aux risques lorsque des données actuarielles et statistiques pertinentes et précises établissent que le sexe est un facteur déterminant dans l'évaluation du risque d'assurance*". Au Québec, comme l'affirme l'article 20.1 de la charte des droits et libertés de la personne, "*la distinction fondée sur l'âge, le sexe ou l'état civil est permise lorsqu'elle repose sur un facteur qui permet de déterminer un risque. Par exemple, une compagnie d'assurance peut vous poser des questions sur votre âge et votre sexe pour fixer votre prime*". En Californie, comme le notait Butler et Butler 1989, de nombreuses variables de tarification ne peuvent être utilisées par les assureurs s'il ne sont pas liés de manière causale au risque d'accident et à leur coût.

Discrimination & assurance, par Rodolphe Bigot

Pour lutter contre les discriminations, il existe des interdictions générales, formulées dans le Code pénal et le Code civil, et des interdictions spéciales en droit des assurances, insérées dans le Code des assurances. Toutes deux connaissent des aménagements à géométrie variable liés aux spécificités de l'activité d'assurance. En **matière civile**, une interdiction générale, applicable à l'assurance, est prévue à l'article 16-13 du Code civil, lequel dispose que « *Nul ne peut faire l'objet de discriminations en raison de ses caractéristiques génétiques* ». En **matière pénale**, le refus ou la subordination de la fourniture d'un service ou d'un bien fondé sur l'un des critères – discriminatoires – figurant à l'article 225-1 ou 225-1-1 du Code pénal constitue une discrimination répréhensible (C. pén., art. 225-1 et s.). Depuis 2008, sont ainsi visées l'ensemble des discriminations directes ou indirectes fondées sur de tels critères (L. 27 mai 2008; modif. par L. 18 nov. 2016 de modernisation de la justice du XXI^{ème}siècle), avec une dérogation générale en présence de différences de traitement « *justifiées par un but légitime* » et si « *les moyens de parvenir à ce but sont nécessaires et appropriés* ». Un renversement de la charge de la preuve est prévu pour les actions portées devant le juge civil (L. 27 mai 2008, art. 4), à la différence des règles de preuve applicables devant le juge pénal où il revient au demandeur de prouver la discrimination.

En premier lieu, **une distinction fondée sur l'âge** notamment (C. pén., art. 225-1 et 225-2) est constitutive d'une discrimination pénalement sanctionnée si elle tend à appliquer une différence de traitement pour l'accès aux garanties d'assurance ou pour la cessation des prestations d'assurance : elle est donc interdite dans les risques de perte d'emploi pour garantir un prêt bancaire par exemple. Mais une dérogation est prévue pour la tarification des contrats d'assurance sur la vie (assurances décès et rentes viagères), par application des tables de mortalité (C. assur., art. A 132-18). Le Défenseur des droits a admis cette discrimination pour un candidat à l'assurance santé âgé de 74 ans dont l'adhésion a été refusée car la police la limitait à 70 ans (avis n° MLD/ 2012-150, 16 novembre 2012 : "*lorsqu'elles sont objectivement justifiées par des éléments actuariels et statistiques, les limites d'âge dans l'accès à un contrat d'assurance de personnes ne constituent pas des discriminations*" (…) "*le caractère aléatoire du contrat d'assurance, les principes de sélection des risques et de leur mutualisation, peuvent justifier la prise en compte du critère de l'âge en matière d'assurance de personnes*". L'article 2 de la loi du 27 mai 2008 prévoit deux autres exceptions qui devraient être confirmées par la proposition de directive relative à l'égalité de traitement entre les personnes (COM 2008/426 du 2 juillet 2008) : 1) les différences de traitement justifiées par un but légitime, si les moyens de parvenir à ce but sont nécessaires et appropriés; 2) lorsque les différences de traitement sont prévues et autorisées par les lois et règlements en vigueur.

En deuxième lieu, sont également prohibées les discriminations fondées sur « **la situation de famille** » ou sur l' « **orientation sexuelle** » (C. pén., art. 225-1 et 225-2). Entrerait dans cette catégorie, en présence d'un couple d'assurés homosexuels, le refus par un employeur de verser le bénéfice d'un capital décès au profit du partenaire pacsé d'un salarié ou au profit du salarié pacsé en cas de décès de son partenaire (…)

Pour reprendre l'analyse de Cummins *et al.* 2013, il existe plusieurs critères pour justifier l'utilisation (ou pas) d'une variable tarifaire : un critère actuariel, opérationnel, d'acceptabilité sociale et légal.

2.1.1 Un critère actuariel

Une variable de classification est considérée comme actuariellement juste, au sens de Cummins *et al.* 2013, si elle est précise, si elle assure l'homogénéité entre les membres, si elle présente une crédibilité statistique et si elle est fiable dans le temps. Une variable de classification sera dite "exacte" si elle répartit les assurés de façon à ce que chacun paie une prime proportionnelle à son coût prévu des sinistres. Pour prévenir l'antisélection, c'est probablement le critère le plus important. L'homogénéité exige que tous les assurés d'une même classe de risque aient le même coût attendu des sinistres. Il faut un grand nombre d'assurés dans chaque groupe pour que l'historique des sinistres du groupe soit statistiquement crédible, et que la mutualisation puisse avoir encore du sens. Un nombre trop faible de membres entraîne des pertes qui varient fortement d'une année à l'autre et fait fluctuer les primes de la même manière. Enfin, une variable de classification fiable produit des différences de coûts entre les différents groupes qui restent relativement stables dans le temps. Être perçu comme un "bon risque" en 2020 puis comme un "mauvais risque" en 2021, sans avoir eu l'impression que sa conduite ait changé entre temps, ne sera généralement pas bien perçu.

Discrimination & assurance, par Rodolphe Bigot

(…) En troisième lieu, les discriminations fondées sur **la grossesse et la maternité des femmes** sont prohibées. En assurance, il ne peut leur être réservé un traitement moins favorable en matière de primes et de prestations par suite des frais liés à la grossesse et à la maternité (C. assur., art. L. 111-7, I, alinéa 2) car par principe les deux sexes sont concernés par la maternité. Autrement dit, les contrats d'assurance, quel que soit le risque couvert (santé, prévoyance…), ne doivent plus comporter de traitements différenciés sur ces facteurs. La clause instituant un délai de carence plus long pour la prise en charge des frais d'hospitalisation, lorsque cette dernière est consécutive à une grossesse, n'est donc plus licite. En revanche, s'il existe des données actuarielles et statistiques pertinentes, autrement dit par des considérations objectives, « *les dispositions contractuelles plus favorables aux femmes pourront subsister. Le non-respect de ces dispositions expose aux sanctions prévues par le Code pénal pour les actes de discrimination* » (Lamy Assurances, 2021, n° 3806).

En quatrième lieu, le refus de la fourniture d'un bien ou d'un service **en raison du lieu de résidence** d'une personne constitue une discrimination au sens pénal (C. pén., art. 225-1). S'il est toujours possible de moduler le montant de la prime en fonction du lieu de résidence de l'assuré, l'assureur se voit interdire de refuser un candidat à l'assurance à raison de ce facteur. Ne sont pas concernées les souscriptions ou adhésions réalisées par des résidents d'États avec lesquels un assureur français n'est pas habilité à contracter (Lamy Assurances, 2021, n° 3808).

En cinquième lieu, est prohibé le refus d'assurance en raison de la **précarité sociale de l'assuré**, qui entre dans le champ de l'infraction de discrimination pour refus de la fourniture d'un bien ou d'un service en raison « *de la particulière vulnérabilité résultant de la situation économique, apparente ou connue de son auteur* » (C. pén., art. 225-1, modifié par L. 24 juin 2016 visant à lutter contre la discrimination à raison de la précarité sociale). L'assureur ne peut rejeter une adhésion ou une souscription sur cette base mais il peut néanmoins tenir compte de ce facteur pour moduler le montant de la prime. Relevons que depuis le 26 juin 2016, une discrimination indirecte peut résulter d'une disposition, un critère ou une pratique neutre en apparence, mais susceptible d'entraîner un désavantage particulier pour des personnes. Il existe par ailleurs un système complexe de protection pour les personnes vulnérables en assurance : "le mouvement pour les personnes vulnérables est une danse qui alterne entre protection spéciale et protection indifférenciée" (Noguéro 2010, op. cit., RGDA, p. 633).

2.1.2 Un critère opérationnel

Certaines variables de classification des risques actuariellement équitables ne peuvent pas être mises en œuvre dans la pratique car elles ne possèdent pas les critères opérationnels d'objectivité, de faible coût de mise en œuvre et de difficulté de manipulation. Les données dont la collecte et

la vérification sont lourdes et coûteuses font rarement de bonnes variables de classification. En lien avec l'objectif de faibles coûts administratifs, les données utilisées dans un autre but font de bonnes variables de classification des risques. L'utilisation d'une variable déclarée, ou collectée par d'autres organismes, réduit la probabilité qu'elle soit manipulée, et comme le soulignent Cummins *et al.* 2013, diminue le coût de la vérification. Une variable de classification doit offrir peu d'ambiguïté entre les assurés, et les classes totales décrites par la variable doivent être mutuellement exclusives et exhaustives.

Discrimination par le sexe & assurance, par Rodolphe Bigot

Les discriminations relatives au **sexe** font l'objet d'un régime renouvelé dédié aux nouveaux contrats conclus à compter du 21 décembre 2012 (sauf les contrats de retraite, maladie et accident souscrits par un employeur et auxquels adhère le salarié à titre obligatoire ; C. séc. soc., art. L. 911-1) et visant à appliquer de manière uniforme la règle unisexe – prohibant toute discrimination directe ou indirecte fondée sur le sexe – aux contrats d'assurance au sein de l'Union européenne (Parléani 2012, p. 563). L'article A. 111-6 du Code des assurances a intégré les lignes directrices de la Commission européenne (Arr. 18 déc. 2012, NOR : EFIT1238658A, relatif à l'égalité entre les hommes et les femmes en assurance, JO 20 déc., mod. par Arr. 3 févr. 2014, NOR : EFIT1400411A, JO 11 févr.).

Le calcul des primes et prestations entre dans le champ d'application de la règle unisexe (dans les opérations d'assurance classées, par référence à l'article R. 321-1, dans les branches : 1 « Accidents (y compris les accidents du travail et les maladies professionnelles) » (art. a. 111-2), 2 « Maladie » (art. A. 111-3), 3 « Corps de véhicules terrestres (autres que ferroviaires) » (art. A. 111-4), 10 «Responsabilité civile véhicules terrestres automoteurs» (art. A. 111-4), 20 «Vie-décès» (art. A. 111-5), 22 « Assurances liées à des fonds d'investissement » (Art. A. 111-5), 23 « Opérations tontinières » (Art. A. 111-5), 26 « Toute opération à caractère collectif définie à la section I du chapitre I^{er} du titre IV du livre IV » (art. A. 111-5)). A titre dérogatoire, le critère du sexe peut être employé « *comme facteur d'évaluation des risques en général pour recueillir, stocker, utiliser des informations sur le sexe ou liées au sexe pour le provisionnement et la tarification internes, le marketing et la publicité, la tarification de la réassurance. L'utilisation du sexe comme discrimination indirecte est également admise pour la tarification de certains risques réels, comme par exemple une différenciation de primes fondée sur la taille du moteur d'une voiture alors même que les voitures les plus puissantes sont de fait davantage achetées par les hommes* » (Lamy Assurances, 2021, 3803).

Pour mettre en conformité la loi française avec les règles européennes, l'article L. 111-7 du Code des assurances a été réécrit avec la loi du 26 juillet 2013. Un alinéa II bis a été ajouté : « *La dérogation prévue au dernier alinéa du I est applicable aux contrats et aux adhésions à des contrats d'assurance de groupe conclues ou effectuées au plus tard le 20 décembre 2012 et à ces contrats et adhésions reconduits tacitement après cette date. La dérogation n'est pas applicable aux contrats et aux adhésions mentionnés au premier alinéa du présent II bis ayant fait l'objet après le 20 décembre 2012 d'une modification substantielle, nécessitant l'accord des parties, autre qu'une modification qu'une au moins des parties ne peut refuser* ». En matière de garanties collectives complémentaires des salariés, aucune discrimination fondée sur le sexe ne peut être réalisée. Mais les assureurs ont toujours la possibilité de proposer des options dans les polices ou des produits d'assurance selon le sexe afin de couvrir les conditions qui concernent exclusivement ou essentiellement les hommes ou les femmes. Une couverture différenciée est donc possible pour le cancer du sein ou de l'utérus en encore le cancer de la prostate.

2.1.3 Un critère d'acceptabilité sociale

Une troisième considération dans la sélection des variables de classification des risques est l'acceptabilité sociale. Selon la classification de Cummins *et al.* 2013, les quatre principaux critères sont le respect de la vie privée, la causalité, la contrôlabilité et le caractère abordable/disponible. Le respect de la vie privée influe sur la volonté des individus de divulguer certaines informations qui, à leur tour, influent sur la précision d'une variable de classification des risques ainsi que sur la facilité avec laquelle elle peut être collectée et vérifiée (nous reviendrons sur ce point à la fin de cette partie). La causalité exige plus qu'une relation intuitive entre la variable de classification

et les pertes attendues. Une bonne variable de classification des risques doit encourager les individus à agir pour réduire la fréquence et/ou la gravité attendues de leurs pertes - le critère de "contrôlabilité". Le critère social d'abordabilité / disponibilité exige que ceux qui doivent acheter une protection d'assurance puissent raisonnablement le faire. L'acceptabilité sociale semble d'autant plus grande que le risque est relié à un critère relevant d'un choix de l'assuré.

2.1.4 Un critère légal

Enfin, en pratique, l'utilisation ou l'interdiction de certaines variables de classification est le plus souvent imposée par une loi (ou un règlement). Au Canada, les lois provinciales, qui sont généralement plus restrictives que dans la plupart des États aux États-Unis, exigent généralement que les variables de classification ne soient pas injustement discriminatoires, c'est-à-dire que le critère d'équité actuarielle doit être démontré. Cependant, les variables de classification ont parfois été interdites parce qu'il n'existe qu'une relation de corrélation et non de causalité entre la variable de classification et les coûts des sinistres prévus, ou parce qu'elles ont été jugées socialement inacceptables. Les critères légaux, évidemment, varient par État et Province, comme nous le mentionnions dans le Tableau 2.1. Des encadrés reviennent sur le critère légal en France, entre la page 43 et la page 45.

2.1.5 Variable protégée

Sans variable explicative, on dispose de la paire (y, p) avec p binaire, par exemple le genre [21], $p \in \{\mathsf{H}, \mathsf{F}\}$, ou plus généralement une distinction de la forme $p \in \{0, 1\}$ [22]. Avec des variables explicatives, on dispose du triplet $(y, \boldsymbol{x}, p)$. Considérons une base de données télématiques, avec 1177 contrats d'assurances, observés pendant une année complète, décrite dans le Tableau 2.2 et la Figure 2.1. On dispose d'un couple de variables (x_1, x_2), respectivement le nombre de kilomètres parcourus en 2019 et l'âge de l'assuré (qui n'est pas vu ici comme un attribut protégé), d'une variable protégée binaire p correspondant au genre du conducteur, et y est la variable binaire indiquant la survenance d'accident en 2020, que l'on cherche à prédire quand on veut calculer un tarif.

Variable protégée binaire, sans variables explicatives

Savoir si le genre p est une variable "prédictive", ou "discriminante" (quand on veut prédire la sinistralité) peut s'interpréter sous la forme d'un test statistique, $H_0 : \mathbb{E}[Y|P = \mathsf{H}] = \mathbb{E}[Y|P = \mathsf{F}]$, qui est un test de moyenne sur deux échantillons. Dans le cas où y est une variable binaire, on a un test de proportion. Le test est alors équivalent à tester l'indépendence entre les deux variables binaires y (à valeurs dans $\{0, 1\}$) et p (à valeurs dans $\{\mathsf{H}, \mathsf{F}\}$), par un test du chi-deux. Ici, le test d'égalité de proportion et le test du chi-deux ont une p-value [23] de 68.45 %.

21. Comme rappelé en préambule du document, conformément à l'usage dans la littérature, le genre sera souvent utilisé comme variable protégée pour illustrer différentes notions en lien avec les discriminations, et sera traité comme une variable binaire (prenant deux valeurs "homme" et "femme"). Il ne s'agit aucunement d'une prise de position.

22. La littérature sur les discrimination utilise les termes "*groupe privilégié*", "*groupe favorisé*" et "*groupe majoritaire*" de façon presque interchangeable, comme discuté à la fin de la section 1.3.4. Encore une fois, ces étiquettes ne sont pas destinées à transmettre un quelconque jugement concernant le mérite réel, mais sont uniquement destinées à être descriptives et à reconnaître qu'historiquement, un groupe a été favorisé, autrement dit, qu'il a historiquement reçu une sorte de statut privilégié que le "*groupe défavorisé*" n'a pas reçu.

23. L'hypothèse alternative étant ici que le genre est (significativement) discriminant, $H_1 : \mathbb{E}[Y|P = \mathsf{H}] \neq \mathbb{E}[Y|P = \mathsf{F}]$.

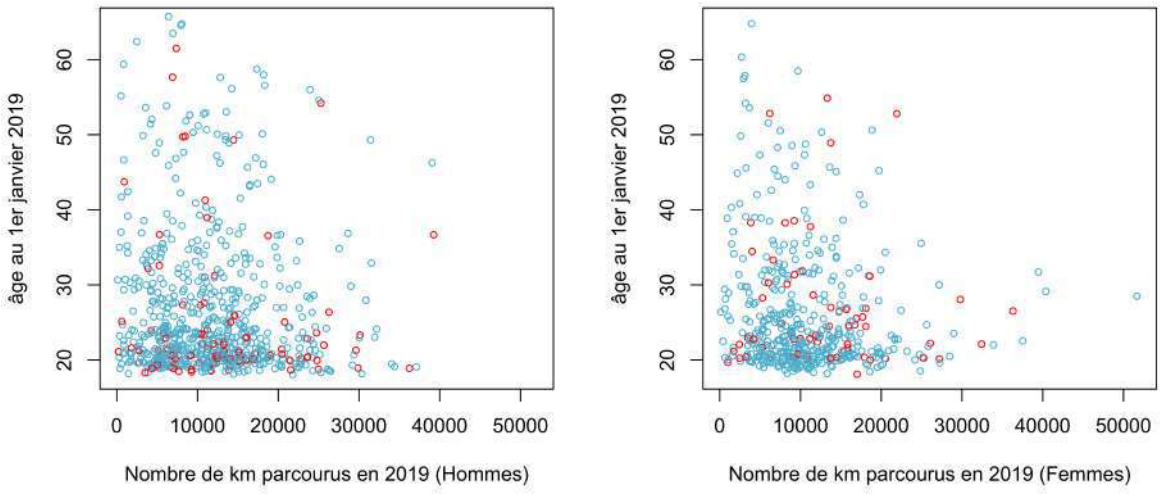


Figure 2.1 – Survenance d'un accident de la route en fonction de l'âge de conducteur, du nombre de kilomètres parcourus pendant l'année 2019, et du genre du conducteur. Les points bleus sont les conducteurs non-sinistrés en 2020, et les points rouges sont les conducteur qui ont eu un sinistre en 2020.

	Total	Femmes	Hommes	Proportions
18-20 ans	26/202~12.9 %	3/69~ 4.3 %	23/133~17.3 %	34.2 %-65.8 %
21-25 ans	83/529~15.7 %	34/229~14.8 %	49/300~16.3 %	43.3 %-56.7 %
plus de 26 ans	49/446~11.0 %	25/185~13.5 %	24/261~ 9.2 %	41.5 %-58.5 %
moins de 10,000 km/an	57/550~10.4 %	25/256~ 9.8 %	32/294~10.9 %	46.5 %-53.5 %
10,000-20,000 km/an	74/504~14.7 %	29/191~15.2 %	45/313~14.4 %	37.9 %-62.1 %
plus de 20,000 km/an	27/123~22.0 %	8/36~22.2 %	19/87~21.8 %	29.3 %-70.7 %
Total	158/1177~13.4 %	62/483~12.8 %	96/694~13.8 %	41.0 %-59.0 %

Table 2.2 – Statistiques descriptives d'accidents sur une base de 1177 assurés couverts pendant l'année 2019. La première variable est l'âge du conducteur (coupée en trois classes), la seconde est le nombre de kilomètres parcourus en 2019 (source : compagnie d'assurance en France).

Variable protégée binaire, avec variables explicatives (x)

Au lieu de regarder l'indépendance entre y et p (comme auparavant), on peut tester l'indépendance entre y et $p \times x$, pour une variable catégorielle x, comme une tranche d'âge, ou la distance parcourue en un an. Pour un seuil d'âge de 25 ans (créant deux catégories, les moins de 25 ans et les plus de 25 ans), et pour un seuil de kilométrages de 10,000 km, on a les tableaux de comptage de la Table 2.3. Les p-values des tests du chi-deux sont respectivement de 1.3 % et de 3.7 %, ce qui nous pousse à rejeter l'hypothèse d'indépendance entre la sinistralité y et l'interaction genre et âge (où on oppose les jeunes conducteurs aux plus expérimentés) et genre et kilométrage (où on oppose

ceux qui conduisent relativement peu aux autres). Dans le dernier cas, on voit que le genre importe peu, la variable principale étant le kilométrage.

$p \times x_1$	$y=0$	$y=1$	prop.
$P=$ F, âge ≤ 26	279	39	12.3 %
$P=$ F, âge > 26	142	23	13.9 %
$P=$ H, âge ≤ 26	374	76	16.9 %
$P=$ H, âge > 26	224	20	8.2 %

$p \times x_2$	$y=0$	$y=1$	prop.
$P=$ F, dist. $\leq 10,000$	231	25	9.8 %
$P=$ F, dist. $> 10,000$	190	37	16.3 %
$P=$ H, dist. $\leq 10,000$	262	32	10.9 %
$P=$ H, dist. $> 10,000$	336	64	16.0 %

Table 2.3 – Statistiques descriptives d'accidents sur une base de 1177 assurés couverts pendant l'année 2019.

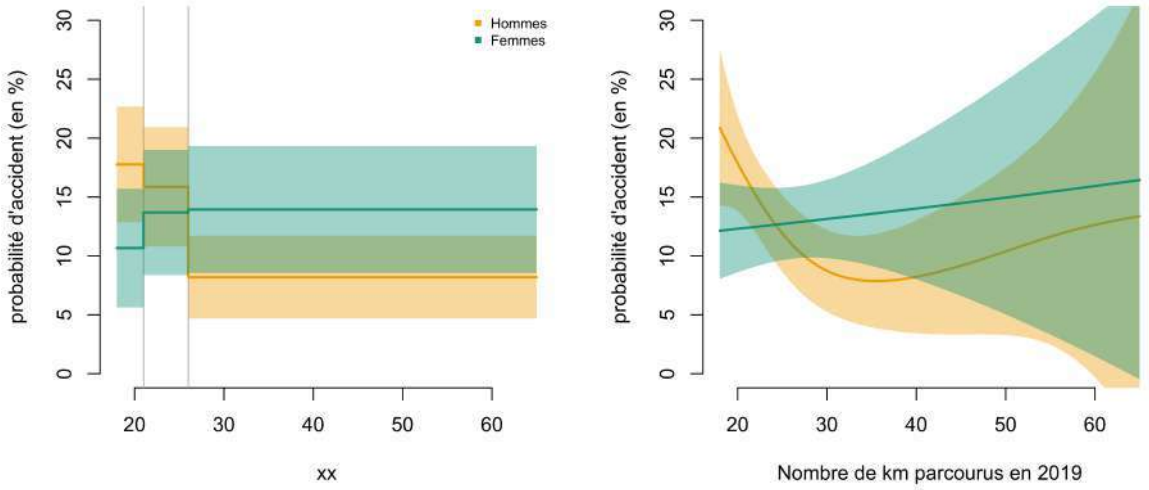


Figure 2.2 – Survenance d'un accident de la route en fonction de l'âge de conducteur et du genre du conducteur, avec l'âge en classes à gauche, et la version continue (lissée) à droite, et avec la distinction entre les hommes et les femmes.

Variable protégée non-binaire

La grande majorité des exemples dans la littérature sur l'équité traite exclusivement du cas où la variable protégée p est binaire, en témoignent Nielsen 2020 ou Agarwal et Mishra 2021 pour des ouvrages récents sur le sujet. Mais dans certains cas, des variables continues p peuvent être considérées, comme nous l'avions vu dans le Tableau 2.1, certaines variables continues peuvent être utilisées.

- l'âge (ou l'ancienneté du permis de conduire), interdit dans plusieurs États, comme la Californie ou le Massachusetts, aux États-Unis,
- l'adresse ou le code postal, correspondant à une information spatiale, interdite dans l'Alberta et l'Ontario, au Canada.

Dans la majorité des cas, la variable continue est discrétisée en classes (au lieu de l'âge, on utilisera des classes d'âges ; au lieu de la localisation géographique, on utilisera des régions prédéfinies).

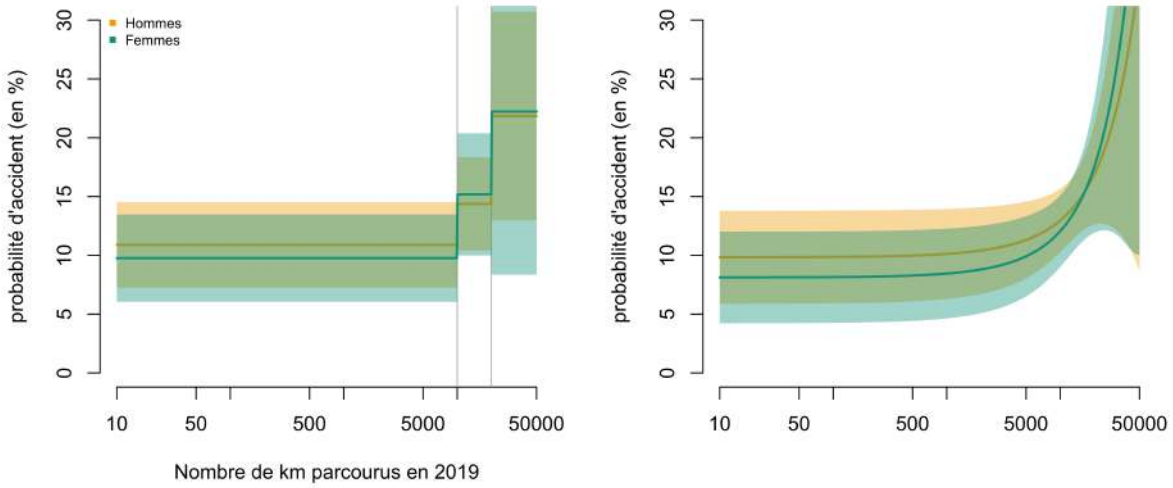


Figure 2.3 – Survenance d'un accident de la route en fonction du nombre de kilomètres parcourus pendant l'année et du genre du conducteur, avec la distance en classes à gauche, et la version continue (lissée et en échelle logarithmique) à droite, et avec la distinction entre les hommes et les femmes.

Mais il est possible de traiter la variable p comme une variable continue, et d'utiliser des tests d'indépendance entre p et la variable cible y pour tester la présence, ou non, de discrimination, comme dans Grari et Detyniecki 2022. Les tests de discrimination seront présentés dans le Chapitre 4.

2.2 Quelques exemples de discriminations

Le 21 août 1789 (comme le raconte Buchez 1846) Alexandre de Lameth, député de la noblesse mais rallié au tiers état, propose ce qui deviendra l'article 5 de la Déclaration des droits de l'homme et du citoyen, affirmant que "*Tout ce qui n'est pas défendu par la Loi ne peut être empêché*" (abusivement traduit [24] "*tout ce qui n'est pas interdit est permis*"). Dans cette section, nous allons revenir sur les principales variables protégées [25], rencontrées en assurance, avec la discrimination raciale, par le genre, par l'âge, par le poids, et des fumeurs.

24. On dirait (en latin) *ubi lex non distinguit, nec nos distinguere debemus* (il n'y a pas lieu de distinguer là où la loi ne distingue pas).

25. Certains auteurs ont, historiquement, suggéré de considérer une discrimination basée de la richesse (ou du revenu), comme Brudno 1976, Gino et Pierce 2010, ou plus récemment Paugam *et al.* 2017, mais ce critère sera ici écarté. Les réflexions de ces auteurs soulèvent toutefois des questions importantes sur les liens entre les discriminations, les inégalités (économiques) et la méritocratie, comme le soulignait aussi Dubet 2014 et 2016.

Mesurer une inégalité et une discrimination, par Avner Bar-Hen [26]

Tout le monde a son idée sur les inégalités et en parle comme si c'était quelque chose de simple. Penchons nous sur la construction d'indicateurs de disparités de genre. Classiquement, il existe deux grandes approches dans la mesure des indices d'inégalité de genre : (i) La première méthode mesure l'égalité des genre à l'aide d'enquêtes, généralement en complément à d'autres questions. Dans ce cas, la connaissance de l'égalité entre les genres n'est pas l'objectif principal. (ii) La deuxième approche est basée sur des objectifs et des statistiques pour les quantifier. Les premiers indices internationaux ne datent que de 1995 et sont fortement basés sur l'indice de développement humain (IDH) du PNUD. L'IDH est un concept bien théorisé et largement diffusé ; ceci permet de savoir assez précisément ce que l'on mesure mais les comparaisons internationales sont difficiles. Les variables complémentaires à l'IDH qui ont pu être proposées sont souvent intéressantes mais compliquent surtout la lisibilité du résultat.

Depuis les années 2000, plusieurs alternatives ont été proposées. Parmi les plus populaires, on peut citer l'indice de parité hommes-femmes (*Gender Equity Index*, GEI, introduit par Social Watch en 2004) ou l'Indice mondial de l'écart entre les genres (GGGI proposé par le Forum Économique Mondial en 2006). Ces indices se revendiquent comme des mesures de l'égalité des genres mais leurs concepts généraux ne sont pas clairement formulés. Par exemple, le GEI ignore les causes sous-jacentes de l'inégalité entre les genres, comme la santé. D'autre propositions comme l'indice des institutions sociales et des genres (*Social Institutions and Gender Index*, SIGI, proposé par l'OCDE en 2007) se concentre sur les institutions sociales qui affectent l'égalité entre hommes et femmes, mais aussi sur le code de la famille ou les droits de propriété. Cet indice peut être vu comme une mesure combinée des désavantages féminins par rapport aux hommes concernant certains droits fondamentaux et l'accomplissement de droits distinctifs pour les femmes. Le manque de symétrie dans les indicateurs du SIGI et les échelles des indicateurs ne permet pas au final de savoir ce qui est mesuré.

Le but de ces différents indices est d'évaluer les multiples aspects des disparités entre les genres, non seulement dans les recherches universitaires sur les causes et les conséquences de l'inégalité entre les genres, mais aussi pour éclairer les débats de politiques publiques. Un indice unique n'est jamais la solution au problème provoqué par le nombre important d'indicateurs : il est donc nécessaire de comparer les différents indices de genre. Or, la description des méthodologies utilise des terminologies différentes, ne décrit pas correctement les choix méthodologiques et reste souvent muette sur les sources potentielles d'erreur de mesure. Enfin, les développeurs d'indice ne sont que rarement explicites sur le concept global qu'ils cherchent à mesurer. Autre point important, l'utilisation des données sur le genre interroge sur la qualité des données. Si elles sont de qualité douteuse, la légitimité de leur utilisation pour des objectifs collectifs est questionnable. Les choix méthodologiques qui sous-tendent la construction des indices montrent souvent que ce que l'indice mesure est différent de ce qu'il prétend déterminer !

Les contributions précieuses apportées par les promoteurs des indices de genre doivent être reconnues. Ils constituent une ressource qui a permis de décrire les disparités entre les genres et de promouvoir les droits des femmes d'une manière plus efficace que ce qui était possible avant 1995. Avoir un indice empirique global est préférable à ne pas en avoir, même si les mesures actuelles souffrent de faiblesses méthodologiques. Cependant, la production d'indices multiples crée également un problème pour les utilisateurs : ils n'ont pas les moyens de les comparer entre eux.

2.2.1 Discrimination raciale

Avant de commencer, rappelons qu'au sens biologique, les races n'existent pas, et qu'au sens sociologique, les races sont construites par le racisme. Le racisme est ici un ensemble de mécanismes qui créent (ou perpétuent) des inégalités sur la base de la racialisation des groupes, avec un groupe "privilégié" qui sera favorisé, et un groupe "racisé" qui sera défavorisé. Compte tenu du lien formé entre le racisme et la perception des discriminations, il est naturel de commencer par cette variable protégée (même si elle n'est pas clairement définie pour l'instant). Historiquement, aux États-Unis, la notion raciale a été centrale dans les discussions sur la discrimination (Anderson

26. Professeur au Conservatoire National des Arts et Métiers à Paris

2004 propose un retour historique aux États-Unis).

Race et origine ethnique, par Elisabeth Vallet [27]

Depuis les débuts de la République, le recensement aux États-Unis permet d'identifier les individus selon de grandes catégories raciales et ethniques. En plus des catégories classiquement définies au niveau fédéral (Blancs/-caucasiens, Noirs/Afro-Américains, Asiatiques-Américains, Amérindiens/Autochtones d'Alaska, Autochtones d'Hawaï/Insulaires du Pacifique, Multiracial), le recensement décennal y ajoute la catégorie Hispaniques/Latino. Il est également prévu qu'à terme, la catégorie Moyen-Orient et Afrique du Nord soit prise en compte par le recensement.

Cette classification raciale, qui sert désormais de base à l'analyse statistique du pays trouve toutefois son ancrage dans l'esclavagisme et une classification visant à établir une forme de pureté raciale, lorsqu'ont été appliquées les règles « de la goutte de sang unique » et du « quantum de sang » (Villazor 2008) pour déterminer l'appartenance à des groupes raciaux, et qui ont contribué à former les groupes de populations concernées et leur propre perception d'elles-mêmes. De surcroît, la complexité de cette classification tient au fait qu'elle n'est pas toujours fondée sur des éléments constants. D'une part les catégories, telles qu'elles sont énoncées dans les formulaires, évoluent dans le temps. Ainsi, la nature du questionnaire et la manière dont les questions sont formulées, parce qu'elles évoluent, altèrent parfois le portrait global de la population sur une longue durée.

D'autre part, cette classification repose désormais sur une auto-identification des individus qui, elle aussi peut varier d'un recensement à l'autre. Ainsi certaines études ont montré des évolutions importantes dans l'auto-identification de certains individus, dont il faut tenir compte lorsque l'on utilise ces données. Selon Liebler *et al.* 2017 par exemple, des variations importantes ont été enregistrées pour de mêmes individus entre le recensement de 2000 et celui de 2010, particulièrement dans des catégories qui pouvaient rapporter plusieurs origines ethniques ou encore hispaniques. Ainsi, l'étude montre que la constance des réponses était plus marquée chez les blancs non hispaniques, les noirs et les asiatiques. L'instabilité d'identification est ainsi plus marquée parmi les personnes qui s'identifient comme autochtones, les insulaires du pacifique, les personnes à origines multiples, et hispaniques. Les raisons sont multiples, mais elles tiennent en partie à l'évolution de l'intégration des communautés immigrantes, ou encore au fait que le métissage peut amener la prévalence d'une identité sur l'autre dans le temps. Ainsi les analyses doivent inclure le fait que l'idée même de race / appartenance ethnique, construction sociale, doit être incluse comme telle dans l'analyse des données statistiques.

Comme le rappelle Wolff 2006, en 1896, Frederick L. Hoffman, actuaire chez Prudential Life Insurance, a publié un ouvrage démontrant, statistiques à l'appui, que le Noir américain n'était pas assurable (voir Hoffman 1896). Du Bois 1896 a noté avec ironie que le taux de mortalité des Noirs aux États-Unis n'était que très légèrement supérieur (mais comparable) à celui des citoyens blancs à Munich, en Allemagne, à la même époque. Mais surtout, la principale critique est d'avoir agrégé toutes sortes de données, empêchant une analyse plus fine d'autres causes de (possible) surmortalité (c'est aussi l'argument avancé par O'Neil 2016). À cette époque, aux États-Unis, plusieurs États adoptaient des lois anti-discrimination, interdisant de demander des primes différentes sur la base d'une information raciale. Par exemple, comme le souligne Wiggins 2013, au cours de l'été 1884, la législature de l'État du Massachusetts a adopté la loi visant à prévenir la discrimination des compagnies d'assurance-vie à l'égard des personnes de couleur. Cette loi empêchait les assureurs-vie opérant dans l'État de faire "*any distinction or discrimination between white persons and colored persons wholly or partially of African descent, as to the premiums or rates charged for policies upon the lives of such persons*". La loi exigeait également que les assureurs paient des indemnités complètes aux assurés afro-américains. C'est sur la base de ces lois que l'argument de non-assurabilité a été invoqué : assurer les Noirs au même tarif que les Blancs serait statistiquement inéquitable, soutenait Hoffman 1896, et ne pas assurer les Noirs était la seule manière de se conformer à la loi (voir aussi Heen 2009). Comme le raconte Bouk 2015

27. Professeure, directrice du centre Raoul Dandurand à Montréal, Canada

"*Industrial insurers operated a high-volume business; so to simplify sales they charged the same nickel to everyone. The home office then calculated benefits according to actuarially defensible discriminations, by age initially and then by race. In November 1881, Metropolitan decided to mimic Prudential, allowing policies to be sold to African Americans once again, but with the understanding that black policyholders' survivors only received two-thirds of the standard benefit*".

Sur le marché du crédit, Bartlett *et al.* 2021, prenant la suite de Bartlett *et al.* 2018, montrent qu'aux États-Unis, la discrimination fondée sur l'origine ethnique a continué d'exister sur le marché hypothécaire américain (pour les Afro-Américains et les Latino-Américains), autant les prêts traditionnels que les prêts basés sur des algorithmes. Mais les algorithmes ont changé la nature de la discrimination, passant d'une discrimination basée sur les préjugés, ou l'aversion humaine, à des applications illégitimes de discrimination statistique. De plus, les algorithmes font de la discrimination non pas en refusant des prêts, comme le font les prêteurs classiques, mais en fixant des prix ou des taux d'intérêt plus élevés. En santé, Obermeyer *et al.* 2019 montrent qu'il existe des discriminations fondées sur l'ethnicité ou sur des préjugés "raciaux", dans un logiciel commercial largement utilisé pour affecter les patients nécessitant des soins médicaux intensifs à un programme de gestion des soins. Les patients blancs étaient plus susceptibles d'être affectés au programme de soin que les patients noirs dans un état de santé comparable. L'affectation a été effectuée à l'aide d'un score de risque généré par un algorithme. Le calcul comprenait des données sur les dépenses médicales totales d'une année donnée et des données fines sur l'utilisation des services de santé au cours de l'année précédente. Le score ne reflète donc pas l'état de santé attendu, mais prédit le coût des traitements. Les préjugés, les stéréotypes préjugés et incertitudes cliniques de la part des prestataires de soins de santé peuvent contribuer aux disparités raciales et ethniques dans les soins de santé, comme le soulignait Nelson 2002. Enfin, en assurance automobile, Heller 2015 a révélé que les quartiers à prédominance afro-américaine paient 70 % de plus, en moyenne, pour les primes d'assurance automobile que les autres quartiers. En réponse, la Property Casualty Insurers Association of America avait répondu [28] en novembre 2015 que "*insurance rates are color-blind and solely based on risk*". Cette position est encore celle d'associations actuarielles aux États-Unis, pour lesquelles le questionnement sur les discriminations n'ont pas de sens. Larson *et al.* 2017 ont obtenu 30 millions de devis de primes, par code postal, pour les principales compagnies d'assurance à travers les États-Unis, et ont confirmé qu'un écart existait, même s'il était plus faible. Aussi, dans l'Illinois, les compagnies d'assurance facturaient en moyenne plus de 10 % de plus les primes de responsabilité civile automobile pour les codes postaux "majoritairement minoritaires" (au sens où le taux de personnes issues de minorités était le plus important [29]) que pour les codes postaux majoritairement blancs. Historiquement, comme le rappelle Squires 2003, de nombreuses institutions financières ont eu recours à une telle discrimination en refusant de desservir des zones géographiques à prédominance afro-américaine.

Si de telles analyses se sont multipliées récemment (Klein 2021 propose une revue relativement complète de la littérature), ce problème de discrimination raciale potentiel avait été analysé par Klein et Grace 2001, par exemple, qui proposaient de tenir compte de covariables corrélées avec la variable raciale, et qui montrait qu'il n'existe aucune preuve statistique de "*redlining*" géographique. Cette conclusion était cohérente avec l'analyse de Harrington et Niehaus 1998, et a été reprise par la suite par Dane 2006, Ong et Stoll 2007 ou Lutton *et al.* 2020 entre autres. Notons ici que le "*redlining*" n'est pas seulement associé à un critère associal, mais bien souvent à un critère

28. En ligne sur leur site, https://www.pciaa.net/pciwebsite/cms/content/viewpage?sitePageId=43349

29. Le terme "minorités" est utilisé pour désigner les personnes discriminées (et non pas d'éventuelles "minorités dominantes").

économique. Un cas récent de discrimination statistique fait actuellement l'objet d'une enquête en Belgique, comme le mentionne Orwat 2020. Dans ce pays, le fournisseur d'énergie EDF Luminus refuse de fournir de l'électricité aux personnes vivant dans une certaine zone de code postal. Pour le fournisseur d'énergie, cette zone de code postal représente une zone où se trouvent de nombreuses personnes ayant de mauvaises habitudes de paiement. Le "*redlining*", qui est basé sur la variable de substitution "lieu de résidence", a été baptisé ainsi parce qu'il encercle des zones avec des lignes rouges comme le notaient Barocas et Selbst 2016.

Si le terme "*statistiques ethniques*" est un sujet sensible en France, les recensements demandent, traditionnellement (depuis plus d'un siècle), la nationalité à la naissance, distinguant ainsi Français de naissance et Français par acquisition. Et, depuis 1992, la variable *pays de naissance des parents* est introduite dans un nombre croissant d'enquêtes publiques. Dans la statistique française, le mot *ethnique*, au sens anthropologique (groupes humains infra-nationaux ou supra-nationaux dont l'existence est attestée bien qu'ils n'aient pas d'État), a sa place depuis longtemps, avec en particulier les enquêtes sur les migrations entre l'Afrique et l'Europe. Néanmoins, dans les textes juridiques, dans lesquels il constitue parfois un substitut euphémique de *racial*. La loi Informatique et Libertés de 1978 utilise ainsi l'expression *origines raciales ou ethniques*. Est *ethnique*, en ce sens, toute référence à une origine étrangère, qu'il s'agisse de nationalité à la naissance, de nationalité des parents, ou de « reconstitutions » fondées sur le patronyme ou sur l'apparence physique. Certains fichiers des renseignements généraux et de la police judiciaire contiennent des informations personnelles sur les caractéristiques physiques des personnes, et en particulier sur leur couleur de peau, comme le rappelle Debet 2007. Certains fichiers de recherche médicale (par exemple en dermatologie) peuvent contenir des informations similaires. L'INSEE avait d'abord refusé d'introduire une question sur le pays de naissance des parents, dans son enquête Famille de 1990 qui aurait pu servir de base de sondage. Il faudra attendre l'enquête de 1999 pour que la question soit enfin posée explicitement, comme le rappelle Tribalat 2016. Un autre soucis qui peut se poser est que la différence qui peut exister sur les primes d'assurance entre origines ethniques n'est pas le reflet de risques différents, mais de traitements différents. Hoffman, Trawalter *et al.* 2016 montrent ainsi que les préjugés raciaux et les fausses croyances sur les différences biologiques entre les Noirs et les Blancs continuent de façonner la façon dont nous percevons et traitons les Noirs - elles sont associées aux disparités raciales dans l'évaluation de la douleur et les recommandations de traitement.

2.2.2 Discrimination par le genre

Un point de terminologie s'impose ici, pour distinguer le sexe et le genre.

Sexe et genre, par Émilie Biland-Curinier [31]

Il est classique de différencier le **sexe**, qui est une caractéristique biologique (en lien avec des caractéristiques physiques et physiologiques, par exemple les chromosomes, l'expression génétique, les niveaux d'hormones et la fonction hormonale, etc.) et le **genre**, qui désigne l'identité sexuelle d'individu. On décrit souvent le sexe et le genre en termes binaires (fille/femme ou garçon/homme). Pourtant la diversité du développement sexué et des formules atypiques est importante, que celles-ci soient d'origine chromosomique, hormonale ou environnementale. En réalité, les débats sexe / genre sont nombreux, et se réfractent dans de nombreux champs du savoir et de l'action publique.

En sciences sociales, le sexe est aujourd'hui moins considéré comme une "réalité biologique" que comme une construction sociale, et surtout légale : le sexe est celui qui est attribué à chaque individu sur son acte de naissance et ensuite sur l'ensemble de ses "papiers" plus ou moins officiels. La plupart des gens conservent ce sexe natif et légal toute leur vie mais certains en changent (une bonne partie des personnes "trans", qu'elles aient ou non procédé à une chirurgie génitale). Dans le cas de personnes inter-sexe / inter-sexuées, l'attribution du sexe légal comporte une bonne part d'arbitraire (ou plus sociologiquement, de pouvoir discrétionnaire médical), puisque les attributs biologiques ne correspondent à aucune des catégories binaires féminin / masculin. La plupart des pays fondent cette catégorie légale de sexe sur cette dualité, mais certains ont récemment ouvert d'autres possibles (exemple : possible de renseigner X aux Pays-Bas, "divers" en Allemagne). Dans ces pays, on parle de sexe / genre neutre, dans d'autres de troisième neutre.

En statistique, la catégorie "sexe" est celle qui est surtout utilisée. Cette catégorie est effectivement déclarative : elle correspond le plus souvent au sexe légal, mais dans le cas où on demande aux individus de la renseigner (exemple : recensement), il peut y avoir des discordances (très minoritaires). Fait intéressant : Statistique Canada [30] a récemment adapté ses catégories pour tenir compte de l'identité de genre (c'est-à-dire pour rendre visible les personnes trans).

Dès lors, c'est principalement à partir de la variable sexe que l'on analyse quantitativement les rapports de genre en sociologie et en économie (et en particulier les inégalités entre femmes et hommes). Pour une brève présentation de cet enjeu dans le contexte français Grobon et Mourlot 2014, sinon Amossé et De Peretti 2011. Enfin, le définition du genre / gender en temps que concept des sciences sociales, on peut citer ces phrases de la philosophe Elsa Dorlin : "*Traduction du terme gender, le concept de genre a permis d'historiciser les identités, les rôles et les attributs symboliques du féminin et du masculin, les définissant, non seulement comme le produit d'une socialisation différenciée des individus, propre à chaque société et variable dans le temps, mais aussi comme l'effet d'une relation asymétrique, d'un rapport de pouvoir. En ce sens, le rapport de genre peut être défini avec Joan Scott de la façon suivante :* Le genre est une façon première de signifier des rapports de pouvoir. *Les catégories du masculin et du féminin, comme les « hommes » et les « femmes » n'ont donc de sens et d'existence que dans leur rapport antagonique et non pas en tant qu'« identités » ou en tant qu'« essences » prises isolément*", Dorlin 2005.

La directive européenne sur les biens et services de 2004, Conseil de l'Union Européenne 2004, visait à réduire les écarts entre les sexes dans l'accès à tous les biens et services, discutée par exemple par Thiery et Van Schoubroeck 2006. Une dérogation spéciale à l'article 5, paragraphe 2, permettant aux assureurs de fixer des prix fondés sur le sexe pour les hommes et les femmes. En effet, "*les États membres peuvent décider (⋯) d'autoriser des différences proportionnelles dans les primes et les prestations des particuliers lorsque l'utilisation du sexe est un facteur déterminant dans l'évaluation du risque, sur la base de données actuarielles et statistiques pertinentes et précises*". Autrement dit, cette clause permettait une exception pour les compagnies d'assurance, à condition qu'elles fournissent des données actuarielles et statistiques qui permettent d'établir

31. Professeure à Science-Po Paris
30. En ligne sur https://www.statcan.gc.ca/fra/concepts/definitions/variable-genre-sexe

que le sexe est un facteur objectif d'évaluation du risque. La Cour de justice de l'Union européenne a annulé cette exception juridique en 2011, dans un arrêté longuement discuté par Schmeiser *et al.* 2014 ou Rebert et Van Hoyweghen 2015, par exemple. Ce règlement, qui a suscité nombre de commentaires en Europe en 2007 puis en 2011, avait aussi soulevé beaucoup de questions aux États-Unis, plusieurs décennies auparavant, comme cette discussion à la fin des années 70, avec G. D. Martin 1977, Hedges 1977 et Myers 1977. Par exemple, dans *City of Los Angeles, Department of Water and Power v. Manhart*, la Cour suprême a examiné un système de retraite dans lequel les employées versaient des cotisations plus élevées que les hommes pour la même prestation mensuelle en raison d'une espérance de vie plus longue. La majorité a finalement déterminé que ce régime violait le titre VII de la loi sur les droits civils de 1964, car il supposait que les individus se conformeraient aux tendances plus larges associées à leur sexe. Une telle discrimination, a suggéré le tribunal, est troublante du point de vue des droits civils car elle ne traite pas les individus comme des individus, par opposition à de simples membres des groupes auxquels ils appartiennent. Ces lois étaient motivées, en partie, par le fait que les décisions en matière d'emploi sont généralement individuelles : une personne spécifique est embauchée, licenciée ou rétrogradée, en fonction de sa contribution passée ou attendue à la mission de l'employeur. En revanche, les stéréotypes sur les individus basés sur les caractéristiques du groupe sont généralement davantage tolérés dans des domaines comme l'assurance, où la prise de décision individualisée n'a pas de sens.

On peut aussi mentionner les (possibles) discriminations contre des personnes transsexuelles. En effet, comme le notait Jacobs et Sommers 2015, des assureurs aux États-Unis ont tenté d'inférer des maladies sur la base de prescription de médicaments. Comment interpréter la décision d'un assureur qui demande une majoration des primes pour des garanties en cas de décès, d'incapacité de travail et d'invalidité sous prétexte que l'assuré s'est vu prescrire un traitement hormonal pour une dysphorie de genre ?

2.2.3 Discrimination par l'âge

Aux États-Unis, l'idée que l'âge peut constituer un motif de discrimination s'est traduit par le "*Age Discrimination in Employment Act*" en 1967, voté en prenant la suite du "*Civil Rights Act*" de 1964, qui se concentrait surtout sur les aspects éthiques et raciaux, comme le montre Macnicol 2006. Dans la majorité des cas, la discrimination par l'âge est envisagée sous l'angle de l'emploi, comme le précisent Duncan et Loretto 2004 ou Adams 2004. On peut noter que certains assureurs font du refus de discriminer suivant l'âge un point important, une forme de "raison d'être" (au sens donné par la loi Pacte de 2019). Ainsi, en France, "*la Mutuelle Générale s'engage à* (·) *renforcer la solidarité entre les génération*", ce qui devrait se traduire par un refus de discriminer, de segmenter, suivant ce critère.

Mais tout comme une distinction existe entre le sexe biologique et le genre, certains suggèrent de distinguer entre l'âge biologique et l'âge perçu (ou subjectif), comme Stephan *et al.* 2015 ou Kotter-Grühn *et al.* 2016. Uotinen *et al.* 2005 montraient que cet âge subjectif serait un meilleur prédicteur de la mortalité que l'âge biologique. Comme le souligne Beider 1987, on peut faire valoir que si les gens n'ont pas une chance équitable en fonction de leur âge, car tout le monde ne vieillit pas de manière identique, les gens mourant à des âges différents. Bidadanure 2017 rappelle que la discrimination suivant l'âge est toujours perçue comme moins "préoccupante" que la plupart des autres. Le processus de vieillissement, de la naissance à l'âge adulte, est corrélé à divers processus de développement cognitifs qui font qu'il est pertinent d'attribuer une responsabilité, une capacité de consentement et une autonomie différentes aux enfants, aux jeunes adultes ou

aux personnes âgées. Mais contrairement au sexe et à la race, l'âge n'est pas une caractéristique discrète et immuable. Comme le dit Macnicol 2006, l'âge n'est pas un club dans lequel on naît. Nous nous attendons à passer par les différentes étapes d'une vie et la vieillesse est un club que nous savons que nous rejoindrons très probablement un jour. Par conséquent, le traitement différentiel en fonction de l'âge ne génère pas nécessairement des inégalités entre les personnes au fil du temps, alors que le traitement différent en fonction de l'origine ethnique et du sexe le fait : "*une société qui discrimine sans relâche les gens en raison de leur âge peut encore les traiter de manière égale tout au long de leur vie* (⋯). *Le tour de chacun [d'être discriminé] vient*" affirmait Gosseries 2014. Citant un arrêt de la cour d'appel de 2008, Mercat-Bruns 2020 rappelle que "*le législateur a pris soin d'opérer une distinction entre l'âge et l'état de santé, il ne peut dès lors être procédé à un amalgame entre ces deux motifs en considérant que l'âge avancé induit nécessairement une santé défaillante*".

En assurance automobile, comme rappelé par Cooper 1990, Liisa 1994 et Clarke *et al.* 2010, l'augmentation de la sinistralité chez les personnes âgées s'explique par une perte d'acuité sensorielle et motrice, par une consommation de médicaments (en particulier des psychotropes), et une diminution des réflexes. Mais les personnes âgées ont aussi tendance à moins conduire, comme le souligne Fontaine 2003. Sur la Figure 2.4, on peut visualiser la fréquence d'accidents et le nombre de décès, par millions de kilomètres parcourus, pour différentes classes d'âge. Le risque de blessure corporelle (ou de décès) dans un accident de voiture augmente significativement dès 60 ans et ne cesse d'augmenter rapidement avec l'âge, comme le montrent Li *et al.* 2003, entre autres.

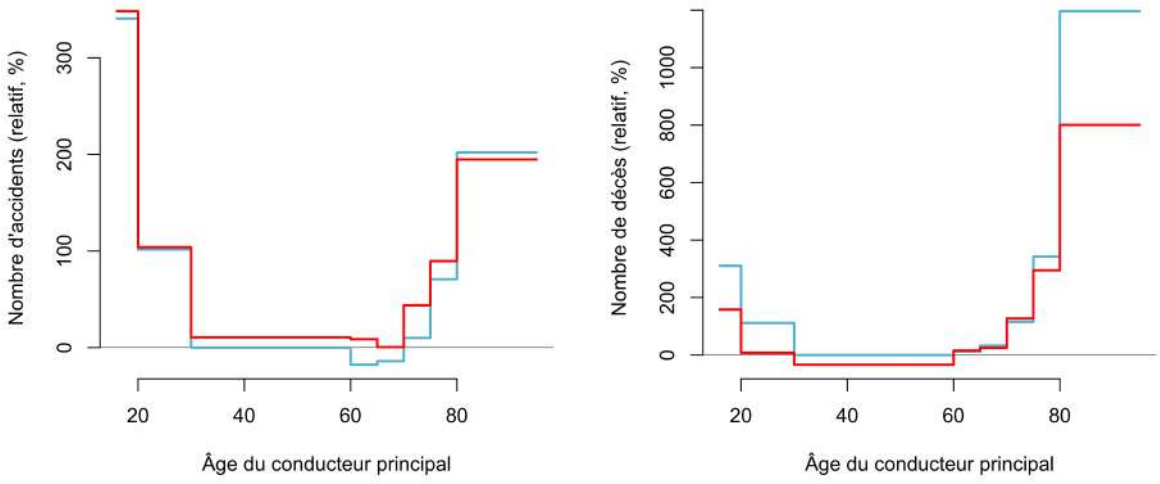


Figure 2.4 – Nombre d'accidents (à gauche) et nombre de décès (à droite), par million de kilomètres parcourus, pour les hommes et les femmes, en fonction de l'âge du conducteur. La référence (0) sont les hommes de 30-60 ans. Le nombre d'accidents est trois fois plus élevé (+200 %) pour les plus de 85 ans, et le nombre de décès plus de dix fois plus élevés (+900 %) (source : Li *et al.* 2003).

Mais dans la majorité des pays, le risque moyen des personnes âgées ne semble pas particulièrement important (tant que les personnes âgées sont vues comme un groupe homogène). La Figure 2.5 montre l'évolution de la fréquence (annuelle) de sinistres, le coût moyen des sinistres et la

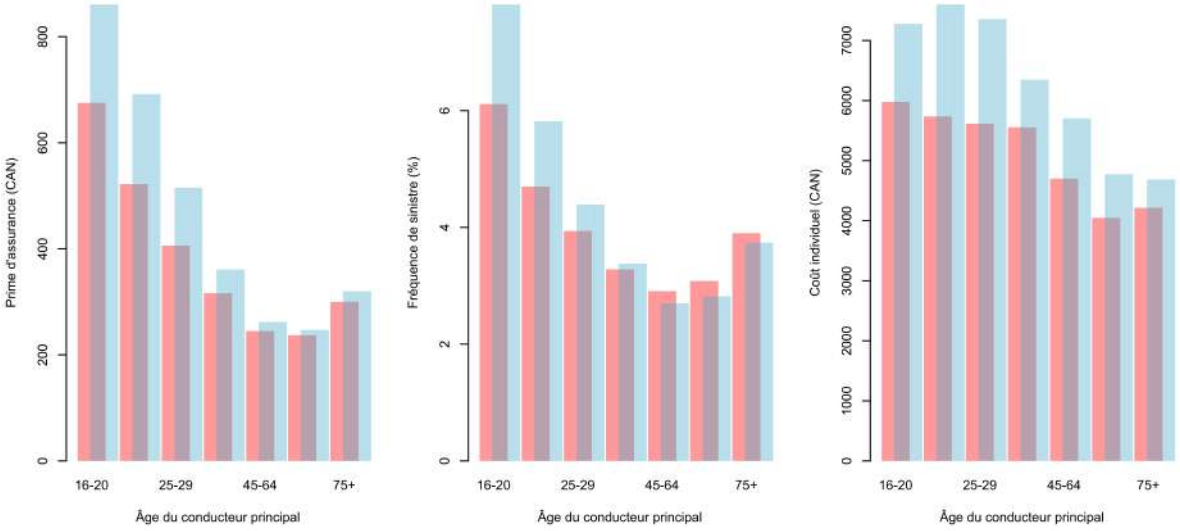


Figure 2.5 – De gauche à droite, prime souscrite moyenne (dollars canadiens), fréquence des sinistres, et coût moyen des sinistres (dollars canadiens), par tranche d'âge (axe des abscisses) et par genre (hommes en bleu et femmes en rouge) au Québec (source Groupement des Assureurs Automobiles 2021).

prime moyenne en assurance automobile, au Québec, en fonction de l'âge des assurés (par tranche d'âge), pour la protection "collision et versement".

Pour aller plus loin, Dulisse 1997, Meuleners *et al.* 2006 et Cheung et McCartt 2011 notent que la part de responsabilité dans les accidents augmente aussi avec l'âge, avec en particulier davantage d'accidents sur le côté droit, traduisant souvent un refus de priorité. De nombreux pays se sont posés la question de la régulation en lien avec les âges très avancés (au-delà de 80 ans). En matière de handicap, les assureurs n'ont pas le droit de discriminer suivant un handicap si la personne a le droit de conduire. Mais pour les maladies dégénératives, rares sont les législations qui interdisent explicitement de conduire, par exemple pour une personne ayant une maladie établie, comme la maladie de Parkison, Crizzle *et al.* 2012. Le fait que les personnes âgées soient davantage responsables des sinistres pose de nombreuses questions morales, car se mettre soi-même en danger en tant que conducteur est une chose, mais potentiellement en blesser voire en tuer d'autres est moins acceptable.

2.2.4 Discrimination par le poids

Parallèlement à l'augmentation des taux de surpoids et d'obésité dans la population, on observe une augmentation quelque peu contre-intuitive des préjugés et de la discrimination à l'égard des personnes perçues comme étant grosses (Latner et Stunkard 2003, Puhl *et al.* 2008, Andreyeva *et al.* 2008 ou encore Sutin et Terracciano 2013). Le stéréotype "gros, c'est mal" existe dans le domaine médical depuis des décennies, comme le rappelle Nordholm 1980. Une étude plus approfondie est nécessaire pour vérifier dans quelle mesure cela affecte la pratique. Il semble que les personnes obèses, en tant que groupe, évitent de rechercher des soins médicaux en raison de leur poids. Cependant, un obstacle à la formulation de conclusions plus poussées est que la

plupart des recherches reposent sur des mesures d'auto-évaluation dont la fiabilité et la validité sont variables. Il est nécessaire d'aller au-delà des rapports d'attitudes pour s'intéresser aux pratiques réelles en matière de soins de santé.

Pour Czerniawski 2007, "*with the rise of actuarial science, weight became a criterion insurance companies used to assess risk. Used originally as a tool to facilitate the standardization of the medical selection process throughout the life insurance industry, these tables later operationalized the notion of ideal weight and became recommended guidelines for body weights*". À la fin des années 50, 26 compagnies d'assurance ont coopéré, afin de déterminer la mortalité des assurés en fonction de la corpulence des assurés, comme le montre Wiehl 1960. La conclusion est claire en ce qui concerne la mortalité : "*studies bring out the clear-cut disadvantage of overweight—mortality ratios rising in every instance with increase in degree of overweight*". C'est aussi ce que disait Baird 1994, quarante ans après "*obesity is regarded by insurance companies as a substantial risk for both life and disability policies. This risk increases proportionally with the degree of obesity*" (la même conclusion se retrouve chez Lew et Garfinkel 1979 ou Must *et al.* 1999).

Bien que les médecins reconnaissent les risques de l'obésité pour la santé et qu'ils perçoivent un grand nombre de leurs patients comme étant en surpoids, ils n'interviennent pas autant qu'ils le devraient, sont ambivalents quant à la façon de gérer les clients obèses et sont peu susceptibles d'orienter officiellement un client vers un programme de perte de poids. Puhl et Brownell 2001 notent que seuls 18 % d'entre eux ont déclaré qu'ils discuteraient de la gestion du poids avec leurs patients en surpoids, ce pourcentage passant à 42 % pour les patients légèrement obèses. D. R. Black *et al.* 1994 ont montré que l'obésité était directement et significativement liée à des coûts de soins de santé plus élevés (un coût supérieur de 8 %), même en tenant compte de l'âge, du sexe et d'un certain nombre de conditions chroniques. Comme l'ont montré Quesenberry *et al.* 1998, par rapport aux personnes ayant un IMC/BMI de 20 à 24,9 kg/m^2, il y avait une augmentation de 25 % à 44 % des coûts annuels chez les personnes en surpoids modéré et sévère, après ajustement pour l'âge et le sexe. Gibbs 1995 note qu'il est courant que les régimes d'assurance maladie excluent explicitement de leur couverture le traitement de l'obésité. En 1998, l'Internal Revenue Service a exclu les programmes de perte de poids de la déduction médicale, même lorsqu'ils sont prescrits par un médecin. Yusuf *et al.* 2005 ont montré une corrélation modeste entre l'indice de masse corporelle (IMC, ou BMI) et le risque d'infarctus du myocarde, alors qu'elle était plus importante avec le rapport tour de taille/hanche. Bien que cette mesure dépende fortement du sexe, il est possible de l'utiliser, comme l'indique explicitement [31] la Commission Européenne : "*L'obésité est un facteur de risque, mesuré par le rapport entre le tour de taille et le tour de hanche, qui est différent pour les femmes et pour les hommes*".

2.2.5 Discrimination des fumeurs

Avant de parler du tabac, rappelons qu'il existe pour le Centre international de recherche sur le cancer (CIRC) cinq grands groupes d'agents : (1) cancérigènes certains (plutonium 239, radium, amiante, butane, gaz moutarde, etc), (2A) probablement cancérigènes (insecticides non arsenicaux, lampes et tables à bronzer, combustion domestique de biomasse (bois), etc.), (2B) peut-être cancérigènes (essence, gaz d'échappement des moteurs à essence, etc.), (3) inclassables (pétrole brut, encres d'imprimerie, etc), (4) probablement pas cancérigènes. Les dangers du tabac sont observés dès le début du XVII^{ème}siècle, et sa cancérogénicité (acteur provoquant, aggravant ou favorisant l'apparition d'un cancer) est soupçonnée au XVIII^{ème}siècle, avant d'être largement

31. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2012:011:0001:0011:FR:PDF

admise au milieu du XIX^{ème}siècle. Toutefois, la responsabilité du tabagisme dans la genèse des cancers (en particulier du poumon) a été longue à établir. Le rôle cancérigène du tabac a été également suspecté dès le lendemain de la Première Guerre mondiale. Patterson 1987 rappelle que dès 1930, les assureurs avaient observé que le tabac était en lien avec certains cancers. En particulier, Frederick L. Hoffman avait commencé à collecter des statistiques à partir de 1915. Hoffman 1931 affirmait ainsi "*Smoking habits unquestionably increase the liability to cancer of the mouth, the throat, the cesophagus, the larynx and the lungs*", tout en soulignant l'incroyable hétérogénéité du groupe "fumeur" (tant sur la quantité que sur la qualité du produit fumé).

Pour Johnston 1945, "*it is clear that smoking is an important cause of mortality*", présentant des tables de mortalités comparant non-fumeurs et fumeurs ("*moderate*" et "*heavy*"). Après la Seconde Guerre Mondiale, Richard Doll (discuté rétrospectivement dans Doll et A. B. Hill 1964 puis 2004), puis des études de grande envergure dans les années 1950 et 1960, ont confirmé bon nombre de liens. En 1954 était lancée une vaste étude prospective dont les résultats, vingt ans plus tard, confirmaient l'association entre consommation chronique de tabac et réduction de l'espérance de vie. Fisher 1958 et 1959 tente l'explication suivante : "*there can therefore be little doubt that the genotype exercises a considerable influence on smoking*", autrement dit, le fait de fumer serait expliqué, génétiquement (Cook 1980 et Stolley 1991 analyseront les données utilisées par Ronald Fisher et arriveront à des conclusions assez différentes). Ferrence 1990, Kluger 1999 ou Brandt 2007 reviennent sur l'histoire des liens entre le tabagisme et la santé. Dans un contexte actuariel, Benjamin et Michaelson 1988 écrivaient "*it seems clear that the heavier mortality resulting from cigarette smoking has not yet as fully caught up with women as it has with men and as, undoubtedly, it is beginning to*". Comme mentionné en introduction, Miller et Gerstein 1983 mentionnent les premières tables de mortalité pour les fumeurs.

2.3 Les proxys, corrélation fallacieuse et stéréotypes

En Statistique, comme l'expliquent Upton et Cook 2014, un "proxy" (on parlera aussi de "variable de substitution") est une variable qui, dans un modèle prédictif, remplace une variable utile mais non observable, non mesurable, ou, dans notre cas, qui ne peut être utilisée car jugée "discriminatoire". Et pour qu'une variable soit un bon proxy, elle doit avoir une bonne corrélation, avec la variable d'intérêt. Un exemple relativement populaire est le fait qu'à l'école primaire, la pointure de chaussure est souvent un bon proxy des capacités de lecture. En réalité, la taille des pieds n'a pas grand chose à voir avec les capacités cognitives, mais chez les enfants, la taille des pieds est très liée à l'âge, qui est lui même très lié aux capacités cognitives. Comme nous le verrons dans la section 4.4.3, ce concept est très lié aux notions de causalité.

Nous avons mentionné plus tôt la vision des économistes sur la discrimination, comme Gary Becker, et le lien avec une forme de rationalité et d'efficacité. Et en effet, pour de nombreux auteurs, ce qui compte, c'est que l'association entre des variables soit suffisamment forte pour constituer un prédicteur fiable. Pour Norman 2003, l'appartenance à un groupe fournit des informations fiables sur le groupe, et par extension sur tout individu qui en est membre, l'utilisation systématique de ces informations (la généralisation et les stéréotypes dont parlent Schauer 2006 et Puddifoot 2021) peut être économiquement efficace. Prenant un contre-pied éthique, Greenland 2002 rappelle que certaines sources d'information devraient être exclues de notre prise de décision parce qu'elles sont non pertinentes, ou non causales, même si elles peuvent fournir des informations assez fiables en raison de leur corrélation forte avec un autre indicateur. L'argument central est que si des variables

Métier	part de femmes
Aides à domicile et aides ménagers et assistants maternels	97.7 %
Secrétaires	97.6 %
Aides-soignants	90.4 %
Infirmiers, sages-femmes	87.7 %
Conducteurs de véhicules	10.5 %
Techniciens et agents de maîtrise de la maintenance	8.9 %
Techniciens et agents de maîtrise du bâtiment et des travaux publics	7.9 %
Ouvriers qualifiés du gros œuvre du bâtiment	2.1 %

Table 2.4 – Part de femmes, par métier, INSEE, 2011.

sont non causales, alors elles sont dépourvues de justification morale. Et la discrimination par proxy pose des questions éthiques complexes. Comme le notait Birnbaum 2020, "*if discriminating intentionally on the basis of prohibited classes is prohibited – e.g., insurers are prohibited from using race, religion or national origin as underwriting, tier placement or rating factors – why would practices that have the same effect be permitted?*". Autrement dit, on ne peut pas se contenter de comparer les primes payées, mais le processus narratif de la modélisation (c'est-à-dire les notions d'interprétabilité et d'explicabilité, ou la "fable" dont nous parlions en introduction) est tout aussi important, pour juger si un modèle est discriminatoire, ou pas. Et la difficulté, comme le disait Seicshnaydre 2007, c'est qu'il ne s'agit pas de chercher un preuve d'une intention, ou d'une motivation raciste, mais d'établir qu'un algorithme discrimine suivant un critère interdit (car protégé).

2.3.1 Quelques exemples

Un exemple relativement simple et classique en France est la profession, qui peut être, dans certains cas, un proxy du genre, comme le montre la Table 2.4,

Obermeyer *et al.* 2019 racontent qu'en 2019, une grande entreprise de soins de santé avait utilisé les dépenses médicales comme un proxy pour la gravité de la condition médicale. L'utilisation de ce proxy a donné lieu à un algorithme discriminatoire sur le plan racial, car si la maladie n'est pas forcément discriminatoire sur le plan racial, alors que les dépenses de santé le sont (aux États-Unis au moins). Plus généralement, toutes sortes de proxys sont utilisés, plus ou moins corrélées à la variable d'intérêt. Par exemple le revenu d'une personne (ou d'un ménage) sera approché par le montant d'impôt sur le revenu, ou par le montant de sa maison (ou du quartier où la personne réside). Kraemer *et al.* 2001 parlera de "facteurs de risques indirect". Nous reviendrons dans la section 4.4 sur l'importance des graphes causaux pour comprendre si une variable en cause une autre, ou si elle est simplement corrélée à cette dernière.

Il existe aussi certaines grandeurs, essentielles pour modéliser des prises de décisions dans un contexte incertain, mais difficiles à mesurer. C'est le cas du concept abstrait qu'est l'"aversion au risque" (largement discuté par Menezes et D. L. Hanson 1970 et Slovic 1987). Hofstede 1995 propose un indice d'évitement de l'incertitude (UAI, "*uncertainty avoidance index*"), calculé à partir de données d'enquêtes. Outreville 1990 et 1996 a suggéré d'utiliser le niveau d'éducation pour

évaluer l'aversion au risque. Selon Outreville 1996, l'éducation favorise une compréhension du risque et donc une augmentation de la demande d'assurance, par exemple (même si une relation inverse pourrait exister, si on suppose que l'augmentation des niveaux d'éducation est associée à une augmentation du capital humain transférable, ce qui induit une plus grande prise de risque).

Pour l'anecdote, on a évoqué l'indice IMC/BMI comme proxy de d'obésité (à partir d'un certain seuil), variable elle-même en lien avec certaines pathologies, comme le diabète de type 2, une hypertension artérielle, l'athérosclérose, la dyslipidémie, des maladies du foie ou des maladies rénales. Mais certains assureurs ont tenté d'utiliser d'autres proxys possibles collectés via des questionnaires de santé, par exemple en demandant "en combien de temps mangez-vous, en moyenne, le midi ?".

2.3.2 La discrimination statistique par proxy

Certaines variables non-protégées sont très corrélées avec des variables protégées. Par exemple, en France, la population active est à 48.3 % féminine, alors que la population travaillant à temps partiel est à 79.5 % féminine (selon Dares 2020). Autrement dit, savoir qu'une personne travaille à temps partiel permet d'affirmer qu'il y quatre fois plus de chance que ce soit une femme qu'un homme. En utilisant des statistiques sur les prénoms [32] on sait que "Brigitte" a, en moyenne 62 ans, "Jean-Pierre", 69 ans et "Arthur", 15 ans. Espeland et Stevens 1998 parleront de "*comesures*" ("*commensuration*") dans le sens où une variable (comme la profession ou le prénom) est utilisée à d'autres fins, correspondant à une "*transformation of different qualities into a common metric*".

"*Humans think in stories rather than facts, numbers or equations — and the simpler the story, the better*" affirmait Harari 2018, mais pour les assureurs, c'est souvent un mélange des deux. Pour Glenn 2000, comme le dieu romain Janus, le processus de sélection des risques d'un assureur possède deux visages : celui qui est présenté aux régulateurs et aux assurés, et l'autre qui est présenté aux souscripteurs. Il y a d'un côté le visage des chiffres, des statistiques et de l'objectivité. De l'autre, il y a le visage des récits, du caractère et du jugement subjectif. La rhétorique de l'exclusion de l'assurance - les chiffres, l'objectivité et les statistiques - forme ce que Brian Glenn appelle "*the myth of the actuary, a powerful rhetorical situation in which decisions appear to be based on objectively determined criteria when they are also largely based on subjective ones*" ou "*the subjective nature of a seemingly objective process*". Et pour Daston 1992, cette prétendue "*objectivité*" du processus est fausse, et dangereuse, comme le soulignait également Desrosières 2016. Glenn 2003 affirmait "*there are many ways to rate accurately, and this is my very point. Insurers can rate risks in many different ways depending on the stories they tell about which characteristics are important and which are not* (⋯) *The fact that the selection of risk factors is subjective and contingent upon narratives of risk and responsibility has in the past played a far larger role than whether or not someone with a wood stove is charged higher premiums*". En allant plus loin, "*virtually every aspect of the insurance industry is predicated on stories first and then numbers*". On se souvient du "*all models are wrong but some models are useful*" de Box *et al.* 2011, autrement dit, tout modèle est au mieux une fable utile.

Stéréotypes et généralisation

Comme le rappelle Bernstein 2013, le mot "stéréotype" fusionne un adjectif grec signifiant solide, στερεός, avec un substantif désignant un moule, τυπός. En combinant les deux termes, le mot

32. https://www.ekintzler.com/projects/age-prediction/

désigne un moulage dur, quelque chose qui peut laisser une marque, qui a donné, un terme d'imprimerie, à savoir la forme imprimante utilisée pour l'impression typographique. En 1802, le dictionnaire de l'Académie française mentionne, pour le mot "stéréotype", "*mot nouveau qui se dit des livres stéréotypés, ou imprimés avec des formes ou planches solides*". Le journaliste et intellectuel public américain Walter Lippmann a donné à ce mot sa signification contemporaine dans Lippmann 1922. Pour Lippmann, il s'agissait de décrire comment les êtres humains font entrer "le monde extérieur" dans "les images que nous avons dans la tête" ("*the World outside and the Pictures in our heads*"), des images qui vont former des catégories descriptives simplifiées par lesquelles nous cherchons à situer autrui ou des groupes d'individus. Lippmann a tenté d'expliquer comment les images qui surgissent spontanément dans l'esprit des gens se concrétisent. Les stéréotypes, a-t-il observé, sont "la plus subtile et la plus envahissante de toutes les influences" ("*subtlest and most pervasive of all influences*"). Quelques années plus tard, Rice 1926, a commencé les premières expériences pour mieux cerner ce concept.

L'importance des stéréotypes pour comprendre bon nombre de prises de décision sera analysée en détails dans Kahneman 2011, inspiré en grande partie de Bruner *et al.* 1957, et plus récemment, Hamilton et Gifford 1976 et surtout Devine 1989. Pour Daniel Kahneman, schématiquement, deux types de mécanismes sont utilisés pour prendre une décision. Le système 1 est utilisé pour la prise de décisions rapides : il nous permet de reconnaître les gens et les objets, nous aide à orienter notre attention, et nous encourage à craindre les araignées. Il est basé sur des connaissances stockées en mémoire et accessibles sans intention, et sans effort. On peut l'opposer au système 2, qui permet une prise de décision dans un contexte plus complexe, exigeant de la discipline et une réflexion séquentielle. Dans le premier cas, notre cerveau utilise les stéréotypes qui régissent les jugements de représentativité, et utilise cette heuristique pour prendre des décisions. Si je cuisine un poisson pour des amis venus manger, j'ouvre une bouteille de vin blanc. Le cliché "*le poisson se marie bien avec le vin blanc*" me permet de prendre une décision rapidement, sans avoir à réfléchir. Les stéréotypes sont des affirmations concernant un groupe qui sont acceptées (du moins provisoirement) comme des faits concernant chaque membre. Qu'ils soient corrects ou justes, les stéréotypes sont l'outil de base pour penser les catégories dans le "système 1". Mais bien souvent, une réflexion plus poussée, plus construite – correspondant au "système 2" – permettra de prendre une décision plus judicieuse, voire optimale. Sans choisir n'importe quel vin rouge, un pinot noir pourrait peut-être aussi parfaitement convenir aux rougets grillés. Comme l'affirmait Fricker 2007, "*stereotypes are [only] widely held associations between a given social group and one or more attributes*". N'est-ce pas ce que font les actuaires au quotidien ?

Généralisation et actuariat

Pour Schauer 2006, cette "généralisation" est probablement la raison d'être de l'actuaire : "*to be an actuary is to be a specialist in generalization, and actuaries engage in a form of decision-making that is sometimes called actuarial*". On retrouvera cette idée en justice actuarielle, chez Harcourt 2008 par exemple. Schauer 2006 rapportait que nous pourrions être amenés à croire qu'il est préférable d'avoir des pilotes de ligne qui ont une bonne vision, plutôt qu'une mauvaise (ce point est également évoqué dans le contexte de la conduite, et de l'assurance automobile, Owsley et McGwin Jr 2010). Ce critère pourrait être utilisé à l'embauche, et constituerait, bien sûr, une forme de discrimination, en distinguant les 'bons" pilotes des "mauvais", les pilotes qui ont une bonne vision des autres. Certaines compagnies aériennes pourraient imposer un âge maximum pour les pilotes de ligne (55 ou 60 ans, par exemple), l'âge étant un indicateur fiable, même s'il est imparfait, d'une mauvaise vision (ou plus généralement d'un mauvais état de santé, avec une

audition déficiente ou des réflexes plus lents). En excluant les personnes âgées de la fonction de pilote d'avion commercial, nous nous retrouverons, *ceteris paribus*, avec une cohorte de pilotes de ligne qui a une meilleure vision, une meilleure audition et des réflexes plus rapides. L'explication donnée ici est clairement causale, et les objectifs sous-jacents à la discrimination semblant alors clairement légitimes, même l'utilisation de l'âge devient alors une discrimination par proxy, au sens de Hellman 1998, que Schauer 2017 appelle discrimination statistique, faute de faire des tests d'état de santé des pilotes.

Pour Thiery et Van Schoubroeck 2006 (mais aussi Wortham 1985), les juristes et les actuaires ont des conceptions fondamentalement différentes de la discrimination et de la segmentation en assurance, l'une étant individuelle, et l'autre collective, telles que stigmatisées aux États-Unis par les affaires Manhart et Norris (Hager et Zimpleman 1982, Bayer 1986). Dans l'affaire Manhart, en 1978, la Cour a jugé illégal un régime de rentes dans lequel les hommes et les femmes recevaient des prestations égales à la retraite, même si les femmes versaient des cotisations plus importantes. En 1983, la Cour suprême a jugé, dans l'affaire Norris, que l'utilisation de facteurs actuariels variant selon le genre en termes de prestations dans le cadre de régimes de retraite était illégale, car elle tombait sous le coup de l'interdiction de la discrimination. Ces deux décisions sont l'affirmation juridique selon laquelle la technique d'assurance ne pourrait pas toujours servir de garantie pour justifier la différence de traitement entre les membres de certains groupes dans un contexte de segmentation de primes d'assurance. En effet, juridiquement, le droit à l'égalité de traitement est octroyé à une personne en sa qualité d'individu, et non en sa qualité de membre d'un groupe racial, sexuel, religieux ou ethnique. Et ainsi, un individu ne peut être traité différemment en raison de son appartenance à un tel groupe, en particulier à un groupe auquel il n'a pas choisi d'appartenir. Ces arrêtés soulignent qu'il est interdit de traiter les individus comme de simples composants d'une classe raciale, religieuse, ou sexuelle, affirmant que l'équité envers les individus l'emporte sur l'équité envers les classes. Mais cette vision s'oppose fondamentalement à l'approche actuarielle, qui, historiquement analyse les risques et calcule les primes en termes de groupes : jusqu'à peu, les actuaires ne considéraient les individus que comme des membres d'un groupe.

L'approche actuarielle est celle évoquée dans le premier paragraphe. Un individu appartenant à un groupe présentant un risque statistique de survie ou de décès plus élevé finit par payer une prime plus élevée ou par recevoir moins de prestations. Dans l'assurance automobile, un individu appartenant à un groupe qui représente un risque statistique d'accident plus élevé doit payer des primes plus élevées. Brilmayer *et al.* 1983 rappelaient que ce sont les différences entre les probabilités d'avoir un accident en fonction du genre (et non les différences individuelles) qui sont prises en compte pour justifier la différence de primes, pour expliquer la différence de prestations, ou pour fonder un mécanisme de sélection. Les systèmes de classification des assurances reposent sur l'hypothèse que les individus répondent aux caractéristiques moyennes, stéréotypées pour reprendre le terme de Schauer 2006, d'un groupe auquel ils appartiennent. Les assureurs font en effet valoir que les statistiques courantes indiquent qu'en moyenne, plus de femmes que d'hommes conduisent sans accident et que, par conséquent, la femme moyenne a une espérance de perte inférieure à celle de l'homme moyen. Sur la base de ces données, les femmes doivent payer une prime inférieure à celle des hommes. Les compagnies d'assurance visent à préserver l'égalité entre les groupes et non entre les individus, ainsi que les raisons pour lesquelles les assureurs pensent en termes de femme moyenne et d'homme moyen.

Un fondement essentiel de l'assurance est l'idée de *mutualisation* des risques, c'est-à-dire la constitution de groupes. Le risque en assurance ne peut s'envisager sans cette notion de

mutualisation, c'est d'ailleurs la grande différence avec les mathématiques financières, où il existe une valeur fondamentale d'un risque (dans un marché). La mutualisation est intrinsèque à la segmentation des risques d'assurance, et impose une forme de solidarité au sein du groupe, la totalité des primes d'un groupe devant être statistiquement entièrement compensée par la totalité des remboursements de ce même groupe. L'assureur impose alors une solidarité entre les assurés qui ont un même profil de risque (avec une probabilité de sinistre et une ampleur de sinistre comparables). On parle alors de solidarité de pure chance ("*luck solidarity*", Barry 2020). Sans segmentation, ou si les groupes qui ne sont pas composés de membres ayant un profil de risque comparable, on observera un phénomène de solidarité subventionnante, au sens où une personne ayant un certain profil de risque paie pour le montant de la perte des personnes ayant une espérance de perte plus élevée.

On retrouve l'opposition (pour reprendre la terminologie de Hume 1739) entre *is* et *ought*, entre ce qui est, et ce qui devrait être, entre la norme statistique de l'actuaire et la norme du législateur, dont nous parlions dans l'introduction. D'un point de vue empirique, descriptif, être dans la norme ne signifie rien d'autre qu'être dans la moyenne, ne pas se détacher trop de cette moyenne. On aura alors tendance à définir la norme comme la fréquence de ce qui se produit le plus souvent, comme l'attitude la plus fréquemment rencontrée ou de la préférence la plus régulièrement manifestée. Mais cette normalité n'est pas la normativité, et "être dans la norme", être exemplaire, relève alors d'une dimension différente, qui se rattache cette fois non plus à une description du réel mais à une identification de ce vers quoi il doit tendre. On passe donc du registre de l'être à celui du devoir-être, du *is* au *ought*. Il est en effet difficile d'envisager le modèle (ou la normalité) sans glisser vers ce second sens qu'on peut trouver au concept de norme, et qui déploie quant à lui une dimension à proprement parler normative. Cette vision mène à une confusion entre normes et lois, même si toute la normativité n'est pas énoncée sous forme de lois. David Hume constate ainsi que, dans tous les systèmes de morale, les auteurs passent de constat de faits, c'est-à-dire énonciatifs de type "il y a", à des propositions qui comportent une expression normative, comme "il faut", "on doit". Ce qu'il conteste, c'est le passage d'un type d'affirmation à un autre : pour lui, ce sont là deux types d'énoncés qui n'ont rien à voir les uns avec les autres, et qu'on ne peut donc pas enchaîner logiquement les uns avec les autres, en particulier d'une norme empirique vers une règle normative. Pour Hume, une affirmation qui ne serait pas normative ne peut pas donner lieu à une conclusion de type normatif. Cette affirmation de David Hume a suscité nombre de commentaires et d'interprétations, en particulier parce qu'énoncée telle quelle, elle semble être un obstacle à toute tentative de naturalisation de la morale (comme le détaillent MacIntyre 1969 ou Rescher 2013). En ce sens, on trouve la distinction forte entre la norme dans la régularité (normalité) et la règle (normativité).

Le principe d'égalité des êtres humains, reconnu comme un droit fondamental, impose l'obligation correspondante de ne pas discriminer. Aussi, tenter de définir ce qu'est la discrimination, c'est tenter de préciser ce que signifie, concrètement, ce principe de l'égalité de tous. La discrimination se définit comme étant un traitement inégal, et défavorable, appliqué à certaines personnes en raison d'un critère prohibé par la loi (à savoir, la race, l'origine, le sexe, etc.) Selon l'article 225-1 du Code Pénal, "*constitue une discrimination toute distinction opérée entre les personnes physiques sur le fondement de leur origine, de leur sexe, …*". Au États-Unis, le projet de loi sur l'interdiction de la discrimination en matière d'assurance automobile, ou PAID (*Prohibit Auto Insurance Discrimination Act*), présenté en juillet 2019 au Congrès [33], interdit à un assureur automobile de tenir compte de

33. https://www.congress.gov/bill/116th-congress/house-bill/3693/

certains facteurs lorsqu'il détermine la prime d'assurance, ou son admissibilité. De manière plus explicite, ces facteurs interdits comprennent le genre du conducteur, sa situation d'emploi, son code postal, son secteur de recensement, son état civil, et sa cote de crédit.

Epstein et King 2002 soulignent que, contrairement aux modèles statistiques traditionnels, un algorithme d'intelligence artificielle ne peut pas fonctionner s'il ne s'appuie que sur l'intuition initiale d'un humain sur les explications causales des liens statistiques entre les données d'entrée et la variable cible. Au lieu de cela, les IA utilisent des données d'entraînement pour découvrir par elles-mêmes quelles caractéristiques peuvent être utilisées pour prédire la variable cible. Bien que ce processus ignore complètement la causalité, il conduit inévitablement les IA à "rechercher" des proxys pour des caractéristiques directement prédictives lorsque les données sur ces caractéristiques ne sont pas mises à la disposition de l'IA en raison d'interdictions légales, Mittelstadt *et al.* 2016. Comme le soulignaient Barocas et Selbst 2016, "*thus, a data mining model with a large number of variables will determine the extent to which membership in a protected class is relevant to the sought-after trait whether or not that information is an input*".

2.3.3 Données massives et proxy

La mutualisation des risques individuels est le principe fondateur de l'activité d'assurance. Elle consiste, pour un groupe d'agents économiques, à mettre en commun une certaine fraction de leurs ressources pour indemniser les membres du groupe qui subiraient des dommages, comme le rappellent Henriet et Rochet 1987. Dans les premières formes d'activité d'assurance, ce principe supposait une participation égale de chaque membre de la communauté, ou éventuellement une participation proportionnelle aux ressources individuelles, mais certainement pas une participation proportionnelle aux risques individuels. La concurrence étant de plus en plus active sur les marchés de l'assurance de plusieurs pays (notamment en assurance automobile), une nouvelle tendance est apparue : celle de la personnalisation des primes. Cette vision antagoniste de la mutualisation considère que chaque assuré doit payer une prime proportionnelle à son risque individuel. De nombreux assureurs s'opposent fermement à cette tendance à la personnalisation, certains d'entre eux rejetant même le concept de "risque individuel mesurable" comme un non-sens statistique. Cependant, la personnalisation est également défendue par d'autres pour des raisons d'équité (chaque membre participe proportionnellement à la charge qu'il impose à la collectivité) et peut même être conciliée avec la mutualisation : si le marché est suffisamment grand, les individus peuvent être regroupés en un très grand nombre de mutualités, chaque mutualité étant homogène du point de vue du risque. Et l'arrivée des données massives et l'utilisation des techniques d'apprentissage machine, dans le monde de l'assurance, semblerait rendre possible cette personnalisation, comme le notent Barry et Charpentier 2020, avec la perspective hypothétique de "*risk 'pool of one'*" pour reprendre l'expression de McFall, Meyers *et al.* 2020.

L'assurance automobile couvre les risques qui sont liés aux habitudes et aux comportements du conducteur, comme le rappelle Landes 2015. L'assurance incendie indemnise les risques qui apparaissent à la suite d'un manque de soins (concernant les appareils électriques ou les fours). L'assurance contre le vol à domicile couvre des risques qui peuvent généralement être évités par des soins appropriés (serrures supplémentaires sur les portes, système de surveillance par caméra, chiens de garde, alarme, etc.) Et les actuaires tentent de capturer ces informations, en enrichissant leurs données. Pour Daniels 1990, "*we are morally obliged to make sure that premiums for insurance of any type reflect the risks of the insured*".

Et dans le cadre de la personnalisation du risque, l'idée que les produits d'assurance doivent avant tout être "équitables" pour les assurés est de plus en plus souvent exprimée par les commentateurs, comme le discutent Meyers et Van Hoyweghen 2018. Mais pour McFall 2019, si le comportement individuel plutôt que l'appartenance à un groupe devait devenir la base de l'évaluation des risques, les conséquences sociales, économiques et politiques seraient considérables. Cela perturberait le caractère distributif et solidaire qui s'exprime dans tous les régimes d'assurance maladie, même dans ceux qui sont nominalement désignés comme privés ou commerciaux. La tarification personnalisée des risques est en contradiction avec les infrastructures qui définissent, réglementent et fournissent actuellement l'assurance maladie. En allant plus loin, Van Lancker 2020 voit dans l'utilisation généralisée de la technologie numérique une modification de l'organisation de l'État-providence d'après-guerre qui affecte son potentiel à garantir un niveau de vie décent pour tous. Moor et Lury 2018 explorent la façon dont la tarification a historiquement été impliquée dans la constitution des personnes et comment la capacité de "personnaliser" la tarification reconfigure la capacité des marchés à discriminer. Car de nombreuses techniques de tarification récentes rendent plus difficile pour les consommateurs de s'identifier comme faisant partie d'un groupe reconnu. O'Neil 2016 en parlait lorsqu'elle écrivait "*now, with the evolution of data science and network computers, insurance is facing fundamental change. With ever more information available (⋯) insurers will increasingly calculate risk for the individual and free themselves from the generalities of the larger pool*".

Et si le "*big data*" peut présenter un danger, les assureurs aiment créer des scores synthétiques. Beniger 2009 souligne que le volume de données est fortement réduit, une unique variable semblant contenir toute l'information en lien avec toutes sortes de risques pour un individu donné. En assurance automobile, le véhicule peut être associé à des dizaines de variables (marque, modèle, vitesse maximale, poids, carburant, couleur, nombre de places, présence d'éléments de sécurité, couleur, valeur à l'Argus, etc.) mais la plupart des assureurs dispose d'un "véhiculier" résumant toute l'information liée au véhicule en une dizaine de classes. En assurance habitation, l'adresse du logement, qui peut être associée à des dizaines de variables (zone inondable, distance à la caserne de pompiers la plus proche, matériaux de construction, etc.), se retrouve associée à un "zonier" constitué de quelques classes.

Prince et Schwarcz 2019 parlaient de "*discrimination par proxy*" (que l'on pourrait traduire "*discrimination par procuration*"), pour décrire cet impact fortuit. La discrimination par procuration se produit lorsqu'un trait facialement neutre est utilisé comme substitut - ou proxy - pour un trait interdit. Austin 1983, citant Works 1977, affirmait "*Although the core concern of the underwriter is the human characteristics of the risk, cheap screening indicators are adopted as surrogates for solid information about the attitudes and values of the prospective insured (⋯) The invitations to underwriters to introduce prejudgments and biases and to indulge amateur psychological stereotypes are apparent. Even generalized underwriting texts include occupational, ethnic, racial, geographic, and cultural characterizations certain to give offense if publicly stated.*" Prince et Schwarcz 2019 considèrent trois types de "*discrimination par procuration*" : la "*discrimination par procuration causale*", la "*discrimination par procuration opaque*" et la "*discrimination par procuration indirecte*". La "*discrimination par procuration opaque*" apparaît lorsqu'on est incapable d'établir formellement un lien causal entre la variable sensible p et la variable cible y. Dans le contexte génétique, même pour de nombreuses variantes génétiques pathogènes, on ignore souvent pourquoi une séquence particulière d'un gène entraîne un risque accru. Dans la discrimination par procuration, aussi bien causale qu'opaque, les caractéristiques interdites sont "directement prédictives" des résultats légitimes d'intérêt. Dans le

cas d'une "*discrimination par procuration indirecte*", une variable x a un pouvoir prédictif important simplement car elle est corrélée à la variable cible y, et que la vraie variable causale n'est pas présente dans la base de données. Un exemple classique est qu'à l'école, la pointure de chaussure est indirectement prédictive du nombre de fautes d'orthographe dans une dictée.

Cette discrimination involontaire par procuration par les IA ne peut être évitée simplement en privant l'IA d'informations sur l'appartenance d'individus à des classes légalement suspectes ou de mandataires évidents pour une telle appartenance à un groupe Gillis et Spiess 2019 "*However, the exclusion of the forbidden input alone may be insufficient when there are other characteristics that are correlated with the forbidden input—an issue that is exacerbated in the context of big data*". Bornstein 2018 évoque la théorie des stéréotypes en matière de responsabilité ("*stereotype theory of liability*", voir aussi Selbst et Barocas 2018).

2.3.4 Le prénom et le nom comme proxy

Comme l'affirmait Bosmajian 1974, "*an individual has no definition, no validity for himself, without a name. His name is his badge of individuality, the means whereby he identifies himself and enters upon a truly subjective existence*". Bien souvent, les noms sont les premières informations dont disposent les gens lors d'une interaction sociale. Parfois, nous connaissons des individus par leur nom avant même de les rencontrer en personne, comme le rappelle Erwin 1995. Le prénom et le nom peuvent être porteurs de beaucoup d'informations, comme le montrent Hargreaves *et al.* 1983, Dinur *et al.* 1996 ou Daniel et Daniel 1998. Pour reprendre une citation de Young *et al.* 1993, "*the name Bertha might be judged as belonging to an older Caucasian women of lower-middle class social status, with attitudes common to those of an older generation* $(\cdots)$ *a person with a name such as Fred, Frank, Edith, or Norma is likely to be judged, at least in the absence of other information, to be either less intelligent, less popular, or less creative than would a person with a name such as Kevin, Patrick, Michelle, or Jennifer*".

Lors d'opérations de "*testing*", Petit *et al.* 2015 évoquent trois profils : le premier profil correspond au candidat dont le nom et le prénom sont à consonance maghrébine ("*par exemple, Abdallah Kaïdi, Soufiane Aazouz ou Medji Ben Chargui*"), celui dont le prénom est à consonance française et le nom à consonance maghrébine ("*par exemple, François El Hadj, Olivier Ait Ourab ou Nicolas Mekhloufi*"), et celui avec un prénom et un nom à consonance française ("*par exemple Julien Dubois, Bruno Martin ou Thomas Lecomte*"). Amadieu 2008 mentionne que les prénoms (masculins) Sébastien, Mohammed et Charles-Henri sont utilisés pour des tests. La Table 2.5 montre les principaux prénoms sur trois générations issues de l'immigration. Pour les prénoms en France, Coulmont 2011 revient sur les informations contenues dans le prénom.

		immigrés	enfants d'immigrés	petits-enfants d'immigrés
europe du sud	hommes	José, Antonio, Manuel	Jean, David, Alexandre	Thomas, Lucas, Enzo
	femmes	Maria, Marie, Ana	Marie, Sandrine, Sandra	Laura, Léa, Camille
maghreb	hommes	Mohamed, Ahmed, Rachid	Mohamed, Karim, Mehdi	Yanis, Nicolas, Mehdi
	femmes	Fatima, Fatiha, Khaduja	Sarah, Nadia, Myriam	Sarah, Inès, Lina

Table 2.5 – Top 3 des prénoms par sexe et générations en France, en fonction de l'origine (Europe du Sud ou Maghreb) de grands-parents, Coulmont et Simon 2019.

Le prénom comme proxy, par Baptiste Coulmont [34]

Dès les années 1930, des historiens comme Marc Bloch, en France, ont eu l'intuition que les prénoms pourraient servir à la recherche historique. "*Le choix même des noms de baptême, leur nature, leur fréquence relative, sont autant de traits qui, convenablement interprétés, révèlent des courants de pensée ou de sentiment*", écrit-il dans Bloch 1932. Le prénom d'une personne est en effet d'un côté le choix intime du couple parental (et de ses proches), et de l'autre côté une information très souvent présente dans des registres de populations. Cette conjonction du privé et du public permet de concevoir le prénom comme une porte d'entrée vers l'étude de la culture d'un groupe, parce qu'il dit des choses sur les donneurs de prénoms et les porteurs de prénoms. Que dit-il ? Marc Bloch est suffisamment prudent pour insister sur la nécessité d'une "*interprétation convenable*". Entre le genre du prénom et le sexe de l'état civil, les associations sont suffisamment fortes pour guider l'interprétation : si les prénoms épicènes existent, ils ne nomment qu'une petite partie de la population. Et il est même possible, pour un grand nombre de prénoms, à partir de ce simple prénom, d'en inférer avec une grande certitude le sexe de la personne qui le porte. C'est ainsi que, après une époque où il était interdit de recueillir le sexe des personnes pacsées, il a été possible de reconstituer l'information manquante, Carrasco 2007. Les phénomènes de mode autour des prénoms, visibles par l'engouement puis l'abandon au cours des années, permettent d'estimer « l'âge moyen » des porteurs de prénom. Aujourd'hui, les Nathalie sont plus âgées que les Emma. Mais cela ne fait pas du prénom un signal transparent de l'âge d'une personne. Une enquête intéressante, ayant demandé à des participants d'estimer l'âge d'une personne à partir de son prénom, indique que "*la perception de l'âge d'un prénom ne correspond que faiblement avec la véritable année moyenne de naissance des personnes portant ce prénom*" Johfre 2020. L'interprétation doit être entreprise avec prudence quand l'étude porte sur d'autres caractéristiques des prénoms. Des parents situés à des endroits différents de l'espace social choisissent des prénoms différents, Besnard et Grange 1993. La fréquence des prénoms varie donc avec le milieu social. Aujourd'hui, les Anouk, Adèle ou Joséphine prises comme un groupe ont des parents plus diplômés que les Anissa, Mégane ou Deborah. Mais cette relation va dépendre du temps, une partie des prénoms se diffusant — en suivant la mode — d'un milieu à un autre. Enfin, quand les caractéristiques associées aux prénoms sont liées à des identités fluides, contextuelles, ou assignées administrativement de l'extérieur, Escafré-Dublet *et al.* 2020, c'est l'ajout d'autres variables qui vient donner sens aux prénoms. De nombreux travaux (Tzioumis 2018 ou Mazieres et Roth 2018) cherchent à exploiter les informations sur l'origine ethnique, nationale ou géographique que contiennent les noms et les prénoms, mais ils ont besoin, pour débuter l'enquête, d'un lien entre le prénom et la variable étudiée. Des annuaires de pays différents, par exemple. La généralisation, à d'autres pays ou d'autres populations, des corrélations repérées entre origine et prénom, doit se faire avec prudence.

Dans "*Are Emily and Greg more employable than Lakisha and Jamal ?*", Bertrand et Mullainathan 2004, pour manipuler la perception de la race, des CV se sont vu attribuer, au hasard, des prénoms et des noms à consonance afro-américaine ou blanche. Les "noms blancs' reçoivent 50 % de rappels en plus pour les entretiens. Voicu 2018 présente le modèle BIFSG ("*Bayesian Improved First Name Surname Geocoding*") afin d'utiliser le prénom pour améliorer la classification de la race et de l'ethnicité dans un contexte de crédit hypothécaire, en s'inspirant de Coldman *et al.* 1988 et Fiscella et Fremont 2006. En analysant les données du panel socio-économique allemand, Tuppat et Gerhards 2021 montrent que les immigrants portant des prénoms considérés comme peu communs dans le pays d'accueil se plaignent de discrimination de manière disproportionnée. Et si le nom est un marqueur indiquant l'ethnicité, les immigrants les plus instruits signalent plus fréquemment qu'ils perçoivent une discrimination dans le pays d'accueil que les immigrants moins instruits ("*discrimination paradox*"). Rubinstein et Brenner 2014 montrent que le marché du travail israélien pratique une discrimination fondée sur l'appartenance ethnique perçue (entre des noms de famille à consonance séfarades et ashkénazes). Carpusor et Loges 2006 analysent l'impact des noms et prénoms sur le marché locatif, alors que Sweeney 2013 analyse leur impact sur la publicité en ligne.

34. Professeur à l'École normale supérieure Paris-Saclay

2.3.5 Le visage comme proxy

Il y a plus d'un siècle, Lombroso 1876 puis Bertillon et Chervin 1909 posaient les fondements de la phrénologie, et la théorie du "criminel-né", qui suppose que les caractéristiques physiques sont corrélées avec des caractères psychologiques et des penchants criminels. L'idée était de bâtir des classifications des types humains sur la base de caractéristiques morphologiques, afin d'expliquer et prévoir des différences de moeurs, et de conduites. On pourrait parler de l'invention d'un "délit de faciès". On peut aussi mentionner les "*ugly laws*", au sens de Schweik 2009, reprenant un terme utilisé par Burgdorf et Burgdorf Jr 1974 pour décrire des lois, en vigueur dans plusieurs villes aux États-Unis jusqu'aux années 1920, mais dont certaines ont perduré jusqu'en 1974. Ces lois autorisaient l'interdiction aux personnes ayant des marques et des cicatrices "disgracieuse" sur le corps, d'aller dans des lieux publics, en particulier dans des parcs.

Ces débats ressurgissent, alors que le nombre d'applications de la technologie de reconnaissance faciale augmente, grâce à l'amélioration de la qualité des images, des algorithmes utilisés, et de la puissance de traitement des ordinateurs. Boczar *et al.* 2021 montrent le potentiel de ces outils de reconnaissance faciale pour effectuer une évaluation de la santé. Historiquement, Anguraj et Padma 2012 avaient proposé un outil de diagnostique de paralysie faciale, et récemment, Hong *et al.* 2021 utilisent le fait que de nombreux syndromes génétiques présentent un dysmorphisme facial ou et des gestuelles faciales qui peuvent être utilisées comme outil de diagnostic pour reconnaître un syndrome.

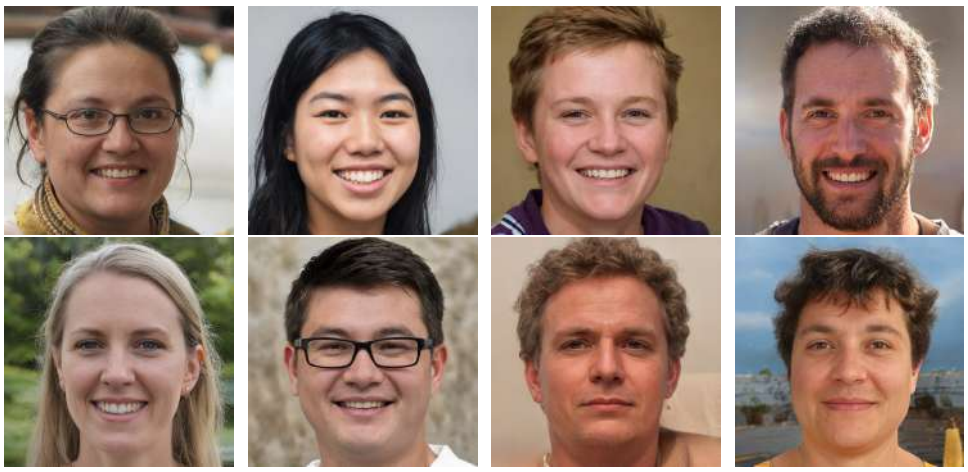


Figure 2.6 – Images de visages généréees par Karras *et al.* 2020, via https://nvlabs-fi-cdn.nvidia.com/stylegan2-ada-pytorch/images/.

Plus généralement, au-delà des considérations médicales, Wolffhechel *et al.* 2014 rappellent que plusieurs traits de personnalité peuvent être lus à partir d'un visage, et que les caractéristiques faciales influencent les premières impressions. Cela dit, le modèle de prédiction considéré ne parvient pas à prédire de manière fiable les traits de personnalité à partir des caractéristiques du visage. Mais Kachur *et al.* 2020 affirment que les développements de techniques récentes, accompagnés du développement de banques d'images importantes, a permis de prédire des profils de personnalité multidimensionnels à partir d'images faciales statiques, en utilisant des réseaux de neurones entraînés sur de grands ensembles de données étiquetées. Il y a une dizaine d'années, Cao *et al.* 2011 proposaient de prédire le sexe d'une personne à partir d'images faciales (Rattani *et al.* 2017 ou Rattani *et al.* 2018), et récemment Kosinski 2021 utilisaient un

algorithme de reconnaissance faciale pour prédire une orientation politique (dans un contexte binaire, opposant libéraux et conservateurs, dans l'esprit de Rule et Ambady 2010). Wang et Kosinski 2018 proposaient d'utiliser ces algorithmes pour prédire l'orientation sexuelle (voir aussi Leuner 2019).

Dans Shikhare 2021, l'idée d'utiliser un "*facial score model*" pour la souscription de contrats d'assurance-vie était évoquée (en lien avec le concept d'AUW, "*accelerated underwriting*"), avec des images comme sur la Figure 2.6. L'idée est de chercher des "*abnormal charactéristics*" en lien (probable) avec une condition particulière (Syndrome de Down, de Cornelia de Lange, de Cushing, acromégalie, etc.), y compris une anormalité de la couleur de peau pour détecter un asthme bronchique ou une hépatite, ou pour inférer le genre de la personne, comme le mentionne Shikhare 2021.

2.3.6 La voix comme proxy

La voix est un élément important, souvent très subjectif lors de rencontres avec des agents, en personne, mais aujourd'hui analysée par des algorithmes, en particulier lorsqu'un assureur utilise des robots conversationnels (on parlera parfois de "*chat bot*"), comme analysé par Hunt 2016, Koetter *et al.* 2019, Nuruzzaman et Hussain 2020, Oza *et al.* 2020 ou Rodríguez Cardona *et al.* 2021. On peut mentionner, en France, l'expérience de Maxime, décrite par Chardenon 2019, l'assistant virtuel lancé par Axa Protection Juridique.

À l'automne 2020, en France, une proposition de loi réprimant les discriminations fondées sur l'accent a été présentée au parlement, comme le rapporte Le Monde 2021b. Les linguistes parleraient de discrimination de phonostyles, comme Léon 1993, ou de "variation diastratique", avec des différences entre les usages pratiqués par genre, par âge et par milieu social (au sens large), dans Gadet 2007, ou encore de "glottophobie", terme introduit par Blanchet 2017. La glottophobie peut être définie comme "*le mépris, la haine, l'agression, le rejet, l'exclusion, de personnes, discrimination négative effectivement ou prétendument fondées sur le fait de considérer incorrectes, inférieures, mauvaises certaines formes linguistiques (perçues comme des langues, des dialectes ou des usages de langues) usitées par ces personnes, en général en focalisant sur les formes linguistiques (et sans toujours avoir pleinement conscience de l'ampleur des effets produits sur les personnes)*". Si Van Parijs 2002 s'intéresse surtout aux personnes discriminées, car le français n'est pas leur langue maternelle ("*mother tongue, in this perspective, is as illegitimate a basis for discrimination as race, gender or faith*"), Blanchet 2017 insiste surtout sur les différences culturelles (l'allongement des voyelles, la distribution des pauses, le débit de parole, l'accentuation de certaines syllabes, richesse du vocabulaire, etc.) On pourra penser aux "accents de banlieue" comme les appelait Fagyal 2010.

Blodgett et O'Connor 2017 ont mis en évidence une frontière importante en matière d'équité algorithmique : la disparité de la qualité des algorithmes de traitement du langage naturel lorsqu'ils sont appliqués au langage d'auteurs appartenant à des groupes sociaux différents. Par exemple, les systèmes actuels analysent parfois plus mal le langage des femmes et des minorités que celui des Blancs et des hommes. Enfin, notons qu'au Canada, l'expression "*speak white*" était une insulte raciste que les anglophones utilisaient pour caractériser ceux qui ne parlaient pas anglais dans les lieux publics. Historiquement, c'est ce qui fut rétorqué au député du Parti libéral Henri Bourassa, en 1889, lorsqu'il tenta de s'exprimer en français au cours des débats à la Chambre des communes du Canada. Comme le dit Michèle Lalonde, auteure du poème "*Speak White*" en 1968 (rapporté

par Dostie 1974) "*Speak White, c'est la protestation des Nègres blancs d'Amérique. La langue ici est l'équivalent de la couleur pour le noir américain. La langue française, c'est notre couleur noire*".

Pour Squires et Chadwick 2006, le "profilage linguistique", c'est-à-dire l'identification de la race d'une personne à partir du son de sa voix et l'utilisation de cette information pour établir une discrimination fondée sur la race, a été documenté sur le marché de la location de maisons, et Squires et Chadwick 2006 analysent son impact dans le secteur de l'assurance habitation. À partir d'une analyse de tests appariés effectués par des organisations de promotion du logement équitable, ils montrent que les agents d'assurance habitation sont généralement capables de détecter la race d'une personne qui les contacte par téléphone et que cette information affecte les services fournis aux personnes qui se renseignent sur l'achat d'une police d'assurance habitation. Enfin, pour l'anecdote, Durry 2001 raconte qu'en 1982, les dirigeants d'une compagnie d'assurance ont été attaqués en justice pour avoir "*refusé de garantir en assurance automobile les individus ne sachant ni lire ni écrire*". Sur la base légale que "*tout ce qui n'est pas interdit est permis*" (mentionnée en introduction), la cour de Rouen a prononcé la relaxe. "*Il est cependant permis de penser que le critère choisi aboutissait plus souvent à éliminer des étrangers que des Français*" soulignait Durry 2001. Plus récemment, pour revenir à notre discussion introductive, les "*chat bot*" ont montré être capables de reproduire des discriminations, majoritairement en lien avec le genre de l'interlocuteur, comme l'expliquent Feine *et al.* 2019, Aran *et al.* 2019, McDonnell et Baxter 2019 ou Maedche 2020, mais aussi être clairement racistes, comme le montraient Schlesinger *et al.* 2018. Il convient alors d'être prudent lorsque la voix est utilisée dans un algorithme "boîte noire".

2.3.7 L'adresse géographique comme proxy

L'adresse permettait, historiquement, d'associer un assuré à une zone urbaine. Sur la Figure 2.7 on peut visualiser, par zone IRIS [35], le revenu médian (par ménage) du quartier, la proportion de personnes de plus de 65 ans, le taux de logements de moins de 40 m^2 et de plus de 1000 m^2. On pourrait avoir des statistiques, par IRIS, sur la vétusté des bâtiments, le taux de cambriolage, etc. Mais il est aujourd'hui possible de quitter un "*redlining*" grossier et imprécis en utilisant de l'imagerie satellite, ou *Google Street View* (dans certaines villes), comme sur la Figure 2.8.

Dans un premier temps, Jean *et al.* 2016 avaient noté qu'à un niveau assez grossier, l'éclairage nocturne était un indicateur approximatif de la richesse économique. En effet, les cartes mondiales nocturnes montrent que de nombreux pays en développement sont faiblement éclairés. Jean *et al.* 2016 ont combiné des cartes nocturnes avec des images satellites diurnes à haute résolution, et les images combinées ont permis – "*with a bit of machine-learning wizardry*" – d'obtenir des estimations précises de la consommation et des biens des ménages (deux grandeurs souvent difficiles à mesurer dans les pays les plus pauvres). Par la suite, Seresinhe *et al.* 2017 ont proposé d'évaluer la quantité d'espaces verts dans différentes localisations sur la base d'images satellites. En allant toujours plus loin, Gebru *et al.* 2017 affirmaient pouvoir quantifier des attributs socio-économiques comme le revenu, l'origine ethnique, le niveau d'éducation et les habitudes de vote à partir des voitures détectées dans les images de Google Street View. Par exemple, aux États-Unis, si le nombre de berlines dans un quartier est supérieur au nombre de camionnettes, ce quartier est susceptible de voter pour un démocrate lors de la prochaine élection présidentielle (88 % de

35. IRIS = *Ilôts Regroupés pour l'Information Statistique*, découpage du territoire français en mailles de taille homogène dont la taille de référence visée par maille élémentaire est de 2,000 habitants. La France compte 16,100 IRIS, dont 650 dans les départements d'outre-mer, et Paris (intra muros) compte 992 IRIS.

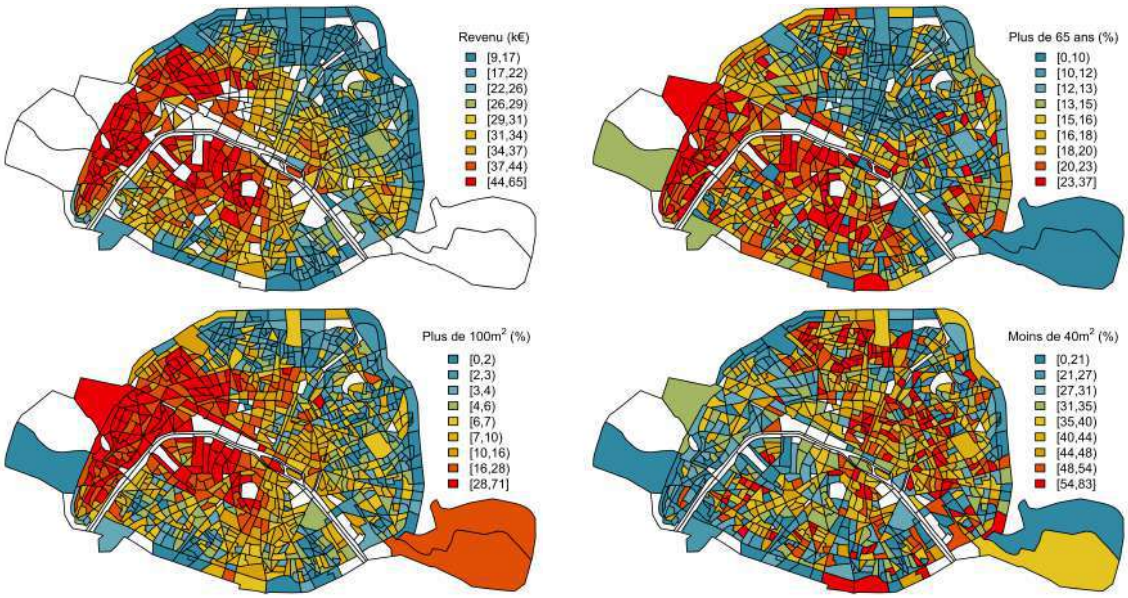


Figure 2.7 – Revenu médian par ménage, proportion de personnes âgées, proportion de logements en fonction de leur superficie, par quartier (IRIS), à Paris. Ces informations statistiques sur le quartier (et non pas l'assuré) peuvent être utilisées pour tarifer un contrat d'assurance habitation, par exemple.

Figure 2.8 – Exemples d'images associées à des logements, avec des photos prises dans la rue à gauche, avec Google Street View (projet lancé en 2007), et de l'imagerie aérienne, avec Google Earth (projet lancé en 2001), à droite.

chances) ; sinon, la ville est susceptible de voter pour un républicain (82 % de chances).

Comme le rappellent Law *et al.* 2019, lorsqu'un individu achète une maison, il achète simultanément ses caractéristiques structurelles, son accessibilité au travail et les commodités du quartier. Certaines commodités, comme la qualité de l'air, sont mesurables, tandis que d'autres, comme le prestige ou l'impression visuelle d'un quartier, sont difficiles à quantifier. Rundle *et al.* 2011 notent que Google Street View permet de se faire une idée de la sécurité du quartier (en lien avec la sécurité routière), si l'on cherche les passages piétons, la présence de parcs et d'espaces verts, etc. À partir de données d'images de rue et d'images par satellite, Law *et al.* 2019 montrent qu'il est possible de capturer ces caractéristiques a priori non-quantifiables et d'améliorer l'estimation des prix des logements, à Londres, au Royaume-Uni. Un réseau de neurones, avec en entrée les caractéristiques traditionnelles des logements telles que l'âge, la taille et l'accessibilité, ainsi que les caractéristiques visuelles des images de Google Street View et des images aériennes, est entraîné pour estimer le prix des maisons. Il est aussi possible d'inférer certaines informations personnelles possiblement sensibles, comme un possible handicap avec la présence d'une rampe d'accès pour la maison (la méthodologie est détaillée dans Hara, Sun, Moore *et al.* 2014), ou une possible

orientation sexuelle avec la présence d'un drapeau arc-en-ciel à la fenêtre, ou politique avec un drapeau confédéré (comme mentionné dans Mas 2020). Ilic *et al.* 2019 évoquent aussi cette "cartographie profonde" des attributs environnementaux. Un réseau de neurones convolutifs siamois (SCNN) entraîné sur des séquences temporelles d'images de Google Street View (GSV) est entraîné. L'évolution au cours du temps permet de confirmer certaines zones urbaines connues pour être en cours de gentrification, tout en révélant des zones en cours de gentrification qui étaient auparavant inconnues. Finalement, Kita et Kidziński 2019 évoquent des applications directes possibles en assurance.

2.3.8 Le score de crédit comme proxy

Les scores de crédit sont une information individuelle importante aux États-Unis ou au Canada, largement utilisés dans de nombreuses branches d'assurance, et pas seulement l'assurance emprunteur. Comme le soulignent Arya *et al.* 2013, un incident de crédit (négatif), tel qu'un défaut (ou un retard) de paiement sur un prêt hypothécaire, ou une faillite, peut avoir un impact sur un particulier pendant une période très importante ("*can haunt a consumer for a considerable period of time*"). Ces scores de crédit sont des nombres qui représentent une évaluation de la solvabilité d'une personne, ou la probabilité qu'elle rembourse ses dettes. Mais de plus en plus, ces scores sont utilisés dans des contextes bien différents, comme l'assurance. Guseva et Rona-Tas 2001 rappellent qu'en Amérique du Nord, Experian, Equifax et TransUnion tiennent des registres des activités d'emprunt et de remboursement d'une personne. Et FICO (Fair Isaac Corporation) a mis au point une formule (non connue) calculant, sur la base de ces registres, un score, fonction de la dette et du crédit disponible, du revenu, ou plutôt de leurs variations (avec l'historique des paiements, le nombre de demandes de crédit récentes et d'événements négatifs – tels qu'une faillite ou une saisie – ainsi que des modifications de revenus dues à des changements d'emploi ou de situation familiale). Le score FICO commence à 300 et monte jusqu'à 850, avec sommairement un mauvais score en dessous de 600, et un score "exceptionnel" au-delà de 800. Le score moyen est [36] aux alentours de 711. Ce score, créé pour les institutions bancaires, est aujourd'hui utilisé lors de vérifications préalables à l'embauche, comme le rappellent Bartik et Nelson 2016. Pour O'Neil 2016, cette utilisation des scores de crédit en matière d'embauche et de promotion crée un cercle vicieux dangereux en termes de pauvreté. Le *Credit-Based Insurance Scores* ne peut utiliser aucune information personnelle pour déterminer le score, autres que celles figurant dans dossier de crédit, en particulier la race et l'origine ethnique, la religion, le genre, l'état civil, l'âge, l'historique d'emploi, la profession, le lieu de résidence, les obligations alimentaires envers les enfants/familles ou les contrats de location. Sur la Table [37] 2.6, on peut voir l'évolution du taux du crédit proposé en fonction du score de crédit, sur 30 ans, pour un "bon risque" (3.2 %) et un "mauvais risque" (4.8 %), avec le montant total du crédit en ordonné (un mauvais risque aura un sur-coût de l'ordre de 20 %), et sur une prime d'assurance, pour un même profil, un "mauvais risque" ayant une sur-prime de l'ordre de 70 %.

Comme le rappelle Lauer 2017, les établissement de crédit ont, depuis longtemps, compilé des données sur les antécédents financiers des individus, tels que la ponctualité dans les remboursements de leurs prêts, le montant total de leurs emprunts, la fréquence des demandes de nouveaux crédits, parmi tant d'autres. Et les assureurs se sont rapidement rendu compte que ces informations étaient très prédictives de toutes sortes de risques. Dès les années 1970, aux États

36. https://www.experian.com/blogs/ask-experian/credit-education/score-basics/850-credit-score/
37. https://www.incharge.org/debt-relief/credit-counseling/credit-score-and-credit-report/

	300-619	620-659	660-699	700-759	760-850
Montant total du crédit	$ 283,319	$ 251,617	$ 245,508	$ 239,479	$ 233,532
Taux	4.8 %	3.8 %	3.6 %	3.4 %	3.2 %
Montant de l'assurance	$ 2580	$ 2400	$ 2150	$ 2000	$ 1500

Table 2.6 – En haut montant d'un crédit sur 30 ans, pour un montant de $150,000, en fonction du score de crédit (avec 5 catégories, allant du plus risqué à gauche au moins risqué à droite), avec le taux moyen proposé. En bas, prime d'assurance demandée à un conducteur de 30 ans, sans infractions constatées, conduisant une voiture moyenne, 12,000 miles par an, en ville, en fonction des mêmes catégories de score de crédit (source : InCharge Debt Solutions).

Unis, la loi a obligé les entreprises à informer les gens lorsque les dossiers de crédit entraînaient des mesures défavorables (ou "*adverse action*"), comme une hausse d'une prime d'assurance ou un crédit. Des décennies plus tard, cette règle a permis aux gens de savoir que les scores de crédit étaient à l'origine de la hausse soudaine de leurs tarifs d'assurance automobile. Pour Kiviat 2019, c'est seulement récemment qu'il est apparu clairement de l'importance du score de crédit en tarification de produit d'assurance, en particulier en assurance automobile. Miller *et al.* 2003 rappellent que les rapports de crédit contiennent une grande variété d'informations de crédit concernant un consommateur individuel. En plus des informations permettant d'identifier une personne en particulier, le rapport contient des données sur les soldes des cartes de crédit et des prêts, les types de crédit, l'état de chaque compte, les jugements, les privilèges, les recouvrements, les faillites et les demandes d'informations sur le crédit. Ces "scores de crédit" sont utilisés depuis longtemps par les institutions de prêt pour prédire le risque associé au remboursement d'un prêt ou à la satisfaction d'une autre responsabilité financière. Les modélisateurs de scores d'assurance ont commencé à combiner et à pondérer des attributs de crédit sélectionnés afin de développer un "score d'assurance" unique. Ces "scores d'assurance" sont ajoutés en tant que facteur de risque pour créer des plans de classification des risques dans le but d'obtenir des résultats plus précis. Même si le "score de crédit" et le "score d'assurance" sont tous deux dérivés du dossier de crédit d'une personne, les deux scores sont différents. Il n'y a aucune raison de croire qu'un score de crédit mesurant la probabilité de remboursement d'un prêt sera basé sur les mêmes attributs (ou que chaque attribut se verra attribuer le même poids) que ceux utilisés pour calculer un score d'assurance, et vice versa. Malheureusement, certains acteurs du secteur de l'assurance en sont venus à appeler les scores d'assurance basés sur le crédit des "scores de crédit". Cet abus de langage a pu amener certains à conclure que les avantages et les inconvénients de l'utilisation des scores de crédit dans le secteur des prêts ont un rapport direct avec les avantages et les inconvénients de l'assurance. Cela a conduit certains à tenter d'appliquer les résultats des études sur les "scores de crédit" (réalisées par les établissements de crédit) à l'utilisation des "scores d'assurance" par les assureurs. Morris *et al.* 2017 soulignent que les scores d'assurance basés sur le crédit ("scores d'assurance") sont peut-être l'exemple le plus important d'une caractéristique de l'assuré qui est réglementée parce qu'elle peut être à l'origine de "*potentially suspect classifications*". Toute corrélation entre les scores d'assurance, d'une part, et la race ou le revenu, d'autre part, est potentiellement troublante. Tout d'abord, la notation d'assurance peut avoir un impact disparate sur les minorités raciales et les ménages à faible revenu, ce qui fait que les membres de ces groupes paient en moyenne des primes plus élevées. Les informations sur le crédit, par exemple, sont utilisées pour la tarification des assurances automobile et habitation depuis plusieurs années, malgré leur potentiel de discrimination par procuration (selon

la classification de Kiviat 2019 et Prince et Schwarcz 2019.

Brockett et Golden 2007 s'étaient interrogés sur l'utilisation des scores de crédit pour tarifer des contrats d'assurance. Comme le note Kiviat 2019, inspiré par Morris *et al.* 2017, le revenu des assurés, en particulier, pourrait être prédictif des futures réclamations d'assurance si les assurés à faible revenu sont plus susceptibles de déposer des réclamations même lorsque les pertes ne sont que légèrement supérieures à leur franchise. Certains assureurs peuvent même ne pas savoir qu'il existe une corrélation entre la variable proxy (cotes de crédit) et la variable suspecte (race et revenu). Et même si les assureurs sont conscients de cette corrélation, ils peuvent ne pas croire que cette corrélation contribue à expliquer le pouvoir des informations de crédit pour prédire les sinistres. Au lieu de cela, ils peuvent croire, comme l'indiquent en fait de nombreuses preuves disponibles, que les informations de crédit sont prédictives des réclamations car elles mesurent les niveaux de soins des assurés. Certains assureurs, comme Root Insurance [38] expliquent qu'ils se refusent d'utiliser ces scores de crédit car ils les jugent discriminatoire : "*For decades, the car insurance industry has used credit score as a major factor in calculating rates. By basing rates on demographic factors like occupation, education, and credit score, the traditional car insurance industry has long relied on unfair, discriminatory biases in its pricing practices. These practices unfairly penalize historically under-resourced communities, immigrants, and those struggling to pay large medical expenses*".

2.3.9 Les réseaux comme proxy

Boyd *et al.* 2014 affirmaient qu'il existe une nouvelle forme de discrimination associée non pas à des caractéristiques personnelles (comme celles dont nous parlions dans la section précédente) mais à des réseaux personnels. Au-delà de leurs caractéristiques personnelles (telles que la race ou le sexe), une source importante d'information est "qui ils connaissent". Dans le contexte de l'assurance automobile basée sur l'utilisation, la nuance est que les réseaux personnels ne sont pas ceux représentés par un comportement de conduite (stricto sensus), mais ceux définis par les lieux où les gens se rendent physiquement.

Dans de nombreux pays, en matière d'emploi, la plupart des entreprises sont tenues de respecter l'égalité des chances : la discrimination fondée sur la race, le sexe, la croyance, la religion, la couleur et l'origine nationale est interdite. Des réglementations supplémentaires interdisent à de nombreux employeurs de pratiquer une discrimination fondée sur l'âge, les handicaps, les informations génétiques, les antécédents militaires et l'orientation sexuelle. Toutefois, rien n'empêche un employeur de pratiquer une discrimination fondée sur le réseau personnel de la personne. Et de plus en plus, comme le rappellent Boyd *et al.* 2014, les outils techniques de prise de décision fournissent de nouveaux mécanismes par lesquels cela peut se produire. Certains employeurs utilisent LinkedIn (et d'autres sites de réseaux sociaux) pour déterminer "l'adéquation culturelle" d'un candidat à l'embauche, et notamment si oui ou non un candidat connaît des personnes déjà connues de l'entreprise. Si l'embauche sur la base de relations personnelles n'est en aucun cas nouveau, elle prend une nouvelle signification quand elle devient automatisée et se produit à grande échelle. Les algorithmes qui identifient nos réseaux, ou qui prédisent notre comportement en fonction de ceux-ci, offrent de nouvelles possibilités de discrimination et de traitement inéquitable. Ces dernières années [39], plusieurs organismes ont proposé d'utiliser des informations sur nos "amis" afin d'en savoir davantage sur nous, suivant un principe d'homophilie important (au sens de Miller *et al.*

38. https://www.joinroot.com/blog/dropping-credit-score-from-car-insurance-by-2025/

39. Paul Verlaine nous avait pourtant mis en garde : "*il ne faut jamais juger les gens sur leurs fréquentations. Tenez, Judas, par exemple, il avait des amis irréprochables*".

2001), car comme le dit l'adage population, "*qui se ressemble s'assemble*". Ainsi Bhattacharya 2015 notait que "*you apply for a loan and your would-be lender somehow examines the credit ratings of your Facebook friends. If the average credit rating of these members is at least a minimum credit score, the lender continues to process the loan application. Otherwise, the loan application is rejected*".

2.4 Pour aller plus loin...

2.4.1 Prouver la discrimination

Comme le notent Parquet et Petit 2021, en France, à la fin des années 2010, le "testing" (selon la terminologie utilisée) s'inscrit au cœur du dispositif des politiques publiques de lutte contre les discriminations. Ces "testing" recouvrent toutefois un grand nombre de procédures. Le testing dit "juridique", pratiqué en France depuis les années 1990 par le milieu associatif, consiste en une série d'opérations-pièges visant à prendre sur le fait des établissements discriminants, tels que des discothèques ou encore des agences immobilières, et à les faire condamner en justice. Cette pratique a été validée par l'arrêt du 11 juin 2002 de la Cour de cassation de Paris (01-85.560). Un testing dit "scientifique" existe également, basé sur des expériences lancées en 1967 en Grande-Bretagne, et appelé "*practice testing*" ou "*situation testing*" (Riach et J. Rich 1991), visant à comparer les chances de succès de trois individus fictifs, citoyens britanniques originaires de Hongrie, des Caraïbes et d'Angleterre dans les domaines de l'emploi, du logement, et des assurances, comme le rappelle Héran 2010. Smith 1977 mentionne également un protocole utilisé en 1974, basé sur un tirage aléatoire d'offres d'emploi et sur l'envoi de candidatures écrites identiques. Aux États-Unis à la même époque, Fix et Turner 1998 Blank *et al.* 2004 mentionnent des "*affirmative action (pair-testing)*". Les testing dits "scientifiques" reposent sur des protocoles stricts, exigeants et souvent coûteux.

Parquet et Petit 2021 rappellent que l'approche classique, pour tester les discriminations à l'emploi, comporte trois volets : la conception d'un protocole, la dimension aléatoire et la construction de données. Dans ce contexte, la première étape vise à identifier et construire le profil-type des candidatures, qui servira de base ensuite (en modifiant une seule caractéristique ensuite). La composante aléatoire consiste à rajouter un peu de bruit (largement contrôlé) pour ne pas avoir des candidats à l'emploi qui soient des clones les uns des autres. Ces procédures permettent de neutraliser les biais de sélection et l'hétérogénéité habituellement inobservée, comme le note Heckman 1992, puisque le contrefactuel est construit *ex nihilo* : les candidats fictifs sont construits de toutes pièces de façon à être des jumeaux parfaits à l'exception de la caractéristique dont on souhaite tester l'effet. Cela dit, cela passe souvent par le biais d'une variable tiers, en cherchant à tester une discrimination par le genre ou raciale, souvent le nom et le prénom du candidat (tel qu'évoqué par Baptiste Coulmont page 68), le lieu d'habitation (avec un "quartier prioritaire de la politique de la ville", dans Bunel *et al.* 2016, en France) ou par un autre signal supposé fort (comme Ahmed *et al.* 2013 qui mentionnaient du travail bénévole dans une organisation gay-lesbienne pour tester la discrimination selon des préférences sexuelles, alors que Valfort 2017 utilisait les langues parlées pour indiquer une religion).

D'un point de vue éthique, il y existe un problème majeur, découlant du droit à la loyauté et à la transparence, associés au droit à l'information des individus. Selon ce principe, l'expérimentateur devrait fournir une information complète en termes clairs sur le traitement, ce qui ruinerait l'idée même du testing. Prouver qu'il y a discrimination est un exercice complexe. Une série de trois

"*testing*", menée pour quantifier les discriminations dans l'accès à la banque et à l'assurance, en France, est présentée dans l'Horty *et al.* 2019. Mais comme le notent Barnard et Hepple 1999 ou Ellis et Watson 2012, trouver des statistiques pour prouver qu'il y a une discrimination est complexe : "*in order to be reliable sophisticated statistical techniques are required which lie beyond the resources of parties in individual legal suits*" affirmait Browne 1993. Les statistiques concernant des questions sensibles, par exemple en rapport avec le handicap, l'orientation sexuelle ou l'origine ethnique, ne sont souvent pas disponibles, voir Tobler 2008 ou Farkas 2011.

l'Horty *et al.* 2019 rappellent que des discriminations liées au sexe et à la couleur de la peau ont été mises en évidence dans le processus de négociation d'achat d'une voiture neuve aux États-Unis, des discriminations raciales ont aussi été mises en évidence aux États-Unis dans l'accès à l'assurance (Wissoker *et al.* 1998), au crédit immobilier (Turner et Skidmore 1999) ou dans l'accès aux soins pour des patients cardiaques. Mais en assurance, la segmentation des prix est-elle pour autant une discrimination ? Sur les marchés du crédit et de l'assurance, les offreurs sont en asymétrie d'information sur la qualité réelle des demandeurs (Stiglitz et Weiss 1981). Pour limiter cette asymétrie, ils utilisent des méthodes de scoring afin d'estimer le risque dans leurs transactions et ajuster leur tarification. Les banquiers et les assureurs pratiquent ainsi une segmentation de la clientèle et une évaluation des risques sur la base de données statistiques et actuarielles. Ces pratiques ne sont pas considérées comme discriminatoires, mais plutôt comme une différenciation de tarification selon l'intensité des risques, dès lors que les offreurs de crédits ou d'assurance ont recours uniquement à des critères de segmentation qui ont une incidence avérée sur le risque. Même si un traitement différent selon l'âge est considéré comme légal, il apparaît intéressant de documenter son ampleur. Cependant, toute différenciation ne constitue pas une discrimination. Ce qui est discriminant n'est pas nécessairement discriminatoire. La discrimination suppose deux éléments : un traitement défavorable et dans une situation comparable. "*Deux personnes d'âge différent ne se trouvent pas dans une situation comparable au regard de l'assurance-vie et que des différences de traitement proportionnées, fondées sur une évaluation solide des risques, ne constituent par conséquent pas une discrimination*". l'Horty *et al.* 2019 ont testé 38 établissements d'assurance automobile, 52 établissements de complémentaires santé et 20 établissements financiers pour solliciter un crédit à la consommation en vue d'acheter une voiture d'occasion, en adressant à chacun les 6 demandes de devis des individus fictifs du testing, soit au total respectivement 228, 312 et 120 demandes envoyées. Sur le marché de l'assurance automobile, 35 établissements ont répondu à au moins un individu fictif du testing (accord de principe pour une assurance automobile) et 26 établissements ont adressé un accord de principe aux six individus. Sur le marché des complémentaires santé, tous les établissements ont répondu favorablement à au moins un individu fictif du testing (accord de principe pour une complémentaire santé) et 41 établissements ont adressé un accord de principe aux six individus. En assurance automobile, l'Horty *et al.* 2019 notaient l'absence de discriminations dans l'accès à ce marché et d'autre part l'existence de discriminations prenant la forme de tarifs différenciés selon le sexe et la réputation du lieu de résidence.

2.4.2 Vie privée, prédiction et attaques

Kelly 2021 rappelle que "*often the location data is used to determine what stores people visit. Things like sexual orientation are used to determine what demographics to target*". Chaque type de données peut révéler quelque chose sur nos intérêts et nos préférences, nos opinions, nos

loisirs et nos interactions sociales. Par exemple, une étude menée par le MIT[40] a démontré comment les métadonnées des courriels peuvent être utilisées pour cartographier nos vies, en montrant la dynamique changeante de nos réseaux professionnels et personnels. Ces données peuvent être utilisées pour déduire des informations personnelles, notamment les antécédents d'une personne, sa religion ou ses croyances, ses opinions politiques, son orientation sexuelle et son identité de genre, ses relations sociales ou sa santé. Par exemple, il est possible de déduire nos conditions de santé spécifiques simplement en reliant les points entre une série d'appels téléphoniques. Pour Mayer *et al.* 2016, la loi traite actuellement séparément le contenu des appels et les métadonnées et facilite l'obtention des métadonnées par les agences gouvernementales, en partie parce qu'elle part du principe qu'il ne devrait pas être possible de déduire des détails sensibles spécifiques sur les personnes à partir des seules métadonnées. Chakraborty *et al.* 2013 rappellent que les approches actuelles de la protection de la vie privée, généralement définies dans des contextes multi-utilisateurs, reposent sur l'anonymisation pour empêcher que de tels comportements sensibles ne puissent être retracés jusqu'à l'utilisateur - une stratégie qui ne s'applique pas si l'identité de l'utilisateur est déjà connue. Au fil du temps, un système de localisation peut être suffisamment précis pour placer une personne à proximité d'une banque, d'un bar, d'une mosquée, d'une clinique ou d'un autre lieu sensible du point de vue de la vie privée. En 2015, comme le raconte Miracle 2016, Noah Deneau s'est demandé s'il serait possible d'identifier les chauffeurs musulmans fervents de la ville de New York en examinant les données anonymes et les chauffeurs inactifs pendant les cinq moments de la journée où ils sont censés prier. Il a rapidement recherché les conducteurs dont l'activité était faible pendant les 30 à 45 minutes des heures de prières musulmanes, et a pu trouver quatre exemples de conducteurs qui pourraient correspondre à ce modèle. Ceci se rapproche de Gambs *et al.* 2010 qui ont mené une attaque sur un ensemble de données contenant les données de mobilité des chauffeurs de taxi dans la région de la baie de San Francisco. En trouvant les points où le capteur GPS du taxi était éteint pendant une longue période (par exemple deux heures), ils ont pu déduire les centres d'intérêt des chauffeurs. Pour 20 des 90 utilisateurs analysés, ils ont pu localiser un domicile plausible dans un petit quartier. Ils ont même confirmé ces résultats pour 10 utilisateurs en utilisant une vue satellite de la zone : elle montrait la présence d'un taxi jaune garé devant le domicile supposé du conducteur. Dalenius 1977 avait introduit un concept intéressant de vie privée : rien concernant un individu ne devrait pouvoir être appris à partir d'un ensemble de données s'il ne peut pas l'être également sans avoir accès à l'ensemble de données. On retrouvera cette idée quand on définira des critères d'équité, et que l'on demandera à ce que la variable protégée p ne puisse pas être prédite à partir des données, et des prédictions.

2.4.3 Multidimensionalité de la discrimination

Il est important de tenir compte du caractère multidimensionel des données (d'autant plus dans un contexte de "données massives"), y compris au regard des discriminiations. Comme le racontait Coluche, dans *le blouson noir*, "*Dieu a dit : il y aura des hommes blancs, il y aura des hommes noirs, il y aura des hommes grands, il y aura des hommes petits, il y aura des hommes beaux et il y aura des hommes moches, et tous seront égaux ; mais ça sera pas facile... Et puis il a ajouté : il y en aura même qui seront noirs, petits et moches et pour eux, ce sera très dur !*"

Au milieu du XIX$^{\text{ème}}$siècle, Adolphe Quételet introduisit le concept d'"homme moyen" dans Quetelet 1846. Dans les années qui suivirent, sous l'impulsion de Francis Galton, l'école anglaise

40. Projet https://immersion.media.mit.edu/

d'eugénisme s'intéressa aux déviations par rapport à cette norme, cette moyenne (déviation vers le haut et déviation vers le bas). Comme le rappelait Bulmer 2003, "*the deviations from that average—upwards towards genius, and downwards towards stupidity—must follow the law that governs deviations from all true averages*. Supposons que l'on s'intéresse à une grandeur d'intérêt (par exemple le poids d'une personne), mesurable, qu'elle soit centrée et réduite de telle sorte qu'elle suive approximativement une loi normale $\mathcal{N}(0,1)$. Supposons que l'"anormalité" corresponde à une personne sur vingt, soit 5 % de la population. Un individu est alors "normal" si la grandeur d'intérêt est entre -2 et +2, car $\mathbb{P}[X \in [-2,+2]] \sim 95\ \%$ si $X \sim \mathcal{N}(0,1)$. Si on considère maintenant une seconde grandeur, et que l'on suppose qu'une personne est "normale" si elle l'est dans les deux dimensions, la proportion de personnes normales sera $\mathbb{P}[-2 \leq X_1, X_2 \leq +2] = \mathbb{P}[X \in [-2,+2]]^2$ si on suppose les deux grandeurs comme étant indépendantes, soit $90\ \%$ de la population. Assez naturellement, la proportion de personnes "anormales" augmente avec la dimension (en $d \mapsto 1-0.95^d$). Si on suppose les grandeurs comme étant indépendantes les unes des autres, on note qu'en dimension 14, une minorité de personnes sont "dans la norme". L'individu moyen devient alors exceptionnel. C'est d'ailleurs ce qu'explique Rose 2016, au travers de deux exemples. Le premier est tiré de problèmes rencontrés par l'armée américaine dans les années 1950. En effet, lors de la conception des postes de pilotage d'avion de chasse, les ingénieurs avaient utilisé les dimensions de plus de 4000 pilotes pour positionner de manière optimale le siège par rapport aux pédales, le manche à balai, la hauteur du pare-brise, mais aussi la forme du siège, du casque, etc. Ces mesures ont permis de calculer les mensurations du pilote "moyen" dans une dizaine de dimensions. Par exemple la taille moyenne des pilotes était de 179 cm, ce qui a permis de définir la taille d'un pilote moyen entre 175 et 185 cm. Si une majorité des pilotes est de taille moyenne, parmi les 4000 pilotes, aucun n'était "moyen" dans toutes les dimensions. Comme l'affirmait Daniel 1952 "*there was no such thing as an average pilot. If you've designed a cockpit to fit the average pilot, you've actually designed it to fit no one*". Le second exemple est lié à deux statues, celles de Norma et Normann (exposées historiquement à Cleveland, aujourd'hui conservées à la bibliothèque de Harvard). L'artiste Abram Belskie et l'obstétricien Robert Latou Dickinson ont réalisé ensemble ces statues, en 1943. La particularité est qu'aucun modèle n'a été représenté. En fait, il s'agissait de représenter une femme et un homme qui avaient les mensurations moyennes de l'époque (à partir de mesures effectuées sur des milliers de sujets). Une fois ces statues réalisées, un concours a été organisé pour trouver qui ces statues pouvaient bien représenter. Plusieurs milliers de personnes de l'Ohio ont envoyé leurs mensurations, mais aucune ne correspondait à celles des statues. Certes, plusieurs centaines avaient la même taille. Plusieurs centaines avaient le même tour de poitrine. Mais aucune n'avait toutes les mêmes mesures. Car comme l'explique Todd Rose, l'homme n'est pas unidimensionnel : c'est sur plusieurs dimensions qu'on le mesure. Et le souci quand on travaille dans un contexte multivarié, c'est que la moyenne perd de son interprétation en terme de "majorité". En fait, d'un point de vue probabiliste, être moyen peut être extraordinaire.

Chapitre 3

Données et Biais

La plupart des ouvrages et articles qui parlent d'équité et de discrimination, dans le contexte de l'intelligence artificielle (O'Neil 2016, Hajian *et al.* 2016, Pleiss *et al.* 2017, Hoffmann 2019, Cowgill et Tucker 2019, Rambachan *et al.* 2020, Mehrabi *et al.* 2021, entre autres) insistent sur l'importance des biais de toutes ces "nouvelles" données que les assureurs souhaitent utiliser (y compris peut-être également dans les données plus traditionnelles). Barocas, Hardt *et al.* 2017 ont proposé plusieurs définitions de biais, étroitement liées aux données utilisées pour entraîner un modèle. Ces données biaisées sont à manipuler avec précaution, si on ne veut pas tomber dans le fameux "*garbage-in, garbarge-out*", expression familière datant de plus d'un demi-siècle (mentionnée dans l'introduction de Mellin 1957) qui signifie qu'utiliser de données de mauvaise qualité en entrée va conduire à des conclusions peu fiables, en sortie.

3.1 Observations et expériences

Pour reprendre la classification de Rosenbaum 2018, il est important de distinguer, lorsque l'on cherche à faire de la modélisation, entre les données d'expérience et les données d'observation. Dans ce dernier cas, on va se contenter de données aussi appelées "administratrives" qui montrent ce que les gens font et non pas ce qu'ils déclarent faire (comme dans des enquêtes, par exemple). Ces données permettent de voir ce que les gens achètent, mangent, où ils voyagent, etc., à partir des traces laissées. Il y a eu 185 milliards de transactions à l'aide de cartes Visa en 2019, et 85 milliards à l'aide de cartes Mastercard. Ces données massives nous donnent des informations sur le comportement de leurs détenteurs, à leur insu pourrait-on penser. Si regarder les transactions permet probablement de mieux quantifier le nombre de voyages par avion d'une personne qu'une enquête ou un rapide sondage, on se doute aussi que ces relevés comportent des biais et ne reflètent pas toutes les transactions et donnent probablement une vision biaisée de leur comportement.

Dans un article intitulé "*Medicaid is worse than no coverage at all*" (Gottlieb 2011) , Scott Gottlieb affirmait que "*dozens of recent medical studies show that Medicaid patients suffer for it. In some cases, they'd do just as well without health insurance*". Plus particulièrement, il s'appuyait sur LaPar *et al.* 2010, qui montraient, statistiques à l'appui, que "*uninsured patients were about 25 percent less likely than those with Medicaid to have an 'in-hospital death'*". Le tableau 3.1 était ainsi donné.

	Medicare	Medicaid	Non assuré	Assurance
Mortalité en milieu hospitalier	4.4 %	3.7 %	3.2 %	1.3 %
Résection pulmonaire	4.3 %	4.3 %	6.2 %	2.0 %
Œsophagectomie	8.7 %	7.5 %	6.5 %	3.0 %
Colectomie	7.5 %	5.4 %	3.9 %	1.8 %
Pancréatectomie	6.1 %	5.8 %	8.4 %	2.7 %
Gastrectomie	10.8 %	5.4 %	5.0 %	3.5 %
Anévrisme de l'aorte	12.4 %	14.5 %	14.8 %	7.0 %
Remplacement de la hanche	0.4 %	0.2 %	0.1 %	0.1 %
Pontage aorto-coronarien	4.0 %	2.8 %	2.3 %	1.4 %
Nombre de cas	491,829	40,259	24,035	337,535
Âge (années)	73.5 ± 8.6	49.8 ± 16.4	51.8 ± 12.8	55.5 ± 11.4
Femmes	49.6 %	48.8 %	35.8 %	39.7 %
Durée du séjour (jours)	9.5 ± 0.1	12.7 ± 0.4	10.1 ± 0.3	7.4 ± 0.1
Coût Total ($)	$76,374 \pm 53.1$	$93,567 \pm 251.4$	$78,279 \pm 231.0$	$63,057 \pm 53.0$
Localisation rurale	10.1 %	8.5 %	9.8 %	6.6 %

Table 3.1 – Mortalité à l'hôpital pour tous les patients subissant une opération chirurgicale majeure, par groupe de payeurs principaux (source LaPar *et al.* 2010, Tables 4, 5).

Plusieurs études ont révélé que les patients bénéficiant de Medicaid et souffrant d'affections spécifiques (par exemple un cancer) ou ayant subi des traitements spécifiques (par exemple une transplantation pulmonaire ou une angioplastie coronaire) présentaient des résultats sanitaires nettement moins bons que les patients ayant une assurance privée, souffrant des mêmes affections, ou subissant les mêmes procédures. Par exemple, dans le cas d'interventions chirurgicales majeures, les patients non assurés avaient environ 25 % de chances de moins que les patients bénéficiant de Medicaid d'être victimes d'un décès à l'hôpital. Être assuré par Medicaid semblerait presque agraver le risque ! Le problème potentiel de ce type d'études est que les groupes de comparaison, les patients de Medicaid et les patients privés assurés ou non assurés, peuvent être victime d'un phénomène d'auto-sélection. Il est probable que de nombreux patients ne s'inscrivent à Medicaid qu'après, parfois longtemps après, l'apparition d'un problème médical grave. Dans ce cas, ceux qui choisissent de s'inscrire à Medicaid peuvent être plus malades que ceux qui ont une assurance privée ou ceux qui ne sont pas assurés. Dans un sens, cet effet d'auto-sélection est une forme de causalité inverse : être malade pousse les gens à s'inscrire à Medicaid, et non l'inverse. C'est le soucis des données d'observations.

En 2008, en raison de contraintes budgétaires très fortes, l'Oregon s'est retrouvé avec une liste d'attente pour Medicaid de 90 000 personnes et seulement assez d'argent pour couvrir 10 000 d'entre elles. L'État a donc créé une loterie pour sélectionner, au hasard, les personnes qui pourraient bénéficier de Medicaid, recréant ainsi les conditions préalables nécessaires[41] à

41. La réalité fut toutefois un peu plus complexe, un grand nombre de les gagnants de la loterie n'étaient pas éligibles pour Medicaid ou ont choisi de ne pas soumettre leurs papiers pour s'inscrire au programme.

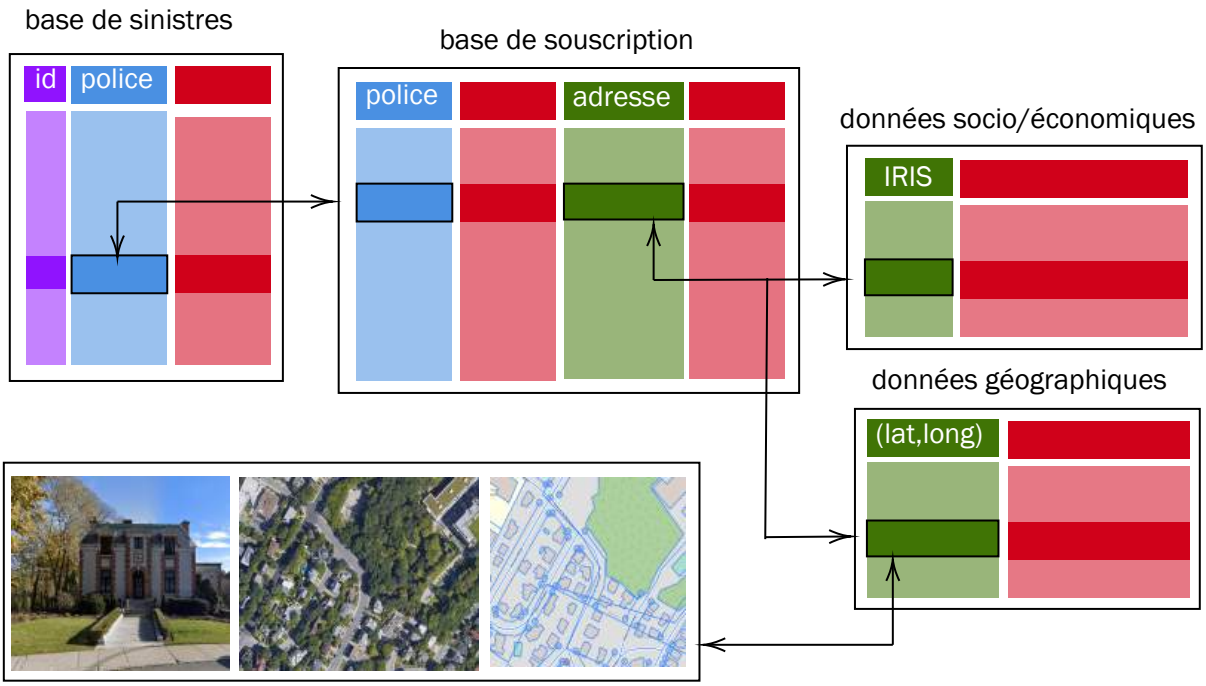
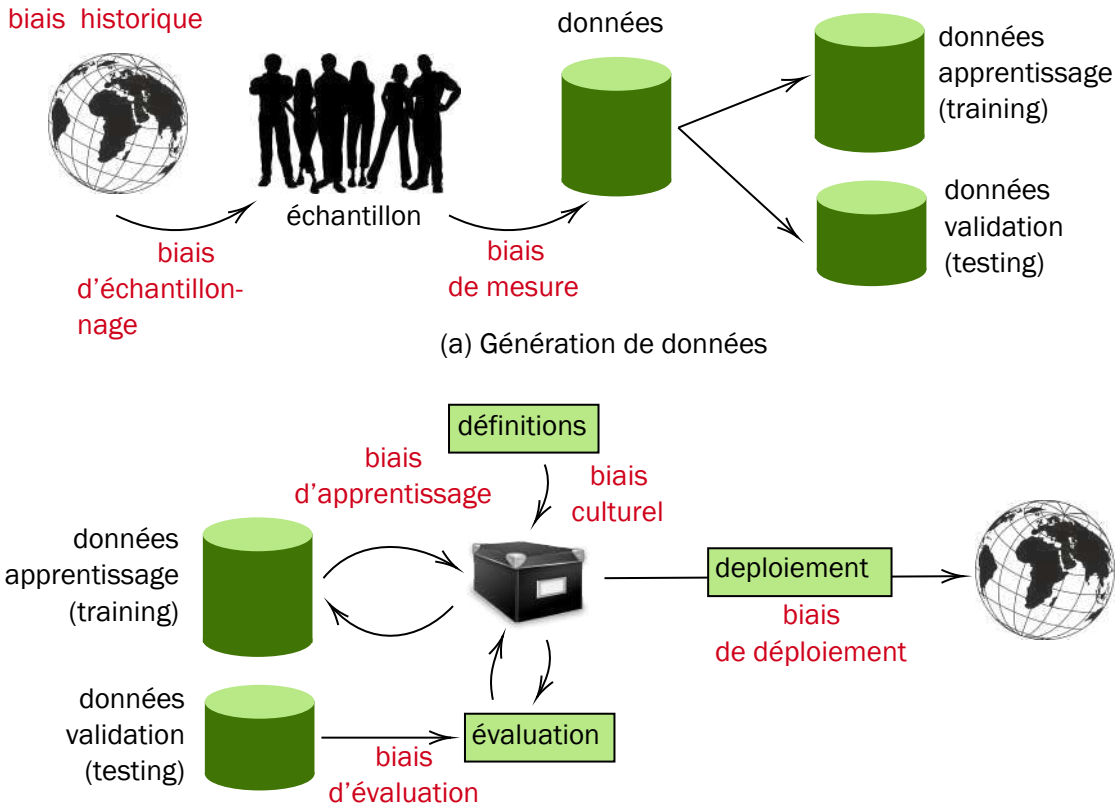


Figure 3.1 – Les bases de données d'un assureur, avec une base de souscription (au centre), avec une ligne par police d'assurance. Cette base sera couplée à la base de sinistres, qui contient tous les sinistres, et un numéro de police associé. D'autres données pourront être utilisées pour enrichir la base, par exemple sur la base de l'adresse d'habitation de l'assurée, avec soit des informations socio-économiques (statiques agrégées par quartier, sur la richesse, le nombre de crimes, etc.), mais aussi d'autres informations extraites de cartes, d'images satellites (distance à la borne incendie la plus proche, présence d'une piscine, etc.) En assurance automobile, il est possible de trouver la valeur d'une voiture à partir de ses caractéristiques, etc.

de véritables expériences, expériences dites "*randomisées*" : en comparant avec le groupe de contrôle (les personnes qui n'ont pas eu accès à Medicaid), A. Finkelstein *et al.* 2012 notent "*the treatment group had substantively and statistically significantly higher health care utilization (including primary and preventive care as well as hospitalizations), lower out-of-pocket medical expenditures and medical debt (including fewer bills sent to collection), and better self-reported physical and mental health*". Ces expériences, souvent difficile à mettre en œuvre (pour des raisons financières, mais aussi parfois éthiques), permettent justement de contourner le biais des données administratives.

3.2 Les données en assurance

Traditionnellement, les entreprises d'assurance utilisent deux bases : une base de souscription (une ligne correspond à une police, avec des renseignements sur l'assuré, sur le bien assuré, etc.) et une base de sinistres (une ligne correspond à un sinistre, avec le numéro de police de l'assuré, et la dernière vision des dépenses associées), comme sur la Figure 3.1. Ces deux bases sont liées



(a) Génération de données

(b) Construction d'un modèle

Figure 3.2 – Biais dans la génération des données, et dans la construction de modèle (librement inspiré de Suresh et Guttag 2019).

par le numéro de police. Ces dernières années, cependant, les entreprises se sont de plus en plus appuyées sur des données obtenues à partir d'un grand nombre de sources externes. Ces données concertant le bien assuré, avec des informations sur le modèle de voiture, ou sur la maison, obtenue à partir de l'adresse, comme sur la Figure 3.1. L'adresse permettait historiquement d'avoir des informations (agrégées) sur le quartier, avec des nombres d'infractions, sur des inondations passées, sur la distance à la caserne de pompiers la plus proche, etc. On peut aussi utiliser des images satellites (via Google Earth) ou les informations d'OpenStreetMap. Et les assureurs s'appuient sur des données qui s'étendent de plus en plus, avec des capteurs déployés partout, dans la voiture, ou les téléphones cellulaires, comme le rappellent Prince et Schwarcz 2019. Cette explosion des données laisse ouverte la question de savoir si un zoom de plus en plus granulaire sur la vie des assurés peut induire une tarification du risque plus précise.

L'évaluation des coûts est un processus technique qui implique l'analyse statistique de l'historique des pertes d'un assureur afin de prévoir les coûts de groupes d'assurés particuliers. La tarification, quant à elle, est un processus commercial par lequel une police est proposée à une certaine prime sur la base d'un large éventail de calculs concernant différents clients, leur coût d'acquisition, leur propension à acheter et leur propension à changer de fournisseur. Par exemple, les nouveaux clients qui effectuent des recherches approfondies et lisent les informations sur la police peuvent

se voir proposer des prix plus bas que les clients fidèles, comme le notait Minty 2016. Les perfectionnements techniques en matière de calcul des coûts ne signifient pas nécessairement que les prix refléteront plus précisément le risque, car les décisions en matière de tarification impliquent aussi toujours des considérations commerciales et marketing, comme l'a fait valoir Swedloff 2014. Les assureurs automobiles savent depuis longtemps que le risque est très fortement corrélé au comportement et à la personnalité de l'automobiliste, mais cette certitude qui ne leur a servi à rien pendant de longues années comme le soulignent Lancaster et Ward 2002, car ces informations étaient souvent non-observables.

Hand 2020 parlait de *dark data*, en notant que toutes les données avaient une face cachée qui pouvait potentiellement engendrer un biais. Nous allons formaliser dans les sections suivantes ces notions de biais, dont une partie peuvent se visualiser sur la Figure 3.2, inspirée de Suresh et Guttag 2019. Le "biais historique", c'est celui qui existe dans le monde tel qu'il est. C'est le biais évoqué dans Garg *et al.* 2018 dans la contextualisation en analyse textuelle ("*word embedding*") où la vectorisation reflète les biais existant entre les hommes et les femmes (le mot "*nurse*" (non genré) est souvent associé à des mots associés aux femmes), ou envers des minorités. Le "biais d'échantillonnage" est celui mentionné dans Ribeiro *et al.* 2016, où un algorithme de classification est entraîné, pour distinguer les chiens des loups, sauf que toutes les images de loups sont prises dans la neige, et l'algorithme se contente de regarder l'arrière plan de l'image pour attribuer un label. Pour le "biais de mesure", Dressel et Farid 2018 évoquent la récidive en justice prédictive qui est parfois mesurée non pas comme une nouvelle condamnation, mais comme une seconde arrestation. Le "biais culturel" (appelé "biais d'agrégation" dans Suresh et Guttag 2019) correspond au problème suivant : "*a particular dataset might represent people or groups with different backgrounds, cultures or norms, and a given variable can mean something quite different across them*". On peut penser à l'ironie en analyse textuelle, ou une référence culturelle (que l'algorithme ne peut comprendre). Hooker *et al.* 2020 notent que "*compression amplifies existing algorithmic bias*" (où la compression s'apparente à l'élagage dans les arbres, lorsque l'on essaye de simplifier les modèles). On peut aussi penser à Bagdasaryan *et al.* 2019 qui notaient que les techniques d'anonymisation des données pouvaient poser problème "*differential privacy training mechanisms such as gradient clipping and noise addition have disproportionate effect on the underrepresented and more complex subgroups*". On parlera alors de "biais d'apprentissage". Le biais d'évaluation se produit lorsque les données de référence utilisées pour une tâche particulière ne sont pas représententives. On peut penser ici aux algorithmes de reconnaissance faciale entraînés sur population très différentes de la vraie population. Buolamwini et Gebru 2018 notent que les femmes à la peau sombre ("*dark-skinned women*") représentent 7.4 % de la base Adience, et ce défaut de représentativité de certaines populations peut poser des problèmes (par exemple pour entraîner un algorithme à détecter des cancers de la peau). Finalement, le "biais de déploiement" désigne le décalage entre le problème qu'un modèle est censé résoudre et la façon dont il est effectivement utilisé. C'est ce que montrent Collins 2018 ou Stevenson 2018, décrivant les conséquences néfastes des outils d'évaluation des risques pour la "détermination actuarielle des peines" (on parlera d'"*actuarial sentencing*"), notamment la justification d'une incarcération accrue sur la base de caractéristiques individuelles.

Biais de décision, par Olivier l'Harridon [42]

Les littératures développées en économie comportementale et en psychologie du risque ont mis en évidence de nombreux "biais" qui affectent les décisions prises en présence d'incertitude. Typiquement, un biais de décision correspond à une déviation observée en référence à une norme de comportement ou de jugement rationnel. Traditionnellement, la théorie économique fournit, avec la théorie de l'utilité espérée de von Neumann, Morgenstern et Savage et le théorème de Bayes, une norme de comportement rationnel face à l'incertitude, que les probabilités associées aux évènements soient connues ou inconnues. Tout écart de comportement constaté par rapport à cette norme clairement définie est en général qualifié de biais de décision face à l'incertitude, même si cet écart peut être justifié par le contexte particulier de décision, rapidité de la décision à prendre, ou manque d'information probabiliste.

Au rang des biais de décision sont comptés ainsi de nombreuses heuristiques de choix qui sont des raccourcis mentaux correspondant à des opérations mentales intuitives et rapides. Il est important de garder à l'esprit que ces raccourcis peuvent être justifiés par la nécessite d'obtenir une réponse satisfaisante, ayant l'avantage d'être rapide sans être nécessairement optimale pour autant. La plus célèbre heuristique de choix dans le domaine de l'incertitude est certainement l'heuristique de représentativité. Elle se produit lorsque les individus mésestiment la fréquence d'un évènement par une généralisation abusive d'un évènement passé similaire. La traduction directe de l'heuristique de représentativité est l'emploi de stéréotypes pour réaliser des prévisions : les évènements qui n'ont été observés que sur un échantillon limité sont généralisés à l'ensemble d'une population, les évènements extrêmes sont banalisés, les phénomènes pour lesquels de plus amples détails sont connus sont considérés comme plus probables, les données sur les événements récents sont favorisées au détriment des données initiales sur la situation etc. Cette dernière caractéristique peut-être renforcée par l'heuristique de disponibilité, qui traduit le fait que les individus ont une meilleure mémoire des évènements qui sont plus facilement disponibles dans leurs représentations, au sens qu'ils sont marquants pour eux, tels que les évènements fréquents, les évènements plus récents, ou encore les évènements de plus grande ampleur. Dernière heuristique majeure identifiée par la littérature, l'heuristique d'ancrage désigne la tendance à fixer les croyances à un certain niveau et à conserver cette ancre dans les évaluations ultérieures de l'incertitude, conduisant à un biais de conservatisme. L'emploi de ces heuristiques est fondamentalement lié au processus de décision : en proposant des retours d'expérience fréquents sur les décisions prises dans le passé, ou en insistant sur l'attention portée à la décision, l'impact de ces heuristiques sur la prise de décision peut être significativement réduit.

Aux côtés de ces heuristiques, les biais motivationnels sont plus délicats à envisager, notamment en termes d'interventions possibles pour les réduire. Ainsi, l'économie comportementale a montré que les individus ont tendance à établir une comptabilité mentale séparée pour chaque petits risques auxquels ils sont soumis, sans envisager leurs compensations potentielles. Par ailleurs, même pour des petits risques, les individus sont nettement plus sensibles à ceux qui conduisent à des pertes qu'à ceux qui conduisent à des gains, démontrant une importante aversion aux pertes. Les individus ont ainsi tendance à se sur-assurer pour ces petits risques mais également à choisir des contrats à faible franchise même s'ils sont moins avantageux pour eux. À noter que l'aversion aux pertes génère également une forme de conservatisme dans le risque en encourageant les individus à rester dans leur statu-quo, leur point de référence, plutôt que de risquer des pertes, même pour bénéficier de perspectives plus importantes de gains. Dans un ordre d'idées différent, les décisions face au risque sont également biaisées par la forte tendance des individus à surestimer les petites probabilités, et à sous-estimer les fortes probabilités, sources importante d'optimisme et de pessimisme quant aux visions subjectives de la chance et de l'incertitude.

Par ailleurs, l'absence d'information, ou la nécessité d'en acquérir avant de prendre une décision se traduisent également par un certain nombre de biais de décision. Par exemple, les individus peuvent réagir fortement à l'ambiguïté, typiquement lorsqu'ils ont conscience de l'absence d'une information qui pourrait être disponible, mais qui ne leur ait pas donnée. Certains individus ont également des difficultés à intégrer des nouvelles informations, positives ou négatives, qu'ils apprennent sur les facteurs de risque, et ces difficultés peuvent avoir des conséquences particulièrement importantes lorsqu'elles touchent à des facteurs de risque affectant la santé, tels que l'hérédité, le tabagisme, ou l'information sur les risques cardiovasculaires.

Enfin, des biais de décision se produisant dans des domaines connexes à l'incertitude, typiquement la perception du présent et du futur qui peuvent interagir fortement avec les décisions prises dans le risque. Ainsi le biais d'immédiateté qui se traduit par une forte variation de l'impatience tournée vers les conséquences immédiates, couplé à l'aversion aux pertes, conduit les individus à surévaluer les coûts présents au détriment du poids des bénéfices futurs et à négliger la prévention et la prévisibilité, que ce soit dans le domaine monétaire ou dans le domaine de la santé.

3.3 Biais de variable omise et paradoxe de Simpson

Le biais de la variable omise se produit lorsqu'un modèle de régression est ajusté sans tenir compte d'une variable (prédictive) importante. Pradier 2011 notait par exemple que des manuels d'actuariat, comme Depoid 1967, affirment que "*la prime pure des femmes sur le marché nord-américain serait égale à celle des hommes si on la conditionnait par le kilométrage*". Mais cette dernière est a priori non-observée.

Variable omise dans un modèle linéaire

La sous-identification correspond au cas où le vrai modèle serait $y_i = \beta_0 + \boldsymbol{x}_1^\top \boldsymbol{\beta}_1 + \boldsymbol{x}_2^\top \boldsymbol{\beta}_2 + \varepsilon_i$ (sous une forme vectorielle classique) mais le modèle estimé est $y_i = b_0 + \boldsymbol{x}_1^\top \boldsymbol{b}_1 + \eta_i$ (autrement dit, les variables contenues dans le vecteur $\boldsymbol{x}_2$ ne sont pas utilisées dans la régression). L'estimateur du maximum de vraisemblance de $\boldsymbol{b}_1$ est (avec l'écriture matricielle classique en économétrie, comme dans Davidson et MacKinnon 2004 ou plus récemment Charpentier, Flachaire *et al.* 2018)

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

de telle sorte que $\mathbb{E}[\widehat{\boldsymbol{b}}_1] = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_{12}$, le biais (ce que nous avons noté $\boldsymbol{\beta}_{12}$) étant nul uniquement dans le cas où $\boldsymbol{X}_1^\top \boldsymbol{X}_2 = \boldsymbol{0}$ (c'est-à-dire $\boldsymbol{X}_1 \perp \boldsymbol{X}_2$) : on retrouve ici une conséquence du théorème de Frisch-Waugh. Si on simplifie un peu, supposons que le véritable modèle sous-jacent des données

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

où x_1 et x_2 désignent des variables explicatives, y est la variable cible, et ε est un bruit aléatoire. Le modèle estimé en enlevant x_2 donne

$$\widehat{y} = \widehat{b}_0 + \widehat{b}_1 x_1.$$

On peut penser à une variable importante manquante x_2, ou au cas où x_2 est une variable protégée. Les estimations des coefficients de régression obtenus par moindre carrés sont (généralement) biaisées, dans le sens où

$$\widehat{b}_1 = \frac{\widehat{\mathsf{cov}}[x_1, y]}{\widehat{\mathsf{Var}}[x_1]} = \frac{\widehat{\mathsf{cov}}[x_1, \beta_0 + \beta_1 x_1 + \beta x_2 + \varepsilon]}{\widehat{\mathsf{Var}}[x_1]},$$

soit

$$\widehat{b}_1 = \beta_1 \cdot \underbrace{\frac{\widehat{\mathsf{cov}}[x_1, x_1]}{\widehat{\mathsf{Var}}[x_1]}}_{=1} + \beta_2 \cdot \frac{\widehat{\mathsf{cov}}[x_1, x_2]}{\widehat{\mathsf{Var}}[x_1]} + \underbrace{\frac{\widehat{\mathsf{cov}}[x_1, \varepsilon]}{\widehat{\mathsf{Var}}[x_1]}}_{=0} = \beta_1 + \beta_2 \cdot \frac{\widehat{\mathsf{cov}}[x_1, x_2]}{\widehat{\mathsf{Var}}[x_1]}.$$

Lorsque x_2 est omise, $\widehat{b}_1$ est biaisé d'autant plus que x_1 et x_2 sont corrélées. Ainsi, dans la plupart des cas réalistes, non seulement la suppression de la variable sensible (si $x_2 = p$) ne rend pas les

42. Professeur à l'Université de Rennes 1.

modèles de régression équitables, mais au contraire, une telle stratégie est susceptible d'amplifier la discrimination. Par exemple, en économie du travail, si les immigrants ont tendance à avoir un niveau d'éducation plus faible, alors le modèle de régression "punirait" encore plus le faible niveau d'éducation en offrant des salaires encore plus bas aux personnes ayant un faible niveau d'éducation (qui sont principalement des immigrants). Žliobaite et Custers 2016 suggèrent qu'une meilleure stratégie pour assainir les modèles de régression consisterait à apprendre un modèle sur des données complètes incluant la variable sensible, puis à supprimer la composante contenant la variable sensible et à la remplacer par une constante qui ne dépend pas de la variable sensible. Une étude sur la prévention de la discrimination pour la régression, Calders et Žliobaitė 2013, est liée au sujet, mais avec un objectif différent. Notre objectif est d'analyser le rôle de la variable sensible dans la suppression de la discrimination, et de démontrer qu'il est nécessaire de l'utiliser pour la prévention de la discrimination. Calders et Žliobaitė 2013 utilisent, bien sûr, la variable sensible pour formuler des contraintes de non-discrimination qui sont appliquées pendant l'ajustement du modèle. Mais une discussion sur le rôle de la variable sensible n'est pas le sujet de leur étude. Des approches similaires ont été discutées dans la modélisation économique, comme dans Pope et Sydnor 2011, où l'accent a été mis sur l'assainissement des modèles de régression, notre objectif dans ce document est sur les implications pour les réglementations des données.

Pour revenir à notre exemple, il existe des cas où $\widehat{b}_1 < 0$ (par exemple) alors que dans le vrai modèle, $\beta_1 > 0$. On parle alors de paradoxe (de Simpson) ou de corrélation fallacieuse (en inférence écologique) au sens où le sens de l'impact d'une variable prédictive n'est pas clair.

Admission scolaire et discrimination positive

Si les exemples de tels paradoxes sont nombreux (Alipourfard *et al.* 2018 dressent une liste importante), le paradoxe de Simpson a été observé tout d'abord sur les admissions universitaires, dans Bickel *et al.* 1975, comme le Tableau 3.2 l'atteste.

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

Table 3.2 – Statistiques d'admission sur les six plus gros programmes gradués à Berkeley, avec le nombre d'admis / le nombre de dossiers reçus ~ le pourcentage d'admission. Les chiffres en gras indique, par ligne qui des hommes ou des femmes ont le taux d'admission le plus important. La colonne proportion indique les proportions hommes-femmes, dans les soumissions de candidatures. Le total correspond aux 12763 candidatures dans 85 programmes gradués; les six plus gros programmes sont détaillés en dessous, et la ligne "top 6" est le total de ces six programmes (soit 4526 candidatures).

Mathématiquement, il n'y a pas vraiment de paradoxe, au sens où

$$\frac{a_1}{c_1} < \frac{a_2}{c_2} \text{ et } \frac{b_1}{d_1} < \frac{b_2}{d_2} \nLeftrightarrow \frac{a_1+b_1}{c_1+d_1} < \frac{a_2+b_2}{c_2+d_2}.$$

La conclusion de Bickel *et al.* 1975 souligne "*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*". Autrement dit, la source des préjugés sexistes dans les admissions était un problème de filière : sans qu'il y ait de faute de la part des départements, les femmes ont été "écartées par leur socialisation" qui s'est produite à un stade antérieur de leur vie.

Survie au naufrage du Titanic

Une autre illustration se retrouve dans les données du Titanic, plus particulièrement quand on utilise des informations sur les membres d'équipage aux passagers. La Table 3.3 montre le même paradoxe, dans le contexte de la survie suite au naufrage.

	Total	Femmes	Hommes
troisième classe	$181/709 \sim$ **25.5** %	$106/216 \sim 49.1$ %	$75/493 \sim 15.2$ %
membre d'équipage	$211/890 \sim 23.7$ %	$20/\ 23 \sim$ **86.9** %	$191/867 \sim$ **22.0** %

Table 3.3 – Statistiques de survie pour les passagers du Titanic et les membres d'équipages (x_1) et leur genre (x_2).

Mais regardons les taux de survie des hommes et des femmes séparément, tels que présentés dans la Table 3.3. Pour les hommes, parmi l'équipage, il y avait 867 hommes, dont 191 ont survécu, soit un taux de 22 %. Parmi les passagers de troisième classe, 493 étaient des hommes, et 75 ont survécu, soit un taux de 15.2 %. Pour les femmes, parmi l'équipage, il y avait 23 femmes, et 20 d'entre elles ont survécu, soit un taux de 87 %. Et parmi les passagers de troisième classe, 216 étaient des femmes, et 106 ont survécu, soit un taux de 49 %. Autrement dit, pour les hommes et les femmes séparément, les membres d'équipage avaient un taux de survie plus élevé que les passagers de troisième classe ; mais dans l'ensemble, les membres d'équipage ont un taux de survie inférieur à celui des passagers de troisième classe. Comme pour les admissions, il n'y a pas de faute de calcul, ou de piège. Il y a simplement une erreur d'interprétation, car le genre x_2 et le statut x_1 (passager/membre d'équipage) ne sont pas des variables indépendantes, tout comme le genre x_2 et la survie y, qui ne sont pas indépendantes. En effet, alors que les femmes représentent 22 % de la population totale, elles représentent plus de 50 % des survivants... et 2.5 % du personnel d'équipage.

Paradoxe de Simpson en assurance

Dans un contexte démographique, Cohen 1986 s'est intéressé à la mortalité au Costa Rica et en Suède (la Suède était alors connue, et mise en avant, pour son excellente espérance de vie). Sans

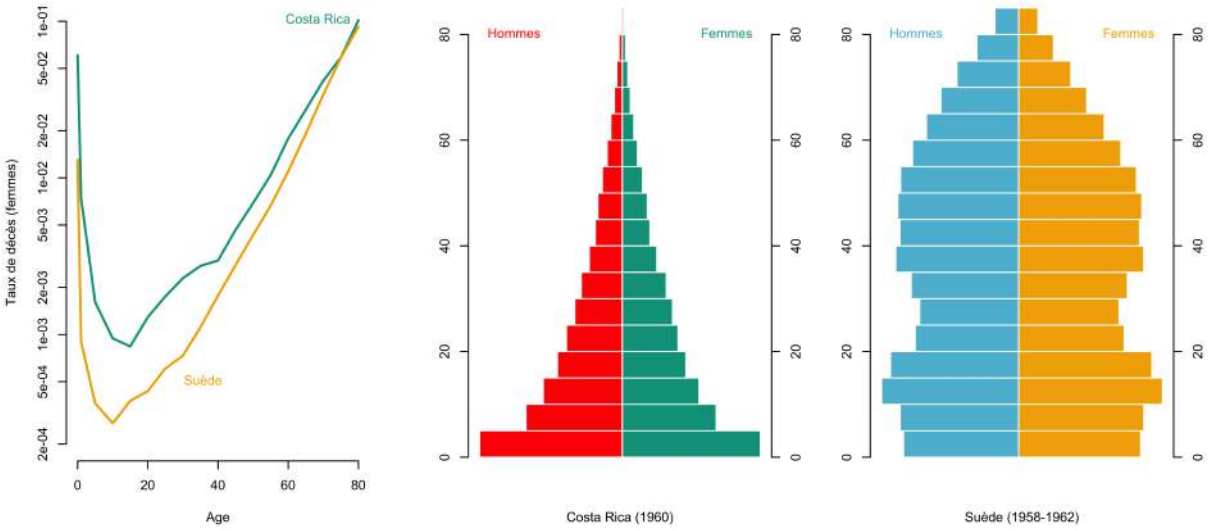


Figure 3.3 – Taux de mortalité annuelle des femmes, au Costa-Rica et en Suède, et pyramide des âges des deux pays (inspiré de Cohen 1986, données Keyfitz, Flieger *et al.* 1968).

surprise, il a trouvé qu'en 1960, le taux de mortalité des femmes dans toutes tranches d'âge était supérieur au Costa Rica, par rapport à la Suède, comme le montre la Figure 3.3. Pourtant, le taux de mortalité général des femmes au Costa Rica était inférieur à celui de la Suède, avec un taux de mortalité de 8.12‰ au Costa-Rica, contre 9.29‰ en Suède. L'explication est liée au paradoxe de Simpson, et elle provient de la structure différente des populations. La population du Costa Rica est beaucoup plus jeune en moyenne que celle de la Suède, et donc les jeunes classes d'âge (qui ont un faible taux de mortalité) pèsent plus dans la moyenne pour le Costa Rica que pour la Suède, conduisant à un taux de mortalité global assez faible au Costa Rica, malgré un taux assez mauvais dans chaque tranche d'âge, comme on le voit sur la pyramide des âges, à droite de la Figure 3.3.

Davis 2004 étudiait la relation entre le nombre d'accidents piéton-véhicule, et la vitesse moyenne des véhicules, en divers endroits d'une ville. Plus précisément, il s'agissait d'évaluer l'intérêt de limiter plus sévèrement la vitesse autorisée des véhicules en ville, et de façon inattendue, le modèle montrait qu'en faisant passer la limite de 30 miles par heure à 25, le nombre d'accidents augmenterait. L'explication est qu'une agrégation malheureuse des données (qui ne prenait pas en compte que le nombre d'accidents était bien plus rare dans les zones résidentielles, par exemple) engendrait une conclusion paradoxale (on s'attendrait à ce que le nombre d'accidents diminue quand la vitesse limite est réduite).

Paradoxe écologique

Des problèmes similaires au paradoxe de Simpson se présentent également sous d'autres formes. Par exemple, le paradoxe écologique (analysé par Freedman 1999, Gelman 2009, King *et al.* 2004) décrit une contradiction entre une corrélation globale et une corrélation au sein de groupes. Un exemple classique a été décrit par Robinson 1950. La corrélation entre le pourcentage de la population née à l'étranger et le pourcentage de personnes alphabétisées dans les 48 États des États-Unis en 1930 était de +53 %. Cela signifie que les États qui comptaient une proportion

plus élevée de personnes nées à l'étranger étaient également plus susceptibles d'avoir un taux d'alphabétisation plus élevé (plus de personnes sachant lire, en anglais américain du moins). Superficiellement, cette valeur suggère que le fait d'être né à l'étranger signifie que les gens sont plus susceptibles d'être alphabétisés. Mais si l'on regarde au niveau des États, le tableau est assez différent. Au sein des États, la corrélation moyenne est de -11 %. La valeur négative indique qu'être né à l'étranger signifie que les gens ont moins de chances d'être alphabétisés. Si les informations au sein des États n'avaient pas été disponibles, une conclusion erronée aurait pu être tirée sur la relation entre le pays de naissance et l'alphabétisation.

On peut également mentionner, dans un contexte récent, le cas de la vaccination, comme l'ont noté Kugelgen *et al.* 2021. En reprenant les statistiques collectées par Morris 2021, on peut obtenir la Table 3.4, où l'efficacité est définie comme le taux d'infectés parmi les personnes non-vaccinées ramené au taux d'infectés parmi les vaccinés. S'il n'y a pas de paradoxe de Simpson stricto sensu, en revanche, on notera que ne pas segmenter en fonction de l'âge conduit à fortement sous-estimer l'efficacité du vaccin.

	Total	jeunes	personnes âgées
vaccinés	301/5634634~ 0.053‰	$11/3501118 \sim 0.003$‰	$290/2133516 \sim 0.136$‰
non vaccinés	214/1302912~ 0.164‰	$43/1116834 \sim 0.039$‰	$171/\ 186078 \sim 0.919$‰
efficacité	$3.1\times$	$13.0\times$	$6.7\times$

Table 3.4 – Statistiques de cas graves de COVID-19, en Israël, (source Morris 2021)

3.4 Biais d'auto-sélection et paradoxe de Berkson

La non-réponse est une forme d'autosélection où les individus refusent de faire partie de l'échantillon d'apprentissage. En fait, cette situation se retrouve de plus en plus dans les fichiers administratifs, comme le souligne Westreich 2012. Jusqu'à très récemment, bien souvent, nos données étaient stockées automatiquement, sans que nous le sachions, et sans que nous n'ayons notre mot à dire. Le Règlement général sur la protection des données (RGPD) de l'Union européenne a changé cela [43]. En application de ce principe, de nombreux pays ont voté des lois donnant la possibilité à ceux qui en font la demande de voir leurs données supprimées. Ce concept de *opting-out* est contraignant et peut fortement biaiser les données conservées. On peut penser à Dilley et Greenwood 2017 qui notaient que que nombre d'appels d'urgence (au 999) abandonnés au Royaume-Uni avaient plus que doublé en 2016. Comment tenir compte de ces appels qui n'ont pas abouti si on souhaite étudier sérieusement le sentiment d'insécurité ?

Ce problème d'auto-sélection peut aussi ressurgir lorsque l'on essaye de modéliser l'admission dans un programme universitaire (pour faire écho au problème présenté dans la section précédente).On considère ici trois variables indicatrices $(y_{1:i}, y_{2:i}, y_{3:i})$ où (1) la première variable indique si une personne a postulé dans un programme, (2) la seconde indique l'admission dans un programme, (3) et la troisième indique l'inscription dans un programme. Supposons de plus que les dossiers sont évalués sur la base de deux grandeurs $(x_{1:i}, x_{2:i})$, notées aussi $\boldsymbol{x}_i$. Dans les exemples

43. Comme vous le réalisez sans doute en raison de toutes les invitations à cocher des cases indiquant que vous comprenez et autorisez l'enregistrement des données personnelles vous concernant lorsque vous naviguez sur les sites internet.

venus des États-Unis, les deux grandeurs principalement étudiées sont le GPA score (*Grade Point Average*, allant souvent de 1 à 4, respectivement pour un D et un A ou un A+) et le SAT score (*Scholastic Aptitude Test*, allant de 400 à 1600). Pour simplifier, supposons que les deux scores sont normalisés sur une échelle de 0 à 100, comme sur la figure 3.4. Supposons que des règles très simples soient adoptées pour chacune des classes (j) : $y_{j:i} = \mathbf{1}[\boldsymbol{x}^\top \boldsymbol{\beta}_j > s_j]$. Par exemple aucun élève dont la somme $x_1 + x_2$ ne dépasse pas 60 ne postule ($y_{1:i} = \mathbf{1}[x_{1:i} + x_{2:i} > 60]$), tout élève dont la somme des scores dépasse 120 est admis ($y_{2:i} = \mathbf{1}[x_{1:i} + x_{2:i} > 120]$) et les élèves "trop bons" qui ne viennent finalement pas ($y_{3:i} = \mathbf{1}[x_{1:i} + x_{2:i} \in (120, 160]]$). Comme on peut le voir sur la figure 3.4, suivant les données utilisées, la corrélation entre les variables x_1 et x_2 peut fortement changer : globalement les variables sont fortement corrélées, positivement, mais sur la sous base des élèves inscrits dans le programme (autrement dit $\{i : y_{3:i} = 1\}$), les variables x_1 et x_2 sont fortement corrélées, mais négativement cette fois.

Cet exemple sur les admissions scolaires se retrouve dans beaucoup de cas, en assurance ou en finance. En risque de crédit, on essaye d'estimer la probabilité qu'un souscripteur ait un défaut de paiement. Mais il convient de s'interroger sur les données à disposition pour estimer cette probabilité (ou construire un score de crédit). La première barrière est l'auto-sélection, certaines personnes ne déposant pas de dossier de crédit, pour toutes sortes de raisons. Cette auto-sélection est d'autant plus gênante que l'on n'a souvent aucune information sur les personnes qui ne postulent pas [44]. On peut aussi penser aux radars pour repérer des zones où les excès de vitesse seraient fréquents : toutes les vitesses des voitures passant sur la route sont-elles mesurées et enregistrées, ou seulement les voitures ayant commis un excès de vitesse ? Les radars fonctionnent-ils en permanence, ou seulement aux horaires où des enfants jouent dans la rue ? L'appareil enregistre-t-il seulement la vitesse ou aussi d'autre comportements suspects ? On peut aussi penser à une base relativement populaire [45] des accidents corporels de la circulation routière, où est enregistrée "tout accident corporel de la circulation routière connu des forces de l'ordre faisant l'objet d'un Bulletin d'Analyse d'Accident Corporel (BAAC)". On peut imaginer que certains accidents ne figurent pas dans la base, et que certaines informations ne sont que partiellement reportées comme les informations en lien avec l'alcoolémie, tel que le soulignent Carnis et Lassarre 2019 : une information manquante signifie-t-il que le test était négatif ou qu'il n'a pas été fait ? Il est indispensable de savoir comment les données ont été collectées avant de commencer à les analyser. Les informations manquantes sont fréquentes aussi en assurance santé, les dossiers médicaux étant souvent relativement complexes, avec des multitudes de codes de procédures (qui varient d'un hôpital à l'autre, d'un assureur à l'autre). En France, la majorité des médicaments étant pré-empaquetés (par boites de 12, 24, 36), il est difficile de quantifier la "consommation" (réelle) de médicaments.

3.5 Biais de rétroaction et biais de déploiement

3.5.1 Biais de rétroaction et loi de Goodhart

Quand on veut mesurer la richesse d'une population, on peut penser aux enquêtes ou aux données fiscales. Les premières sont coûteuses et complexes à organiser. Les secondes sont biaisées, en

44. Pour poursuivre l'analogie, en risque de crédit, on retrouve les trois niveaux précédents, avec (1) ceux qui ne sollicitent pas de crédit, (2) ceux à qui l'établissement n'offre pas de crédit et (3) ceux qui ne sont pas intéressés par l'offre faite.

45. https://www.data.gouv.fr/fr/datasets/accidents-corporels-de-la-circulation-routiere/

i

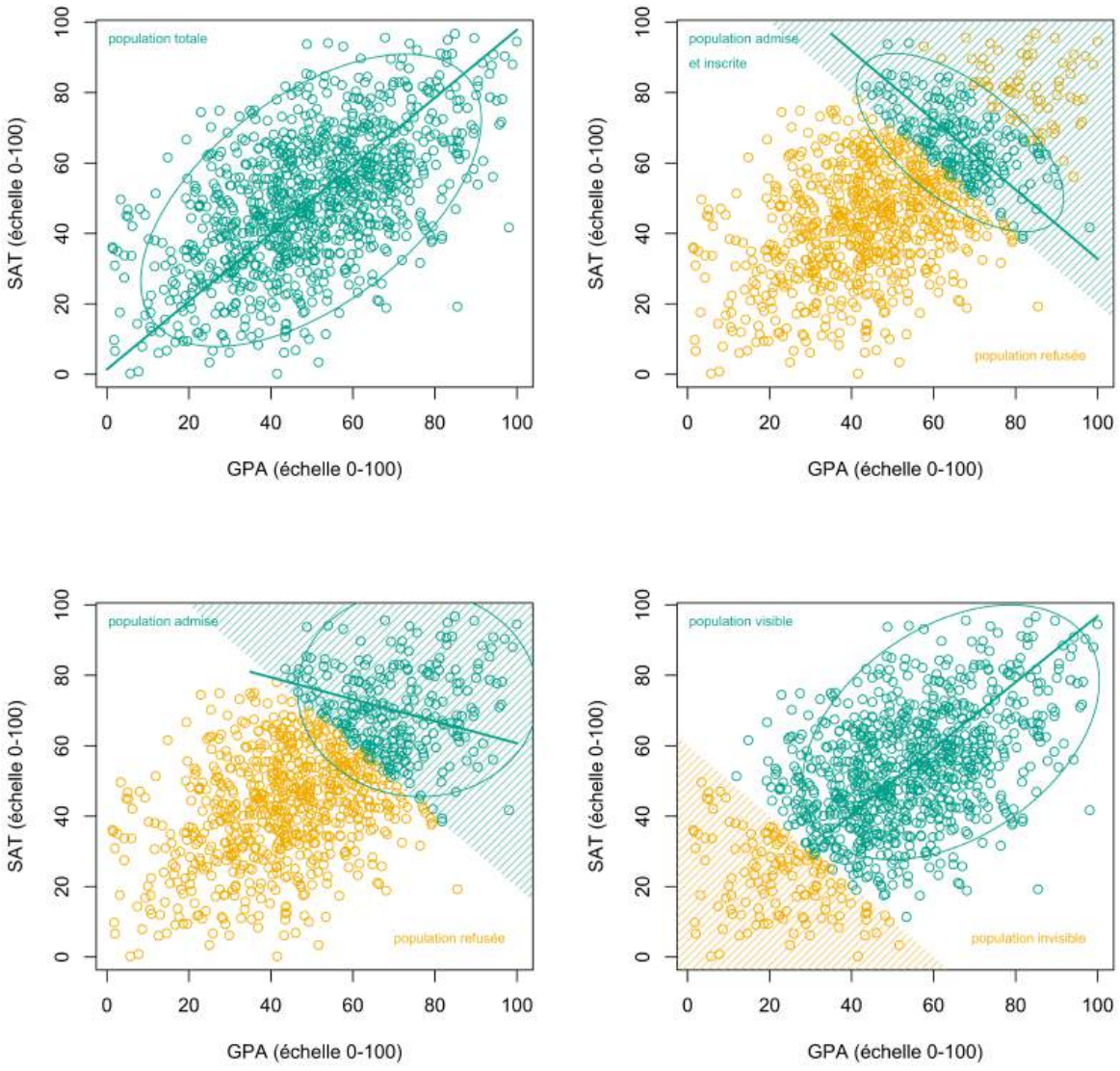


Figure 3.4 – Relation entre les deux variables, x_1 (GPA) et x_2 (SAT), en fonction de la population étudiée : population totale en haut à gauche ($r \sim 0.55$), population observable en haut à droite, i.e. élèves ayant postulé au programme ($r \sim 0.45$), population des élèves admis au programme en bas à gauche ($r \sim -0.05$), et population des élèves finalement inscrits au programme en bas à droite ($r \sim -0.7$) (source : *données fictives*).

particulier pour les très hauts revenus, qui pratiquent de l'optimisation fiscale (et qui vont changer au cours du temps). En la matière, l'imagination est grande. Par exemple, au Royaume-Uni, on peut éviter les droits de succession en empruntant sur un bien imposable (par exemple, votre maison) et en investissant le prêt dans un bien non imposable, tel qu'un bois ; on peut acheter un bien immobilier par l'intermédiaire d'une société *off-shore*, puisque les sociétés et les résidents non britanniques ne paient pas d'impôts au Royaume-Uni. Lorsque les lacunes d'un système fiscal sont découvertes et que les gens commencent à en faire un usage intensif, mais souvent cela ne fait que conduire à des structures encore plus élaborées qui présentent à leur tour leurs propres lacunes. C'est la loi de Goodhard.

Comme l'a énoncé Marilyn Strathern, la loi de Goodhart dit que "*when a measure becomes a target, it ceases to be a good measure*" (lorsqu'une mesure devient un objectif, elle cesse d'être une bonne mesure). Dans le domaine de la santé, aux États-Unis, Poku 2016 note qu'à partir de 2012, en vertu de la loi "*Affordable Care Act*" , Medicare a commencé à imposer des sanctions financières aux hôpitaux présentant des taux de réadmission de 30 jours "plus élevés que prévus". Dès lors, le taux moyen de réadmission à l'hôpital dans les 30 jours pour les bénéficiaires de l'assurance "*fee-for-service*" a diminué. Est-ce dû à une amélioration des efforts de transition et de coordination des soins menés par les hôpitaux, ou bien faut-il relier cette baisse à une hausse des séjours 'en observation' pendant la même période ? Car bien souvent, fixer un objectif sur la base d'une mesure précise (ici le taux de réadmission de 30 jours) rend cette grandeur totalement inutilisable pour quantifier le risque de retomber malade, mais a aussi une influence directe sur d'autres grandeurs (ici le nombre de séjours 'en observation'), qu'il devient alors difficile de suivre sur la durée. Sur internet, on demande de plus en plus à des algorithmes de trier du contenu, de juger le caractère diffamatoire ou raciste des tweets, de voir si une vidéo relève du "*deepfake*", de calculer un score de fiabilité associé à un compte Facebook, etc. Et nombreux sont ceux qui réclament une transparence des algorithmes, pour savoir comment ces scores sont créés. Malheureusement, comme le notait Dwoskin 2018 "*not knowing how [Facebook is] judging us is what makes us uncomfortable. But the irony is that they can't tell us how they are judging us — because if they do, the algorithms that they built will be gamed* " , exactement comme le suppose la loi de Goodhart.

Comme l'écrivait Desrosières 2016, "les indicateurs quantitatifs rétroagissent sur les acteurs quantifiés". Au printemps 2020, les chaînes télévisées d'information donnaient, en temps continu, le nombre de personnes en soins intensifs, et le nombre de morts dans les hôpitaux, mesures que l'on retrouvera ensuite sous forme de graphiques, mis à jour toutes les semaines, voire tous les soirs, sur des sites Internet dédiés. En cette période de crise, au plus fort de la saturation des hôpitaux, le National Health Service (NHS) en Angleterre avait demandé à chaque hôpital d'estimer ses capacités de lits, afin de réallouer les ressources globalement. Annoncer que peu de lits étaient disponibles était la meilleure stratégie pour obtenir davantage de financement. On peut alors s'interroger sur le niveau de saturation réelle du système, chaque hôpital ayant compris la règle, et manipulant la mesure à sa guise. Et tout aussi préoccupant, alors que les gouvernements se concentraient sur les hôpitaux (fournissant les données officielles utilisées pour construire la plupart des indicateurs), les maisons de retraite ont connu des hécatombes désastreuses, qui ont mis beaucoup de temps à être quantifiées, comme le raconte Giles 2020.

Mais, hormis les erreurs de données, l'idée des cartes de la criminalité soulève des problèmes plus subtils, liés aux données cachées et aux réactions. L'attention a été attirée sur ces problèmes lorsque le groupe d'assurance britannique Direct Line a réalisé une enquête qui a révélé que "*10 % de tous les adultes britanniques envisageraient certainement ou probablement de ne pas signaler*

un crime à la police parce qu'il apparaîtrait sur une carte de la criminalité en ligne, ce qui pourrait avoir un impact négatif sur leur capacité à louer/vendre ou réduire la valeur de leur propriété". Au lieu de montrer où les incidents se sont produits, les cartes risquent de montrer où les gens ne se soucient pas de les signaler. Ce n'est pas du tout la même chose, et toute personne qui prendrait des décisions sur la base de ces données pourrait facilement être induite en erreur. O'Neil 2016 rappelait aussi le biais de sélection dans les premiers systèmes de données télématiques : "*in these early days, the auto insurers' tracking systems are opt-in. Only those willing to be tracked have to turn on their black boxes. They get rewarded with a discount of between 5 and 50 percent and the promise of more down the road. (And the rest of us subsidize those discounts with higher rates.)*". Il y a plus d'un vingtaine d'années, ce problème avait déjà été mentionné, par exemple par Morrison 1996, qui rappelait (en citant Stein 1994), que certaines compagnies d'assurance, aux États-Unis, utilisaient les preuves de violence conjugale pour discriminer la victime en lui refusant l'accès à toute forme d'assurance. L'argument était le même que celui évoqué voilà plusieurs siècles pour interdire l'assurance sur la vie : "*there is some fear that if the beneficiary is the batterer, we would be providing a financial incentive, if it's life insurance, for the proceeds to be paid for him to kill her*" rapportait ainsi Seelye 1994. Ainsi, les victimes réalisaient que si elles demandaient une protection médicale ou policière, les dossiers qui en résulteraient pourraient compromettre leur assurabilité. En conséquence, Morrison 1996 affirme que les victimes pouvaient cesser de chercher de l'aide, ou de signaler les incidents de violence, afin de préserver leur couverture d'assurance.

Un autre exemple est celui des objets connectés, permettant enfin de connaître le comportement des assurés. Les assureurs automobiles savent depuis longtemps que le risque est très fortement lié au comportement de l'automobiliste, mais, comme le soulignaient Lancaster et Ward 2002, cette intuition n'a jamais pu être utilisée. "*As certainty replaces uncertainty*" pour reprendre les mots de Zuboff 2019, "*premiums that once reflected the necessary unknowns of everyday life can now rise and fall from millisecond to millisecond, informed by the precise knowledge of how fast you drive to work after an unexpectedly hectic early morning caring for a sick child or if you perform wheelies in the parking lot behind the supermarket*". Les assurés sont récompensés lorsqu'ils améliorent leur comportement de conducteur, "*relative to the broader policy holder pool*" comme l'affirmaient S. Friedman et Canaan 2014. Cette approche, parfois appelée "*gamification*", peut même inciter les automobilistes à changer leurs comportements, leurs risques. Jarvis *et al.* 2019 vont jusqu'à affirmer que "*insurers can eliminate uncertainty by shaping behavior*".

3.5.2 Biais de déploiement

Tel qu'évoqué sur la Figure 3.2, le biais de déploiement peut signifier qu'une fois confronté au marché, un "bon" modèle peut s'avérer désastreux. Formellement, cela signifie qu'un modèle estimé pour minimiser une erreur de précision sur une base d'apprentissage "biaisée" peut donner toutes sortes de résultats en le mettant en compétition face à d'autres modèles. Par exemple, sur la Figure 3.5, on suppose que chaque assureur dispose de données $(y_i, x_{1,i}, x_{2,i})$, où y désigne le coût d'un sinistre, x_1 est une caractéristique des assurés (schématiquement, $x_1 \in \{\square, \circ\}$, des carrés ou des ronds), et x_2 désigne l'âge des assurés. Et comme on le voit, x_1 et x_2 sont très corrélées (parfaitement corrélées dans cet exemple, mais on pourrait relâcher cette hypothèse sans trop de difficultés). Supposons de plus qu'il y a 2 assureurs, et que pour des raisons historiques, les carrés $\square$ étaient assurés par l'assureur A alors que les ronds $\circ$ étaient assurés par l'assureur B. Chaque assureur avait son modèle linéaire de tarification, $\pi_A(x_2) = \alpha_0 + \alpha_1 x_2$ et $\pi_B(x)_2 = \beta_0 + \beta_1 x_2$ (comme ils n'ont qu'un type de x_1, ils n'utilisent pas cette variable comme critère tarifaire). Comme le montre la droite de la Figure 3.5, chaque assureur avait un "bon" modèle, lui permettant d'être

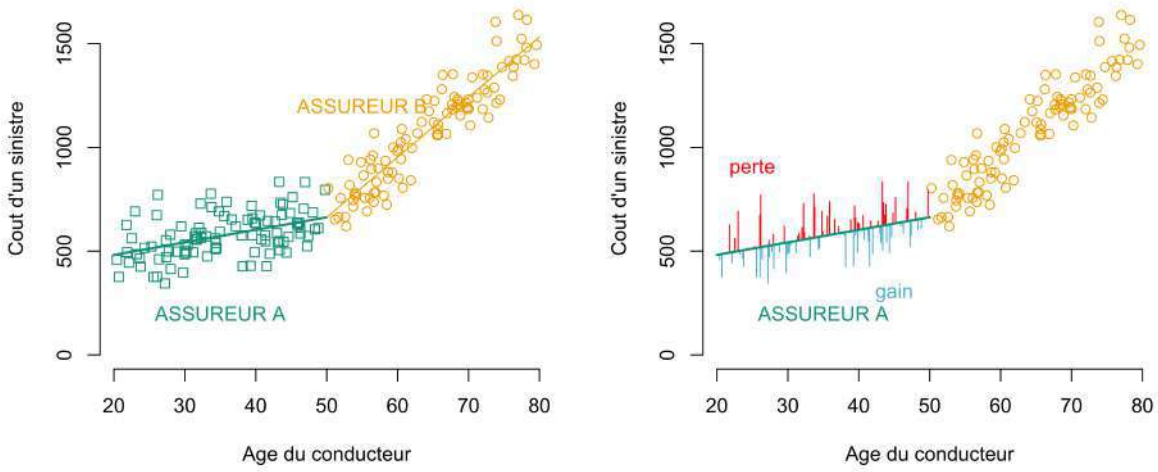


Figure 3.5 – Données $(y_i, x_{1,i}, x_{2,i})$, où y désigne le coût d'un sinistre, x_1 est une caractéristique des assurés ($x_1 \in \{\square, \circ\}$, des carrés ou des ronds), et x_2 désigne l'âge des assurés, avec deux assureurs ayant chacun un modèle (linéaire) $\pi_A(x_2) = \alpha_0 + \alpha_1 x_2$ et $\pi_B(x)_2 = \beta_0 + \beta_1 x_2$ (source : *données fictives*).

en moyenne à l'équilibre, au sens où

$$\sum_{i:\, x_{1,i}=\square} y_i \approx \sum_{i:\, x_{1,i}=\square} \pi_A(x_{2,i}) \text{ et } \sum_{i:\, x_{1,i}=\circ} y_i \approx \sum_{i:\, x_{1,i}=\circ} \pi_B(x_{2,i}).$$

Supposons maintenant soit que la concurrence devient plus sévère, ou simplement que les deux modèles, développés indépendemment soient lancé sur un même marché. L'avantage des modèles linéaires est de proposer des prix pour tous les âges, même s'ils n'ont pas été observés dans les données, comme on le voit sur la gauche de la Figure 3.6. On le voit, pour les carrés $\square$, $\pi_A(x_i) > \pi_B(x_i)$, autrement dit, tous les carrés devraient (rationnellement) choisir d'être assurés par l'assureur B. Et de manière symmétrique, pour les ronds $\circ$ on a $\pi_A(x_i) < \pi_B(x_i)$ autrement dit, tous devraient être assurés par l'assureur A. Aussi, lors du déploiement, les portefeuilles des deux assureurs vont radicalement changés. Et pire, comme on le voit sur la droite de la Figure 3.6, l'assureur A va perdre de l'argent sur presque tous ses assurés (ainsi que B).

3.6 Autres biais et "*dark data*"

En tentant de faire une typographie des biais des "*dark data*", Hand 2020 avait énuméré des dizaines d'autre biais existants. Au-delà des variables manquantes évoquées précédemment, il y a un biais de sélection qui peut être particulièrement important. En 2017, lors d'un des débats à la conférence NeurIPS sur l'interprétabilité [46], un exemple sur la détection de la pneumonie est mentionné : un réseau de neurones profonds est entraîné pour distinguer les patients à faible

46. Appelé "*The Great AI Debate : Interpetability is necessary for machine learning*", opposant Rich Caruana et Patrice Simard (pour) à Kilian Weinberger et Yann LeCun (contre) https://youtu.be/93Xv8vJ2acI.

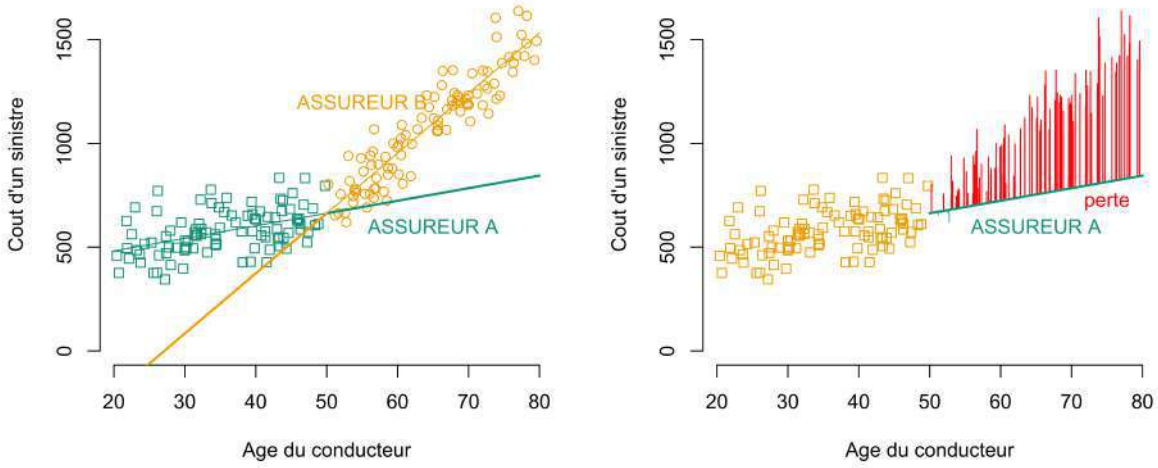


Figure 3.6 – Mise en compétition des modèles de la Figure 3.5, (source : *données fictives*).

risque des patients à haut risque, afin de déterminer qui traiter en premier. Le modèle était extrêmement précis sur les données d'entraînement. Après une inspection approfondie, il s'avère que le réseau neuronal a appris que les patients ayant des antécédents d'asthme présentaient un risque extrêmement faible et ne nécessitaient pas de traitement immédiat. Ce qui peut sembler contraire à l'intuition, car la pneumonie est une maladie pulmonaire et les patients asthmatiques ont tendance à y être plus sensibles (ce qui en ferait classiquement des patients à haut risque). En regardant plus en détail, les patients asthmatiques présents dans les données d'entraînement présentaient effectivement un faible risque de pneumonie, car ils avaient tendance à se faire soigner par des médecins beaucoup plus tôt que les patients non asthmatiques. En revanche, les personnes non asthmatiques ont tendance à attendre que le problème devienne plus grave avant de se faire soigner.

Le biais de survie est un autre type de biais relativement connu et documenté. L'exemple le plus connu est celui présenté par Mangel et Samaniego 1984 : pendant la seconde guerre mondiale, il a été demandé à des ingénieurs et des statisticiens (britanniques) comment renforcer les bombardiers qui subissaient le feu ennemi. Le statisticien Abraham Wald a commencé à collecter des données sur les impacts dans la carlingue. À la surprise générale, il a recommandé de blinder les endroits des appareils qui présentaient le moins de dommages. En effet, les avions utilisés dans l'échantillon présentaient un biais important : seuls les avions revenus du théâtre des opérations ont été pris en compte. S'ils ont pu revenir avec des trous au bout des ailes, c'est que ces parties sont suffisamment solides. Et comme aucun avion n'est revenu avec des trous au niveau des moteurs et des hélices, ce sont ces parties qu'il convenait de renforcer. Un autre exemple est celui des patients atteints d'un cancer avancé. Pour déterminer lequel des deux traitements est le plus efficace pour prolonger la vie, les patients seront répartis au hasard entre les deux traitements et les durées moyennes de survie dans les deux groupes seront comparées. Mais inévitablement, certains patients survivront longtemps - peut-être des décennies - et nous ne voudrions pas attendre des décennies avant de savoir quel traitement est le meilleur. Ainsi, l'étude sera probablement terminée avant que tous les patients ne soient décédés. Cela signifie que nous ne connaîtrons pas

les temps de survie des patients qui ont vécu au-delà de la date de fin de l'étude. Un autre soucis est qu'au cours du temps, des patients pourraient mourir d'une autre cause que le cancer. Et là encore, les données nous indiquant combien de temps ils auraient survécu avant de mourir "*du cancer*" sont manquantes. Enfin, des patients pourraient abandonner (pour des raisons sans rapport avec l'étude, ou pas). Et à nouveau, leurs durées de survie seraient à nouveau des données manquantes. Lié à cet exemple, on peut revenir sur un autre exemple important : pourquoi plus de personnes meurent de maladies liées à la maladie d'Alzheimer que par le passé ? une réponse peut sembler paradoxale : elle découle des progrès de la science médicale. Grâce aux progrès de la médecine, des personnes qui seraient mortes jeunes survivent aujourd'hui suffisamment longtemps pour être vulnérables à des maladies potentiellement longues, comme la maladie d'Alzheimer. Cela soulève toutes sortes de questions intéressantes, notamment sur les conséquences de l'allongement de la vie.

Un autre exemple plus complexe est lié à la définition des variables : la variable observée est-elle réellement la grandeur qu'on cherche à quantifier ? Un exemple assez classique est celui de la durée moyenne de détention d'une action [47] : il y a plus de dix ans, Creswell 2010 affirmait que la durée moyenne était de 11 secondes, alors que Zweig 2010 évoquait une durée moyenne de 17 mois. Aussi surprenant que cela puisse paraître, les deux avaient probablement raison, la différence était la manière dont cette durée moyenne est calculée (et définie). En effet, considérons une action (A) achetée en 2020, et revendue 3 ans plus tard, alors que pendant la même période, une action (B) a changé de main toutes les 5 secondes (et donc a été vendue et achetée $n = 18$ millions de fois). La durée moyenne de possession calculée sur les deux titres donne une durée moyenne $\overline{y}$ de l'ordre de 18 mois (la moyenne entre 3 ans et 5 secondes), alors que si la moyenne est calculée sur l'ensemble des $n + 1$ transactions, on obtient

$$\overline{y} = \frac{n \times 5 \text{ secondes} + 3 \text{ ans}}{n+1} \approx 10 \text{ secondes}.$$

Cet exemple peut facilement se généraliser, par exemple sur la durée moyenne d'occupation d'une chambre d'hôpital, ou des longueurs moyennes de trajets automobiles. On le voit, la manière dont sont calculées les différentes grandeurs n'est pas neutre, loin de là.

Parmi les autres biais classiques, on peut mentionner les erreurs dans les données qui sont souvent importantes. Comme le rappelait, Klein 1997 il existe des preuves solides que les éléments stockés dans les bases de données organisationnelles présentent un taux d'erreurs significatif. Si elles ne sont pas détectées lors de l'utilisation, les erreurs dans les données peuvent affecter de manière significative les résultats de l'entreprise. Comme l'affirmait le statisticien William Kruskal, "*a reasonably perceptive person, with some common sense and a head for figures, can sit down with almost any structured and substantial data set or statistical compilation and find strange-looking numbers in less than an hour*". Ainsi, un rapport sur les prix des médicaments publié dans le Times (Londres) du 26 mai 2018 mentionne la pharmacie du Shropshire, au Royaume-Uni, qui a été payée 6,030 £ pour un médicament qui aurait dû coûter 60.30 £, et une autre de Greenwich, au Royaume-Uni, qui a été payée 7,450 £ pour des analgésiques coûtant 74.50 £, selon Kenber *et al.* 2018, cité par Hand 2020.

47. Au sens financier du terme, autrement dit le titre de propriété d'une société de capitaux

Chapitre 4

Équité et modèles prédictifs

Dans un problème de classification, à partir des données prédictives $\boldsymbol{x}$, on va utiliser un modèle d'apprentissage (ou une régression logistique) pour calculer un score $s(\boldsymbol{x})$ qui sera utilisé pour prédire un résultat binaire $y \in \{0,1\}$. L'affectation aux classes pour $\widehat{y}$ se fera en fonction de l'interprétation de y et s,

- en risque de crédit, y désigne un défaut de remboursement d'un crédit ($y = 1$ en cas de défaut), et souvent le "*score de crédit*" (au sens FICO [48]) sera d'autant plus élevé que le risque sera faible. Autrement dit s et y (et $\widehat{y}$) évoluent en sens inverse : avec cette interprétation du score, $\widehat{y} = \mathbf{1}(s(\boldsymbol{x}) < \text{seuil})$,
- en assurance-décès, y désigne un décès ($y = 1$ en cas de décès) et le score s sera la probabilité de décéder. Autrement dit s et y (et $\widehat{y}$) évoluent dans le même sens : avec cette interprétation du score, $\widehat{y} = \mathbf{1}(s(\boldsymbol{x}) > \text{seuil})$,

Pour un problème de régression, on va prédire $y \in \mathbb{R}$ à partir d'un modèle $m(\boldsymbol{x})$, et on posera $\widehat{y} = m(\boldsymbol{x})$. La subtilité, dans le problème de classification, est le passage intermédiaire par ce score s (qui ne sera pas à valeurs dans $\{0,1\}$ - comme y - mais dans $[0,1]$ ou dans $\mathbb{R}$). On supposera enfin qu'il existe un attribut protégé p (nous considérons ici qu'il est binaire, avec $p \in \{0,1\}$).

4.1 Complément sur les classifieurs

Comme nous l'avons évoqué, la classification binaire (lorsque $y \in \{0,1\}$) est un peu particulière de part la construction intermédiaire d'un score s, avant de construire le classifieur $\widehat{y}$ *stricto sensu* ($\widehat{y} \in \{0,1\}$). Par analogie avec les régressions logistique ou probit, dans toute la suite, le score s sera une fonction $\mathcal{X} \to [0,1]$, où $\mathcal{X} \subset \mathbb{R}^k$, et avec, par exemple [49]

$$s(x) = \frac{\exp\left[\boldsymbol{x}^\top \boldsymbol{\beta}\right]}{1+\exp\left[\boldsymbol{x}^\top \boldsymbol{\beta}\right]} \text{ ou } s(\boldsymbol{x}) = \Phi(\boldsymbol{x}^\top \boldsymbol{\beta}),$$

respectivement pour le modèle logistique et le modèle probit. La Figure 4.1 reprend schématiquement l'analyse d'un classifieur (linéaire), avec la matrice de confusion, qui va servir de base pour

48. Évoqué dans la section 2.3.8, voir aussi https://www.fico.com/en/products/fico-score

49. Certains ouvrages définissent le score comme $\boldsymbol{x}^\top \boldsymbol{\beta}$, qui est à valeurs dans $\mathbb{R}$. Le score que l'on définit est une fonction croissante de cette combinaison linéaire.

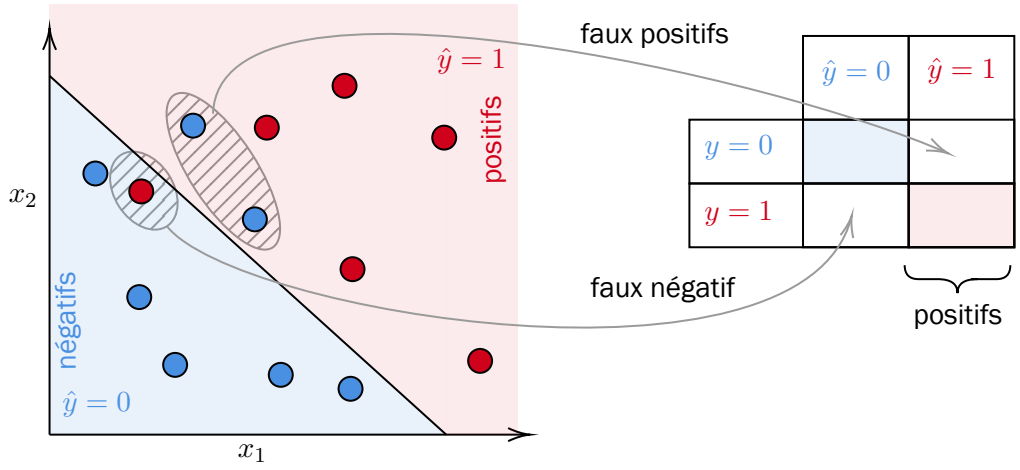


Figure 4.1 – Construction de la matrice de confusion pour un classifieur, $\widehat{y} = \mathbf{1}(x_1 + x_2 > t)$, où les points bleus • représentent des points $y = 0$ et les points rouges • représentent des points $y = 1$. La zone bleue correspond aux prédictions $\widehat{y} = 0$ et la zone rouge aux prédictions $\widehat{y} = 1$. Les points rouges en zone bleue, et les points bleus en zone rouge sont des mauvais classifications, correspondant à des erreurs, respectivement des " faux négatifs" et des "faux positifs".

construire la courbe ROC,

$$(\mathbb{P}[S > t|Y = 0], \mathbb{P}[S > t|Y = 1])_{t\in[0,1]},$$

où $S = s(\boldsymbol{X})$, soit, en notant $\widehat{Y} = \mathbf{1}(S > t)$ pour un seuil t,

$$(\mathbb{P}[\widehat{Y} = 1|Y = 0], \mathbb{P}[\widehat{Y} = 1|Y = 1]) = (\mathsf{FPR}, \mathsf{TPR}),$$

qui est la courbe des taux de vrais positifs (TPR, "*true positive rate*") en fonction des taux de faux positifs (FPR, "*false positive rate*") quand le seuil t varie. Dans l'exemple de la Figure 4.1, le taux de faux positifs (FPR) est de 2 sur 7 (sur les 7 points bleus •, 2 points sont mal classés, et annoncés positifs), soit 28.57 %; le taux de vrais positifs (TPR) est de 5 sur 6 (sur les 6 points rouges •, 1 point est mal classé et annoncé négatif), soit 83.33 %.

Sur la Figure 4.3, on peut visualiser un score de crédit (conformément à l'approche en économétrie, en statistique et en apprentissage machine, un score s faible désigne un bon risque, et donc moins de chance d'avoir un défaut). On peut visualiser à gauche les distributions du score, conditionnellement à y, avec respectivement en pointillés la densité de S quand $y = 1$ et en trait plein, celle de S quand $y = 0$. Supposons que le seuil permettant de changer de classe soit à 60 %, de telle sorte que $\widehat{y} = 1$ si $s > 60$ %. On peut observer sur la gauche que

$$\begin{cases} \mathbb{P}[S > 60\ \%|Y = 1] = \mathbb{P}[\widehat{Y} = 1|Y = 1] \sim 66.3\ \% \text{ taux de vrai positifs} \\ \mathbb{P}[S > 60\ \%|Y = 0] = \mathbb{P}[\widehat{Y} = 1|Y = 0] \sim 9.6\ \% \text{ taux de faux positifs,} \end{cases}$$

(car $\widehat{y} = 1$ correspondant à un "positif"). La courbe ROC est la courbe obtenue en représentant les taux de vrais positifs en fonction de taux de faux positifs, en changeant le seuil. C'est donc la courbe paramétrique

$$C = \{\mathbb{P}[S > t|Y = 0], \mathbb{P}[S > t|Y = 1]\}, \text{ pour } t \in [0, 1],$$

lorsque le score S et Y évoluent dans le même sens (un score élevé indique un risque élevé). On définira l'enveloppe convexe $\mathcal{C}$, comme dans Hardt *et al.* 2016. L'enveloppe convexe est intéressante car elle permet de décrire l'ensemble des classifieurs qui peuvent être construits à partir du score s. À gauche de la Figure 4.2, on peut voir l'enveloppe convexe $\mathcal{C}$ de la courbe ROC de l'exemple de la Figure 4.7. $\mathcal{C}$ est ici un quadrilatère, les bords étant constitués de quatre segments. Par exemple le segment [AB] est obtenu en utilisant le classifieur s, mais en tirant le seuil au hasard : soit le seuil associé au point A, soit le seuil associé au point B.

Pour la partie droite de la Figure 4.2, rappelons que l'*accuracy* (notée a) associée à une courbe ROC est la proportion de bonne prédiction, $\mathbb{P}[\widehat{Y} = Y]$, soit

$$a = \frac{\mathsf{TP} + \mathsf{TN}}{\mathsf{P} + \mathsf{N}} = \frac{\mathsf{TPR} \cdot \mathsf{TPR} + (1 - \mathsf{FPR}) \cdot \mathsf{N}}{\mathsf{P} + \mathsf{N}}.$$

Les courbes d'iso-performance (ou "*iso-accuracy*") ont pour équation

$$\mathsf{TPR} = \frac{\mathsf{N}}{\mathsf{P}} \cdot \mathsf{FPR} + \frac{a \cdot [\mathsf{P} + \mathsf{N}] - \mathsf{N}}{\mathsf{P}},$$

qui est linéaire en FPR : ce sont des droites (parallèles) de pente N/P, correspondant au ratio $\mathbb{P}[Y = 0]/\mathbb{P}[Y = 1]$. La courbe ayant l'*accuracy* la plus élevée sera la plus haute, et elle est "tangente" à la courbe ROC en B.

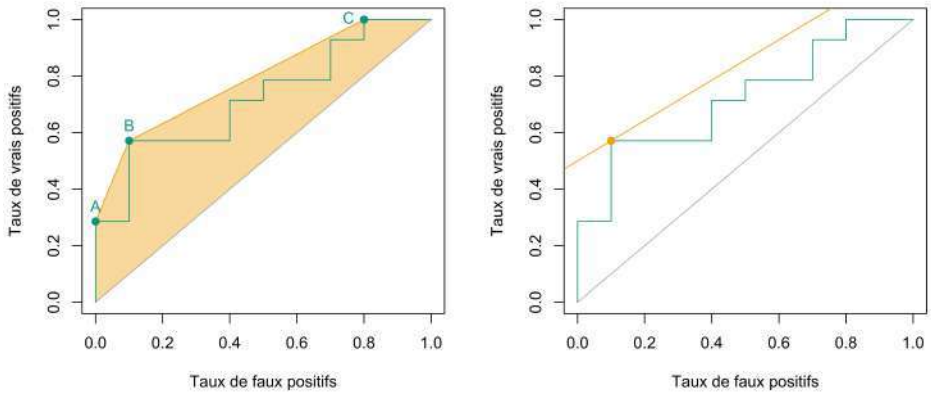


Figure 4.2 – À gauche, courbe ROC C et son enveloppe convexe $\mathcal{C}$. À droite, précision optimale à partir de ratio N/P.

On peut également construire les courbes ROC p-conditionnelles,

$$C_p(t) = \{\mathbb{P}[S > t | Y = 0, P = p], \mathbb{P}[S > t | Y = 1, P = p]\}, \text{ pour } t \in [0, 1],$$

pour les deux classes $p = 0$ (les points bleus ●) et $p = 1$ (les points rouges ●), ainsi que leur enveloppe convexe $\mathcal{C}_p$ (comme sur la Figure 4.4).

Notons qu'on peut aussi écrire $C(t) = \mathsf{TPR} \circ \mathsf{FPR}^{-1}(t)$, où $\mathsf{FRP}(t) = \mathbb{P}[S > t | Y = 0]$ et $\mathsf{TPR}(t) = \mathbb{P}[S > t | Y = 1]$. Autrement dit, la courbe ROC est la différence entre deux distributions, FPR et TPR. Pour la courbe ROC p-conditionnelles, $C_p(t) = \mathsf{TPR}_p \circ \mathsf{FPR}_p^{-1}(t)$, où $\mathsf{FRP}(t) = \mathbb{P}[S > t | Y = 0, P = p]$

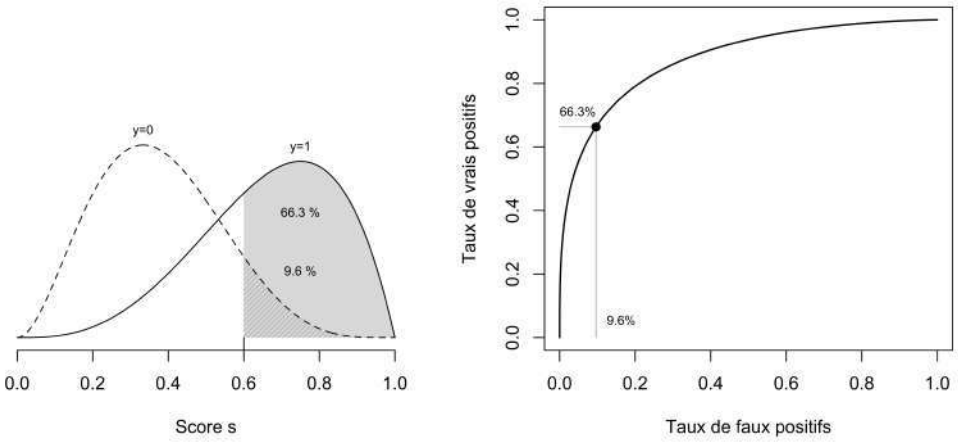


Figure 4.3 – Distributions du score S conditionnelles à $y = 1$ et à $y = 0$, à gauche, et courbe ROC, à droite, pour un score de (risque de) crédit s, $y = 1$ désignant un incident. Les aires à gauche et le point à droite sur la courbe ROC correspondent à un seuil t de 60 %.

et $\mathsf{TPR}(t) = \mathbb{P}[S > t | Y = 1, P = p]$. Notons que le AUC, l'aire sous la courbe, s'écrit alors

$$\mathsf{AUC} = \int_0^1 C(t)dt = \int_0^1 \mathsf{TPR} \circ \mathsf{FPR}^{-1}(t)dt.$$

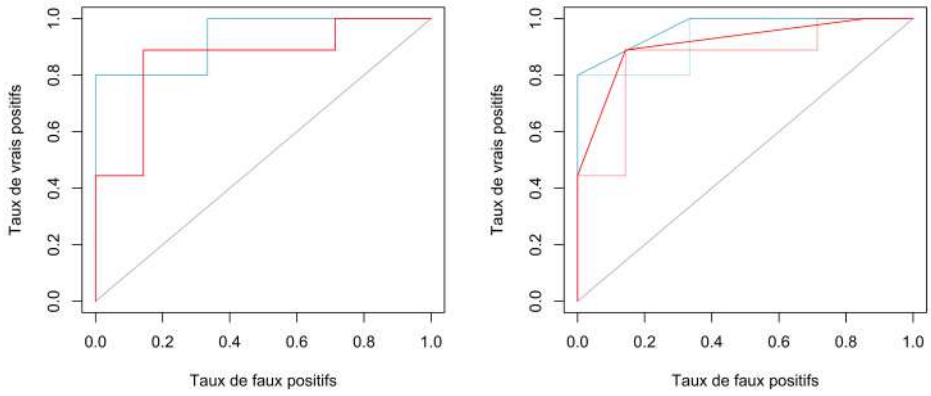


Figure 4.4 – Courbes ROC C_p à gauche, et l'enveloppe convexe à droite $\mathcal{C}_p$, pour $p = 0$ et $p = 1$.

Dans ce chapitre, lorsque l'on s'intéressera à un problème de classification, avec une variable binaire $y \in \{0, 1\}$, on conditionnera également par la variable protégée p.

Si y est à valeurs dans $\mathcal{Y}$ (correspondant è $\{0, 1\}$ pour un classifieur, $\mathbb{N}$ pour un comptage, $\mathbb{R}$ pour un problème de régression), on définit une fonction de perte comme étant une fonction définie sur $\mathcal{Y} \times \mathcal{Y}$, à valeurs réelles, telle que $\ell(y, y') \geq 0$ et $\ell(y, y) = 0$. Ainsi, le risque d'un prédicteur m est

$$\mathcal{R}(m) = \mathbb{E}[\ell(Y, m(\boldsymbol{X}))],$$

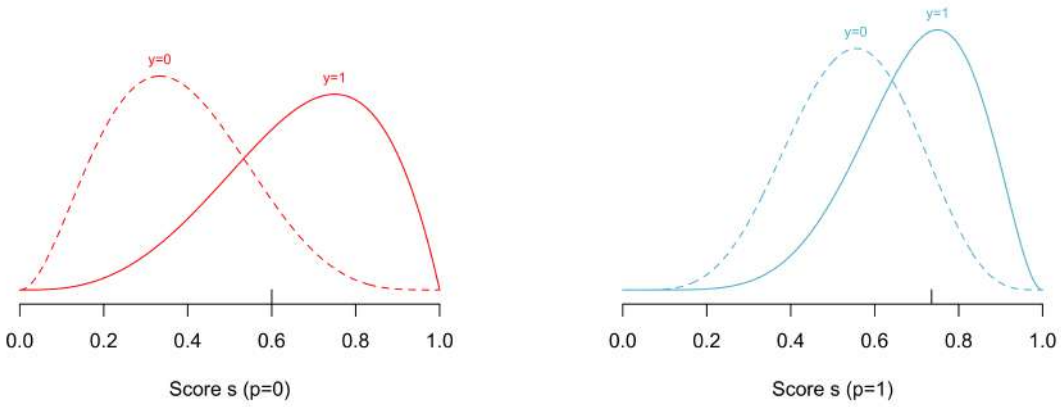


Figure 4.5 – Distributions conditionnelles de $S|Y=1,P=0$ et $S|Y=0,P=0$, à gauche (population supposée favorisée), et distributions conditionnelles de $S|Y=1,P=1$ et $S|Y=0,P=1$, à droite (population supposée défavorisée).

En régression, par exemple on travaille généralement avec une fonction de coût quadratique, $\ell(y,y') = (y-y')^2$, et on obtient alors la fonction de risque

$$\mathcal{R}(m) = \mathbb{E}[(Y - m(\boldsymbol{X}))^2],$$

appelée risque quadratique. Et comme la distribution exacte de Y n'est pas connue, il n'est donc pas possible de calculer la fonction de risque. On construit la fonction de risque empirique

$$\widehat{\mathcal{R}}_n(m) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i, m(\boldsymbol{x}_i)\right),$$

où $\{\boldsymbol{x}_i, y_i\}_{i=1,\cdots,n}$ est un échantillon aléatoire (on parlera de "*in-sample risk*"). Dans un contexte de régression, on travaille généralement avec l'erreur quadratique moyenne (ou "*mean squared error*", MSE), donnée par

$$\text{EQM}_n = \text{MSE}_n = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - m(\boldsymbol{x}_i)\right)^2.$$

Dans le contexte d'un classifieur, $\mathcal{Y} = \{0,1\}$, le classifieur de Bayes est

$$\widehat{y} = \underset{y\in\mathcal{Y}}{\text{argmax}}\{\mathbb{P}(Y=y|\boldsymbol{X}=\boldsymbol{x})\} \text{ noté } m^\star(\boldsymbol{x}).$$

Notons que

$$\mathbb{P}(\widehat{Y} \neq Y) = \mathbb{E}[\mathbb{P}(\widehat{Y} \neq Y|\boldsymbol{X})] = 1 - \mathbb{E}[\max_{y\in\mathcal{Y}} \mathbb{P}(Y=y|\boldsymbol{X})].$$

Le risque de (mauvaise) classification (ou taux d'erreur) d'un classifieur [50] m est

$$\mathcal{R}(m) = \mathbb{P}[m(\boldsymbol{X}) \neq Y],$$

50. Avec les notations précédentes, $m(\boldsymbol{x}) = \mathbf{1}(s(\boldsymbol{x}) > t)$.

et le risque empirique associé de m s'écrit

$$\widehat{\mathcal{R}}(m) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(m(\boldsymbol{x}_i) \neq y_i),$$

qui est minimisé pour le classifieur de Bayes $m^\star$,

$$m^\star(\boldsymbol{x}) = \mathbf{1}(\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}] \geq \frac{1}{2}) = \mathbf{1}(\frac{\mathbb{P}[\boldsymbol{X}=\boldsymbol{x}|Y=1]}{\mathbb{P}[\boldsymbol{X}=\boldsymbol{x}|Y=0]} > \frac{\mathbb{P}[Y=0]}{\mathbb{P}[Y=1]}),$$

(on retrouvera le ratio N/P à droite).

4.2 Le critère d'aveuglement

La notion de discrimination étant étroitement liée à une indépendance entre la variable protégée p et la variable d'intérêt y (ou le score s). Rappelons que l'indépendance entre deux variables X et Y s'écrit

$$Y \perp\!\!\!\perp X \iff \begin{cases} \mathbb{P}(Y=y|X=x) = \mathbb{P}(Y=y)\ \forall x,\ \forall y,\ \text{ si } y \in \{0,1\} \\ \mathbb{P}(Y\leq y|X=x) = \mathbb{P}(Y\leq y)\ \forall x,\ \forall y,\ \text{ si } y \in \mathbb{R}, \end{cases}$$

et l'indépendance entre deux variables X et Y conditionnellement à une troisième Z s'écrit

$$Y \perp\!\!\!\perp X \mid Z \iff \begin{cases} \mathbb{P}(Y=y|X=x,Z=z) = \mathbb{P}(Y=y|Z=z)\ \forall x,z,\ \forall y,\ \text{ si } y \in \{0,1\} \\ \mathbb{P}(Y\leq y|X=x,Z=z) = \mathbb{P}(Y\leq y|Z=z)\ \forall x,z,\ \forall y,\ \text{ si } y \in \mathbb{R} \end{cases}$$

Une conséquence (ça ne serait équivalent que dans le cas d'une variable y binaire), sera respectivement

$$\mathbb{E}[Y|X=x] = \mathbb{E}[Y],\ \text{ou } \mathbb{E}[Y|X=x,Z=z] = \mathbb{E}[Y|Z=z],\ \forall x,z.$$

Dans la littérature sur les discriminations, on note soit p (pour variable *p*rotégée), s (pour variable sensible), voire a (pour *a*ttribut protégé). Dans la littérature sur l'inférence causale, on utilisera t (pour *t*raitement). Nous utiliserons ici la notation p pour la variable protégée, s sera utilisée pour le score, permettant de classer. Nous ferons référence à $p=0$ comme la population (supposée) favorisée, et $p=1$ comme la population défavorisée.

L'approche la plus populaire pour avoir un classifieur équitable consiste à interdire l'utilisation d'une variable protégée dans un modèle prédictif. Cette approche est appelée "équité par ignorance".

Équité par ignorance - Fairness Through Unawareness (Kusner *et al.* 2017)

On parlera d'équité par ignorance si l'attribut protégé p n'est pas explicitement utilisé dans la fonction de décision $\widehat{y}$, c'est-à-dire si l'attribut protégé p n'est utilisé ni dans la construction du score s, ni dans le choix du niveau du seuil, permettant de passer de s à $\widehat{y}$.

On suppose ici que le choix du seuil ne dépend pas du critère protégé p, comme sur la Figure 4.6. Dans ce cas, pour la population favorisée (courbe rouge à gauche, $p=0$), davantage de personnes

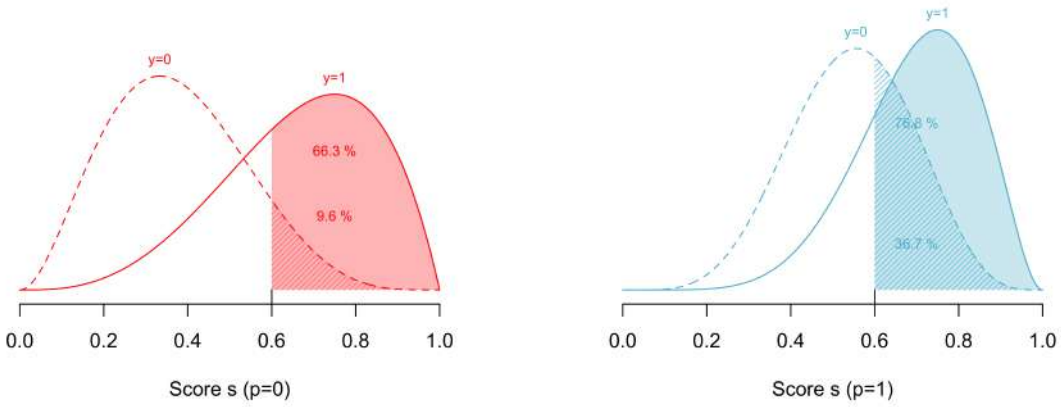


Figure 4.6 – Distributions de s conditionnelle à $y = 1, p = 0$ et s conditionnelle à $y = 0, p = 0$ à gauche (population supposée favorisée), et distributions de s conditionnelle à $y = 1, p = 1$ et s conditionnelle à $y = 0, p = 1$, à droite (population supposée défavorisée). Le seuil permettant de passer de $\widehat{y} = 0$ à $\widehat{y} = 1$ est fixé à 60 %.

dans cette sous-population ont un score (et donc un risque) plus faible que pour la population défavorisée (courbe bleue à droite, $p = 1$).

Supprimer une variable protégée de la base d'apprentissage pourrait sembler être un pas en avant. Cependant, certaines des variables prédictives, non-protégées, peuvent en fait être (très) corrélées avec la variable protégée, ce qui permet à la discrimination de se poursuivre. On pourrait poursuivre, et éliminer également toutes les variables fortement corrélées. Mais cela a un prix, car chaque suppression d'une variable supprime également des informations précieuses pour la tâche de prédiction.

Gajane et Pechenizkiy 2017, Žliobaitė 2017 Verma et Rubin 2018, ou Friedler, Scheidegger, Venkatasubramanian *et al.* 2019 ont recensé plusieurs concepts d'équité algorithmique. La plupart des définitions de l'équité sont basées sur l'équité de groupe qui traite de l'équité statistique dans l'ensemble de la population. En complément, l'équité individuelle stipule que des individus similaires doivent être traités de manière similaire, indépendamment de leur appartenance à un groupe. Dans cette section, nous nous concentrerons principalement sur l'équité de groupe, dont les trois définitions sont les suivantes : (i) la parité démographique, (ii) l'égalité des chances et (iii) l'égalité des opportunités. Nous allons maintenant les examiner l'une après l'autre, avant de proposer quelques extensions, inspirées de Žliobaitė 2017 (qui mentionne une vingtaine de mesures) et Gajane et Pechenizkiy 2017 (qui considèrent sept grandes approches).

4.3 Équité au niveau du groupe

Pour illustrer, considérons un jeu de données fictif, comme celui de Kearns et Roth 2019 (dans leur exemple, il existe deux types d'individus, les cercles et les carrés, comme dans Ruillier 2004),

comme sur la Figure 4.7. Dans notre cas, on considère des individus bleus • et des individus rouges •, ce qui sera notre variable protégée p. On dispose aussi d'une variable d'intérêt y, binaire ($y \in \{0, 1\}$), ainsi que d'un ensemble de variable $\boldsymbol{x} \in \mathcal{X}$ qui ont permis de construire un score $s(\boldsymbol{x}) \in [0, 1]$. Sur la Figure 4.7, au niveau $s(\boldsymbol{x}_i) \in [0, 1]$, on observe un point bleu • ou rouge • suivant la valeur de p_i, ainsi que $y_i \in \{0, 1\}$.

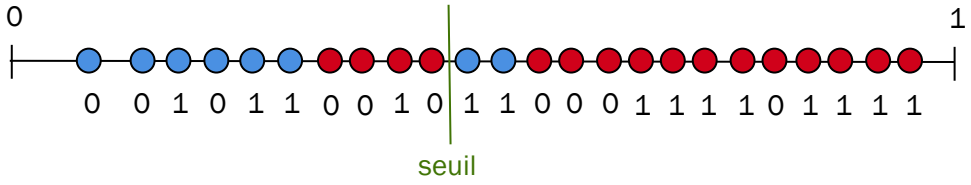


Figure 4.7 – Un score $s(\boldsymbol{x}) \in [0, 1]$, une variable protégée $p \in \{•, •\}$, et une variable d'intérêt $y \in \{0, 1\}$ (inspiré de Kearns et Roth 2019).

4.3.1 Parité démographique

Comme le soulignent Caton et Haas 2020, il existe plusieurs manières de définir (formellement) l'équité d'un classifieur, ou d'un modèle. Par exemple, on peut souhaiter l'indépendance entre le score et l'appartenance à un groupe, $S = s(\boldsymbol{X}) \perp\!\!\!\perp P$, ou entre la prédiction et la variable protégé $\widehat{Y} \perp\!\!\!\perp P$.

Équité démographique - Demographic Parity (Corbett-Davies *et al.* 2017, Agarwal 2021)

Une fonction de décision $\widehat{y}$ satisfait à la parité démographique si $\widehat{Y} \perp\!\!\!\perp P$, soit

$$\mathbb{P}[\widehat{Y} = y | P = 0] = \mathbb{P}[\widehat{Y} = y | P = 1], \ \forall y \in \{0, 1\},$$

ou

$$\mathbb{P}[\widehat{Y} \leq y | P = 0] = \mathbb{P}[\widehat{Y} \leq y | P = 1], \ \forall y \in \mathbb{R}.$$

Une implication sera

$$\mathbb{E}[\widehat{Y} | P = 0] = \mathbb{E}[\widehat{Y} | P = 1].$$

La dernière caractérisation est équivalente aux deux autres si y et $\widehat{y}$ prennent deux valeurs. Dans le cas où y est continue, la seconde propriété correspond à une notion d'équité démographique "forte" alors que la dernière correspond à une notion d'équité démographique "faible" (la seconde impliquant la troisième, mais pas l'inverse).

Cette équité démographique, aussi appelée "parité statistique", impose simplement que la fraction des demandeurs bleus qui se voient accorder un crédit soit approximativement la même que la fraction des demandeurs rouge qui se voient accorder un crédit. Par symétrie, les proportions de rejet doivent être identique. En utilisant un même seuil sur le score, pour accorder un crédit, comme sur la Figure 4.7, on voit que la parité statistique n'est pas atteinte :

$$\mathbb{P}[\widehat{Y} = 1 | P = •] = \mathbb{P}[S > \text{seuil} | P = •] = \frac{2}{8} = 26\ \% \text{ et } \mathbb{P}[S > \text{seuil} | P = •] = \frac{12}{16} = 76\ \%,$$

de telle sorte que

$$\mathbb{P}[\widehat{Y} = 1 | P = •] \neq \mathbb{P}[\widehat{Y} = 1 | P = •].$$

La parité statistique est certainement une forme d'équité, mais elle est généralement faible et imparfaite. Et comme le montre la partie gauche de la Figure 4.8, elle n'a rien à voir avec la qualité du modèle prédictif.

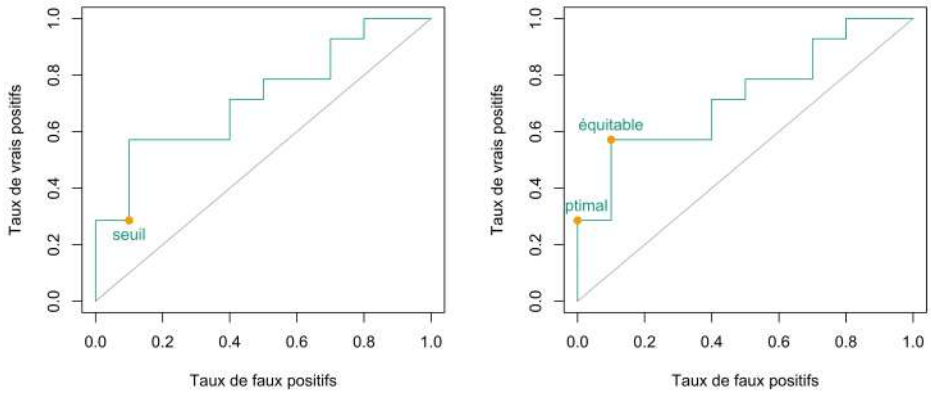


Figure 4.8 – Courbe ROC sur les données de l'exemple de la Figure 4.7, avec à droite deux niveaux pour le seuil, respectivement le seuil "équitable" (Figure 4.9) et "optimal" (Figure 4.11).

Supposons que

$$\mathbb{P}[Y = 1|P = \bullet] = \frac{1}{4} = 26\,\% \text{ et } \mathbb{P}[Y = 1|P = \bullet] = \frac{3}{4} = 76\,\%,$$

et que la loi de Y dépende *seulement* de P. Dans ce cas, imposer la parité statistique revient à choisir un mauvais modèle, car le modèle parfait donnerait

$$\mathbb{P}[\widehat{Y} = 1|P = \bullet] = \frac{1}{4} = 26\,\% \text{ et } \mathbb{P}[\widehat{Y} = 1|P = \bullet] = \frac{3}{4} = 76\,\%.$$

Sur la Figure 4.9, on peut voir le seuil "optimal", au sens du pouvoir prédictif maximal, minimisant le taux d'erreur commise, visible sur la Figure 4.10, à gauche.

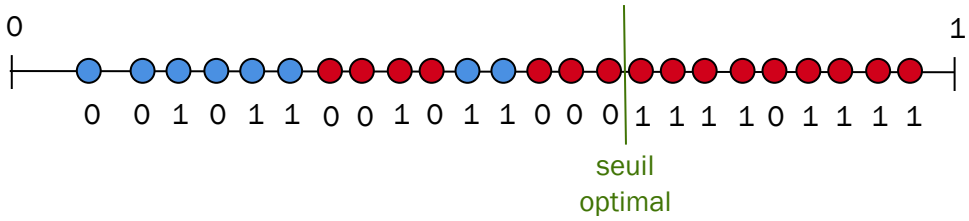


Figure 4.9 – $p \in \{\bullet, \bullet\}$, $y \in \{0, 1\}$, via Kearns et Roth 2019.

Sur la Figure 4.11, on peut voir le seuil "équitable" (ou au moins "plus équitable"), au sens où $s \mapsto \mathbb{P}[S > s|P = \bullet]/\mathbb{P}[S > s|P = \bullet]$ est aussi grand que possible (autrement dit $\mathbb{P}[\widehat{Y} = 1|P = \bullet]/\mathbb{P}[\widehat{Y} = 1|P = \bullet]$ est aussi grand que possible).

Pour résumer, minimiser le taux d'erreur (et donc augmenter la précision) et maximiser l'équité,

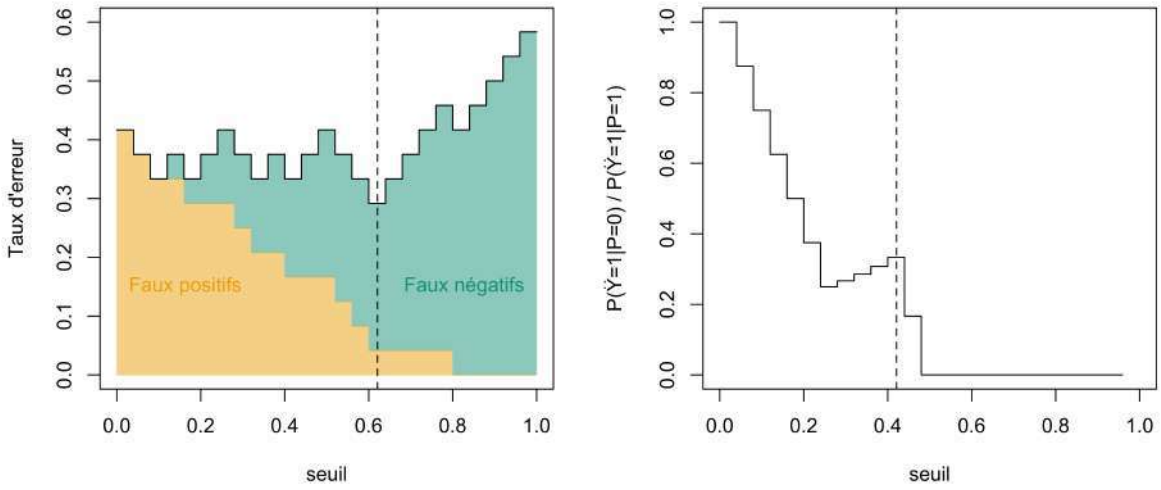


Figure 4.10 – Taux de faux positifs et de faux négatifs, à gauche, et évolution du ratio $\mathbb{P}[\widehat{Y}=1|P=0]/\mathbb{P}[\widehat{Y}=1|P=1]$, en fonction du niveau du seuil utilisé.

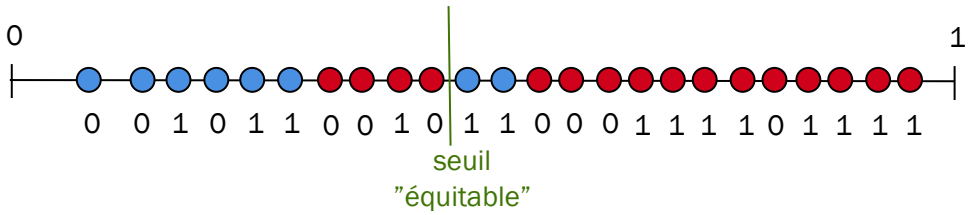


Figure 4.11 – Choix d'un seuil d'équité "équitable", via Kearns et Roth 2019.

sont souvent deux objectifs inconciliables.

$$\begin{cases} \text{seuil optimal} & : \mathbb{P}[\widehat{Y} \neq Y] = \dfrac{6+1}{24} = 29.17\,\% \text{ et } \dfrac{\mathbb{P}[\widehat{Y}=1|P=\bullet]}{\mathbb{P}[\widehat{Y}=1|P=\bullet]} = \dfrac{0}{24} = 0\,\% \\ \text{seuil "équitable"} & : \mathbb{P}[\widehat{Y} \neq Y] = \dfrac{4+4}{24} = 33.33\,\% \text{ et } \dfrac{\mathbb{P}[\widehat{Y}=1|P=\bullet]}{\mathbb{P}[\widehat{Y}=1|P=\bullet]} = \dfrac{2 \cdot 16}{12 \cdot 4} = 33.33\,\% \end{cases}$$

Nous reviendrons par la suite sur le compromis qui existera bien souvent entre l'équité des modèles, et leur précision (ou leur pouvoir prédictif), et la frontière d'efficience associée.

Un autre défaut de cette approche est que l'indépendance souhaitée, entre la variable protégée p et la prédiction $\widehat{y}$, ne tient pas compte du fait que le résultat y peut être corrélé avec la variable sensible p. Autrement dit, si les groupes induits par p ont des distributions sous-jacentes différentes pour y, le fait de ne pas tenir compte de ces dépendances peut conduire à des résultats qui seraient considérés comme équitables, mais pas pour les groupes eux-mêmes. Assez naturellement, une extension de la propriété d'indépendance est le critère de "séparation" qui demande l'indépendance entre la prédiction $\widehat{y}$ et la variable sensible P, conditionnellement à la valeur de la variable cible y, soit $\widehat{Y} \perp\!\!\!\perp P$ conditionnement à Y.

Cette approche peut revenir à choisir des seuils différents, avec un seuil plus bas pour les individus bleus • que pour les individus rouges •, comme sur la Figure 4.12. En l'occurrence,

$$\mathbb{P}[\widehat{Y} = 1 | P = \bullet] = \mathbb{P}[S > \text{seuil}_{\bullet} | P = \bullet] = 50\,\% = \mathbb{P}[\widehat{Y} = 1 | P = \bullet] = \mathbb{P}[S > \text{seuil}_{\bullet} | P = \bullet].$$

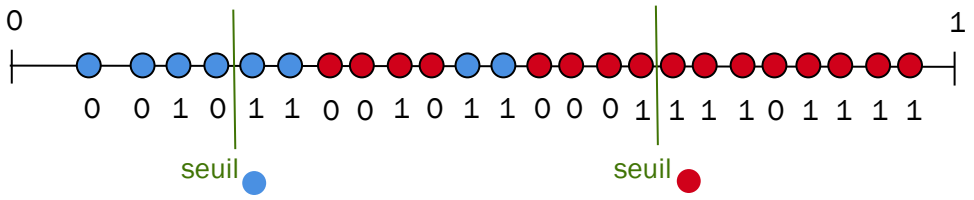


Figure 4.12 – $p \in \{\bullet, \bullet\}$, $y \in \{0, 1\}$, via Kearns et Roth 2019.

Sur la Figure 4.13, on peut visualiser le taux de faux positifs dans chacune des classes, en fonction du seuil, à gauche, et le taux de vrais positifs à droite. Ce dernier cas est appelé "égalité des opportunités" par Hardt *et al.* 2016.

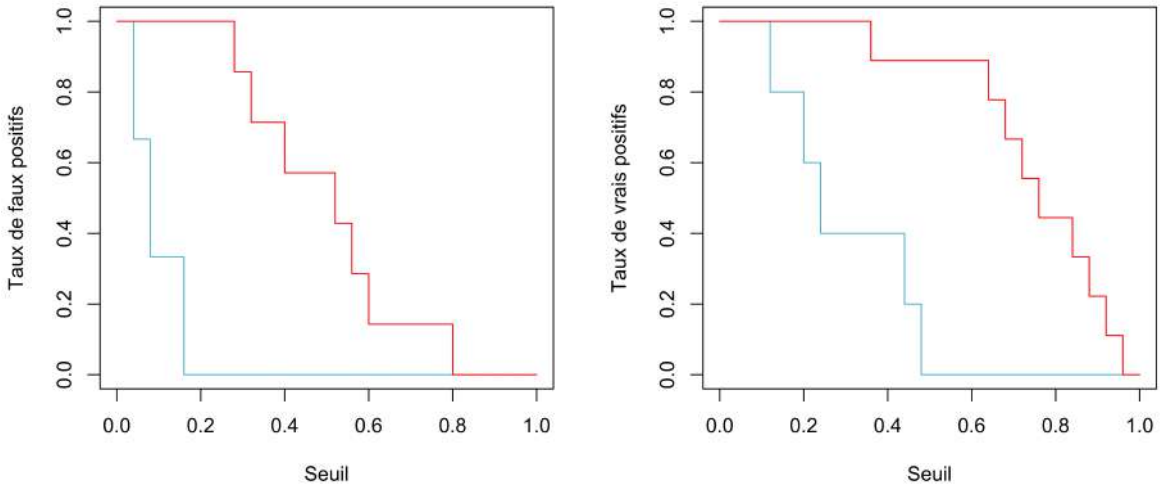


Figure 4.13 – Taux de faux positifs, à gauche, et taux de vrais positifs, à droite.

La parité démographique (ou parité statistique) suggère qu'un prédicteur est sans biais si la prédiction $\widehat{y}$ est indépendante de l'attribut protégé p,

$$\widehat{Y} \perp\!\!\!\perp P$$

de sorte que

$$\begin{cases} \mathbb{P}(\widehat{Y} = 1 | P = 0) = \mathbb{P}(\widehat{Y} = 1 | P = 1) = \mathbb{P}(\widehat{Y} = 1), \text{ si } y \in \{0, 1\} \\ \mathbb{P}(\widehat{Y} \leq \gamma | P = 0) = \mathbb{P}(\widehat{Y} \leq \gamma | P = 1) = \mathbb{P}(\widehat{Y} \leq \gamma), \ \forall \gamma, \text{ si } y \in \mathbb{R} \end{cases}$$

On parlera aussi de "*group fairness*". Notons que dans le premier cas, la condition est équivalente à avoir

$$\mathbb{E}(\widehat{Y} | P = 0) = \mathbb{E}(\widehat{Y} | P = 1) = \mathbb{E}(\widehat{Y})$$

mais pas dans le second, quand y est continue. Cette dernière sera vraie, mais elle ne sera pas suffisante. Ici, la même proportion de chaque population est classée comme positive. Cependant, cela peut entraîner des taux de faux positifs et de vrais positifs différents si le vrai résultat y varie effectivement avec l'attribut protégé p.

Une alternative à l'hypothèse d'indépendance $\widehat{Y} \perp\!\!\!\perp P$ est de demander à ce que $\widehat{Y}$ et P aient une une information mutuelle nulle,

$$IM(\widehat{Y}, P) = \mathcal{E}(\widehat{Y}) + \mathcal{E}(P) - \mathcal{E}(\widehat{Y}, P) = 0,$$

où $\mathcal{E}$ désigne l'entropie, soit

$$IM(\widehat{Y}, P) = \sum_{\widehat{y},p} \mathbb{P}(\widehat{y}, p) \log \frac{\mathbb{P}(\widehat{y}, p)}{\mathbb{P}(\widehat{y})},$$

Sur notre exemple illustratif, on a une situation équitable (au sens démographique) si on a choisi le seuil de manière à ce que la même proportion de chaque groupe soit classée comme $\widehat{y} = 0$ (ou $\widehat{y} = 1$), et reçoive un prêt, comme sur la Figure 4.14. Par exemple, si on veut maintenir un seuil à 60 % pour la population favorisée (courbe rouge à gauche, $p = 0$), on doit légèrement baisser le seuil pour la population défavorisée (courbe bleue à droite, $p = 1$), avec ici un seuil légèrement inférieur à 50 %. Autrement dit, avec un tel choix $\mathbb{P}(\widehat{Y} = 1|P = 1) = \mathbb{P}(\widehat{Y} = 1|P = 0)$.

Cette méthode présente un inconvénient : les taux de vrais et de faux positifs peuvent être complètement différents dans les deux sous-populations.

4.3.2 Égalité des opportunités et des chances

L'égalité des chances et l'égalité des opportunités ne représentent pas tant une mesure d'équité qu'une définition potentielle de l'équité. L'égalité des chances est atteinte lorsque la variable cible prédite d'un modèle $\widehat{y}$ et le label d'une catégorie protégée p sont statistiquement indépendants l'un de l'autre, conditionnellement à la valeur réelle de la variable cible y. Dans une tâche de classification binaire, cela peut être simplifié en exigeant que les taux de vrais positifs et les taux de faux positifs soient égaux entre les groupes, où les groupes sont déterminés par la catégorie protégée. Un critère d'équité légèrement moins exigeant est l'égalité des chances, dans lequel seule la probabilité du vrai positif est égalisée entre les différents groupes d'une catégorie protégée.

Formellement, on a les définitions suivantes, où on demande la parité des faux ou des vrais positifs (Figures 4.14 et 4.15 respectivement).

Égalité des opportunités, Equal Opportunity (Hardt *et al.* 2016)

On parlera d'égalité des opportunités, ou parité des vrais positifs, si

$$\mathbb{P}[\widehat{Y} = 1|P = 0, Y = 1] = \mathbb{P}[\widehat{Y} = 1|P = 1, Y = 1].$$

Égalité des faux positifs, Equalized Opportunity (Hardt *et al.* 2016)

On parlera d'égalité des faux positifs si

$$\mathbb{P}[\widehat{Y} = 1|P = 0, Y = 0] = \mathbb{P}[\widehat{Y} = 1|P = 1, Y = 0].$$

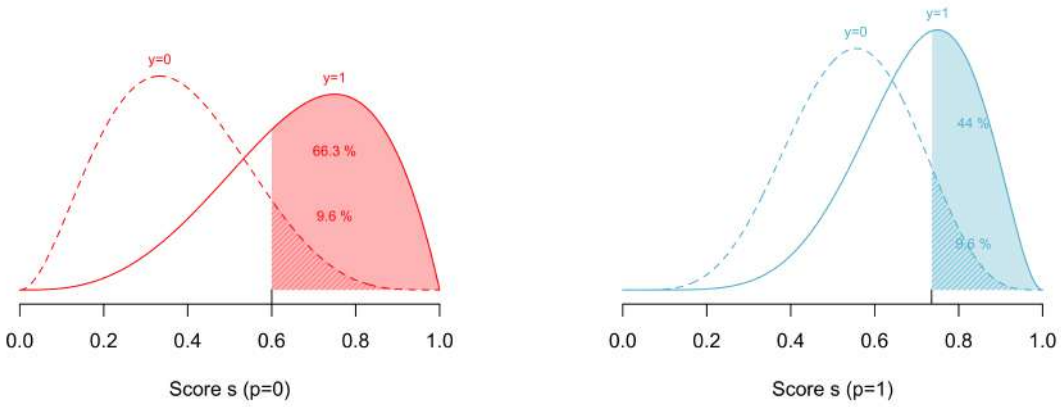


Figure 4.14 – Distributions de s conditionnelle à $y = 1, p = 0$ et de s conditionnelle à $y = 0, p = 0$, à gauche (population supposée favorisée), et distributions de s conditionnelle à $y = 1, p = 1$ et conditionnelle à $y = 0, p = 1$, à droite (population supposée défavorisée), avec $\mathbb{P}(\widehat{Y} = 1|P = 1, Y = 0) = \mathbb{P}(\widehat{Y} = 1|P = 0, Y = 0)$ (ici de l'ordre de 9.6 % de faux positifs).

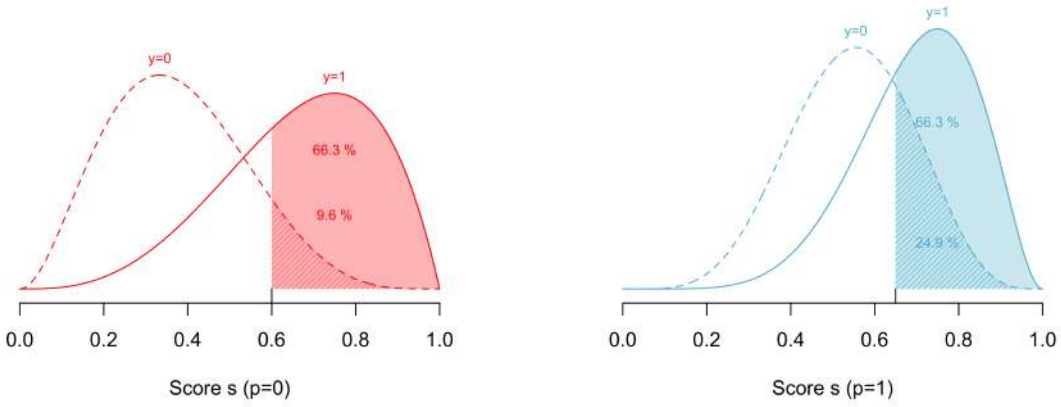


Figure 4.15 – Distributions de s conditionnelle à $y = 1, p = 0$ et à $y = 0, p = 0$, à gauche (population supposée favorisée), et distributions de s conditionnelle à $y = 1, p = 1$ et à $y = 0, p = 1$, à droite (population supposée défavorisée), avec $\mathbb{P}(\widehat{Y} = 1|P = 1, Y = 1) = \mathbb{P}(\widehat{Y} = 1|P = 0, Y = 1)$ (ici de l'ordre de 66.3 % de vrais positifs).

Égalité des chances, Equalized Odds (Hardt *et al.* 2016)

La parité des faux positifs et des vrais positifs est appelée égalité des chances,

$$\begin{cases}\mathbb{P}[\widehat{Y}=1|P=0,Y=1]=\mathbb{P}[\widehat{Y}=1|P=1,Y=1]\\ \mathbb{P}[\widehat{Y}=1|P=0,Y=0]=\mathbb{P}[\widehat{Y}=1|P=1,Y=0]\end{cases}$$

ou

$$\mathbb{P}[\widehat{Y}=1|P=0,Y=y]=\mathbb{P}[\widehat{Y}=1|P=1,Y=y],\ \forall y\in\{0,1\}$$

autrement dit, $\widehat{Y}\perp\!\!\!\perp P$ conditionnellement à Y.

Équité d'AUC (Borkan *et al.* 2019)

On aura une équité d'AUC si $\text{AUC}_1=\text{AUC}_0$, où AUC_p est l'AUC pour le groupe p.

On retrouve une idée proche chez Beutel, Chen, Doshi *et al.* 2019. Le soucis de l'AUC est qu'on peut avoir des AUC identiques, mais des courbes ROC sous-jacentes très différentes. Aussi, il peut être intéressant de considérer une notion d'équité basée sur les courbes ROC. Pour rappel, nous avions défini la courbe ROC comme $C(s)=\text{TPR}\circ\text{FPR}^{-1}(s)$, où $\text{FRP}(t)=\mathbb{P}[S>t|Y=0]$ et $\text{TPR}(t)=\mathbb{P}[S>t|Y=1]$.

Équité de courbes ROC (Vogel *et al.* 2021)

Soient $\text{FRP}_p(t)=\mathbb{P}[S>t|Y=0,P=p]$ et $\text{TPR}_p(t)=\mathbb{P}[S>t|Y=1,P=p]$. Posons $\Delta_{TPR}(s)=\text{TPR}_1\circ\text{TPR}_0^{-1}(s)-s$ et $\Delta_{FRP}(s)=\text{FPR}_1\circ\text{FPR}_0^{-1}(s)-s$. On aura une équité de courbes ROC si $\|\Delta_{TPR}\|_\infty=\|\Delta_{FPR}\|_\infty=0$.

Une hypothèse (implicite) faite ici est que la classe 0 (ou • dans notre illustration) dans l'attribut protégé P représente un groupe socialement protégé, c'est-à-dire un groupe minoritaire qui fait l'objet de discriminations), de telle sorte que l'impact disparate est défini par des résultats positifs (c'est-à-dire des résultats souhaitables).

L'égalité des chances est satisfaite si la prédiction $\widehat{Y}$ est conditionnellement indépendante de l'attribut protégé P, étant donné la valeur réelle Y,

$$\forall y:\ \widehat{Y}\perp\!\!\!\perp P\mid Y=y,$$

soit

$$\begin{cases}\text{classifieur}: \mathbb{P}(\widehat{Y}=1|P=0,Y=y)=\mathbb{P}(\widehat{Y}=1|P=1,Y=y)=\mathbb{P}(\widehat{Y}=1|Y=y),\ \forall y\in\{0,1\}\\ \text{régression}: \mathbb{P}(\widehat{Y}\leq\gamma|p=0,Y=y)=\mathbb{P}(\widehat{Y}\leq\gamma|p=1,Y=y)=\mathbb{P}(\widehat{y}\leq\gamma|y),\ \forall\gamma\ \forall y\in\mathbb{R}\end{cases}$$

Là encore, dans le premier cas, la condition est équivalente à avoir

$$\mathbb{E}(\widehat{y}|p=0,Y=y)=\mathbb{E}(\widehat{y}|p=1,Y=y)=\mathbb{E}(\widehat{y}|Y=y),\ \forall y\in\{0,1\},$$

mais pas dans le second, quand y est continue. Cette dernière sera vraie, mais elle ne sera pas suffisante. Cela signifie que le taux de vrais positifs et le taux de faux positifs seront les mêmes pour chaque population ; chaque type d'erreur est apparié entre chaque groupe.

Dans notre illustration, l'égalité des chances est impossible à atteindre. En effet, cette définition de l'équité propose que les taux de faux positifs et de vrais positifs soient les mêmes pour les deux populations. Cela peut sembler raisonnable, mais, dans l'exemple illustratif, c'est impossible car les deux courbes ROC ne se croisent pas. Notons que si les courbes se croisaient, cela pourrait imposer des choix de seuils qui ne seraient pas intéressants, en pratique (avec des taux d'acceptation potentiellement beaucoup trop faibles, ou trop élevés).

L'égalité des opportunités, définie par Hardt *et al.* 2016, a la même formulation mathématique que l'égalité des chances, pour un classifieur, mais elle se concentre sur une étiquette particulière,

$$\exists y : \widehat{Y} \perp\!\!\!\perp P \mid Y = y.$$

Classiquement, on se concentrera sur le label 1 de la vraie valeur y, pour définir l'égalité des opportunités, de sorte que

$$\mathbb{P}(\widehat{Y} = 0|Y = 1, P = p) = \mathbb{P}(\widehat{Y} = 0|Y = 1),\ \forall p \in \{0, 1\},$$

ce qui revient à comparer les taux de faux négatifs. Dans ce cas, nous voulons que le taux de vrais positifs $\mathbb{P}(\widehat{Y} = 1|Y = 1)$ soit le même pour chaque population sans tenir compte des erreurs lorsque $y = 0$. En effet, cela signifie que la même proportion de chaque population reçoit le "bon" résultat $y = 1$.

La déviation de l'égalité des chances est mesurée par la différence d'égalité des chances :

$$EOD = \mathbb{P}(\widehat{Y} = 1|Y = 1, P = 1) - \mathbb{P}(\widehat{Y} = 1|Y = 1, P = 0).$$

Comme on a ici une variable y binaire, la condition sur la probabilité de $\widehat{y}$ sera ici équivalente à la condition sur l'espérance

$$\mathbb{E}(\widehat{Y}|Y = 1, P = p) = \mathbb{E}(\widehat{Y}|Y = 1),\ \forall p \in \{0, 1\}.$$

Dans notre illustration, l'égalité des opprtunités revient à trouver des niveaux équivalents de taux de vrais positifs sur les courbes ROC.

Comme on peut le visualiser à droite de la Figure 4.16, les seuils sont choisis de façon à ce que le taux de vrais positifs soit le même pour les deux populations. Autrement dit, on doit avoir la même proportion qui se voit proposer un crédit, dans chacun des groupes (favorisé et défavorisé). Par exemple ici, on garde le seuil de 60 % sur le score pour la population favorisée (correspond à un taux de vrais positifs de l'ordre de 63 %), et on doit utiliser un seuil de l'ordre de 70 % sur le score pour la population défavorisée (correspond là aussi à un taux de vrais positifs de l'ordre de 63 %).

4.3.3 Parité démographique conditionnelle

Toute la discussion que nous venons d'avoir peut être étendue en conditionnant également par les variables explicatives (ou un sous-ensemble d'entre elles).

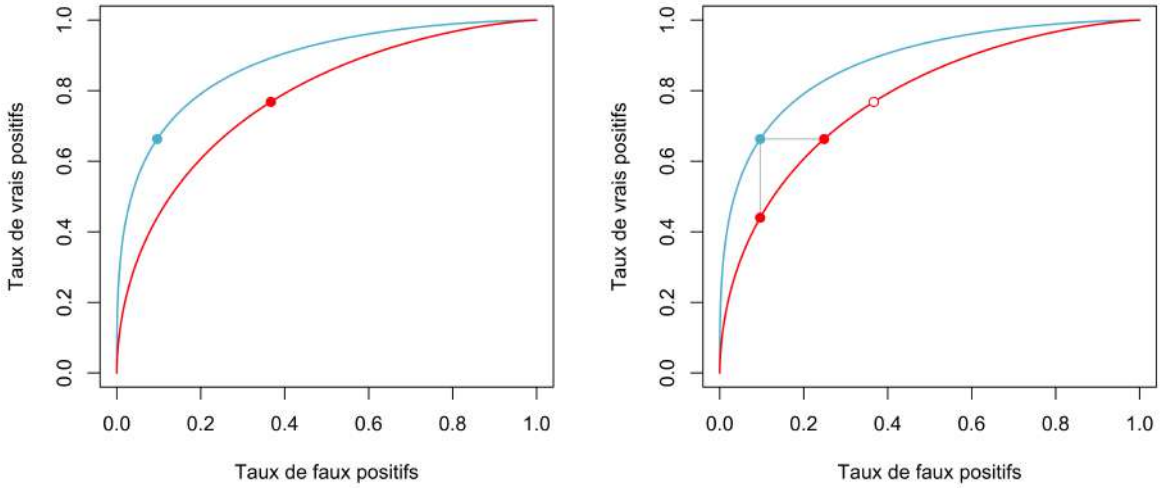


Figure 4.16 – Courbes ROC pour les deux sous-populations, celle supposée favorisée (ou $p = 1$) et celle supposée défavorisée (ou $p = 0$). À gauche, les deux points correspondent à un seuil de 60 %, identique pour les deux populations (stratégie d'aveuglement au critère protégé). À droite, le cas d'égalité d'opportunité, où le seuil pour la population défavorisée est la seuil induit par •, choisi de manière à avoir le même taux de vrais positifs que sur la population favorisée •.

Parité démographique conditionnelle (Corbett-Davies *et al.* 2017)

On aura une parité démographique conditionnelle si (au choix)

$$\begin{cases} \mathbb{P}[\widehat{Y} = y | \boldsymbol{X}_L = \boldsymbol{x}, P = 0] = \mathbb{P}[\widehat{Y} = y | \boldsymbol{X}_L = \boldsymbol{x}, P = 1], \ \forall y \in \{0, 1\} \\ \mathbb{P}[\widehat{Y} \leq y | \boldsymbol{X}_L = \boldsymbol{x}, P = 0] = \mathbb{P}[\widehat{Y} \leq y | \boldsymbol{X}_L = \boldsymbol{x}, P = 1], \ \forall y \in \mathbb{R} \\ \mathbb{E}[\widehat{Y} | \boldsymbol{X}_L = \boldsymbol{x}, P = 0] = \mathbb{E}[\widehat{Y} | \boldsymbol{X}_L = \boldsymbol{x}, P = 1], \end{cases}$$

où L désigne un sous-ensemble "légitime" de covariables non-protégées.

4.3.4 Équilibre des classes

Au lieu de prédire la valeur de $\widehat{y}$ (conditionnellement à y et p), Kleinberg, Mullainathan et Raghavan 2016 avaient suggéré de prédire la valeur moyenne de $s(\boldsymbol{x})$,

Équilibre des classes - Balance Class (Kleinberg, Mullainathan et Raghavan 2016)

On aura un équilibre des classes au sens faible si

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

ou au sens fort si

$$\mathbb{P}[S(\boldsymbol{X}) \leq s | Y = 1, P = 0] = \mathbb{P}[S(\boldsymbol{X}) \leq s | Y = 1, P = 1], \ \forall s \in [0, 1], \ \forall y \in \{0, 1\}.$$

4.3.5 Égalité de traitement

Égalité de traitement : (Berk, Heidari *et al.* 2021)

On a un égalité de traitement, le taux des faux positifs et des faux négatifs sont identiques dans les groupes protégés,

$$\frac{\mathbb{P}[\widehat{Y}=1|P=0,Y=0]}{\mathbb{P}[\widehat{Y}=0|P=0,Y=1]}=\frac{\mathbb{P}[\widehat{Y}=1|P=1,Y=0]}{\mathbb{P}[\widehat{Y}=0|P=1,Y=1]}.$$

Berk, Heidari *et al.* 2021 utilisent le terme de traitement en lien avec l'inférence causale dont nous parlerons ensuite. Si le classifieur produit plus de faux négatifs que de faux positifs pour le groupe supposé privilégié, cela signifie que plus d'individus défavorisés reçoivent un résultat favorable que l'inverse.

Une version un peu différente avait été proposée par Jung *et al.* 2020,

Égalité des désincitations - Equalizing Disincentives (Jung *et al.* 2020)

La différence entre le taux de vrais positifs et le taux de faux positifs doit être identique dans les groupes protégés,

$$\mathbb{P}[\widehat{Y}=1|P=0,Y=1]-\mathbb{P}[\widehat{Y}=1|P=0,Y=0]=\mathbb{P}[\widehat{Y}=1|P=1,Y=1]-\mathbb{P}[\widehat{Y}=1|P=1,Y=0].$$

4.3.6 La calibration

Un troisième critère couramment utilisé est parfois appelé "suffisance", qui requiert l'indépendance entre la cible Y et de la variable sensible P, conditionnellement à un score donné $s(\boldsymbol{X})$ ou à une prévision donnée, $\widehat{Y}$, introduit par Sokolova *et al.* 2006, et repris plus tard par Kleinberg, Mullainathan et Raghavan 2016 et Zafar *et al.* 2017.

Parité de calibration, accuracy parity : (Kleinberg, Mullainathan et Raghavan 2016, Zafar *et al.* 2017)

On a parité de calibration si

$$\mathbb{P}[Y=1|s(\boldsymbol{X})=t,P=0]=\mathbb{P}[Y=1|s(\boldsymbol{X})=t,P=1],\ \forall t.$$

On peut aller plus loin en demandant un peu plus, en demandant non seulement la parité, mais aussi une bonne calibration :

Bonne calibration (Kleinberg, Lakkaraju *et al.* 2017)

On a une équité de bonne calibration si

$$\mathbb{P}[Y=1|s(\boldsymbol{X})=t,P=0]=\mathbb{P}[Y=1|s(\boldsymbol{X})=t,P=1]=t,\ \forall t.$$

Dans la plupart des définitions que nous avions vues, nous nous intéressions à $\widehat{y}$, mais il est aussi

possible d'utiliser le score s. On peut ainsi définir $S = s(\boldsymbol{X})$ et viser comme objectif

$$Y \perp\!\!\!\perp P \mid S,$$

de sorte que

$$\mathbb{P}(Y = 1|S = t, p = 0) = \mathbb{P}(Y = 1|S = t, p = 1) = \mathbb{P}(Y = 1|S = t),\ \forall t \in [0, 1].$$

Or quand nous avions défini le biais, nous avions noté qu'un modèle était sans biais local (bien calibré) si $\mathbb{E}[Y|\widehat{Y} = y]$ est égale à y pour un modèle de régression, et $\mathbb{P}[Y = 1|S = t] = t$ pour un classifieur. Autrement dit, la condition de "bonne" calibration rajoute une condition dans l'écriture précédente,

$$\mathbb{P}(Y = 1|S = t, p = 0) = \mathbb{P}(Y = 1|S = t, p = 1) = \mathbb{P}(Y = 1|S = t) = t,\ \forall t \in [0, 1],$$

comme l'ont défini Verma et Rubin 2018.

Parité prédictive (1) - outcome test (Chouldechova 2017)

On a une parité prédictive si

$$\mathbb{P}[Y = 1|\widehat{Y} = 1, P = 0] = \mathbb{P}[Y = 1|\widehat{Y} = 1, P = 1].$$

Notons que si $\widehat{y}$ n'est pas un classifieur parfait ($\mathbb{P}[\widehat{Y} \neq Y] > 0$), et si les deux groupes ne sont pas équilibrés ($\mathbb{P}[P = 0] \neq \mathbb{P}[P = 1]$), il est impossible d'avoir en même temps la parité prédictive et l'égalité des opportunités. Notons que

$$\text{PPV}_p = \frac{\text{TPR} \cdot \mathbb{P}[P = p]}{\text{TPR} \cdot \mathbb{P}[P = p] + \text{FPR} \cdot (1 - \mathbb{P}[P = p])},\ \forall p \in \{0, 1\},$$

de telle sorte que $\text{PPV}_0 = \text{PPV}_1$ implique que soit TPR, soit FPR est nul, et comme

$$\text{NPV}_p = \frac{(1 - \text{FPR}) \cdot (1 - \mathbb{P}[P = p])}{(1 - \text{TPR}) \cdot \mathbb{P}[P = p] + (1 - \text{FPR}) \cdot (1 - \mathbb{P}[P = p])},\ \forall p \in \{0, 1\},$$

de telle sorte que $\text{NPV}_0 \neq \text{NPV}_1$, et la parité prédictive ne peut être atteinte.

Poursuivant le formalisme de Chouldechova 2017, Barocas, Hardt *et al.* 2019 ont proposé une extension de la parité prédictive

Parité prédictive (2) (Barocas, Hardt *et al.* 2019)

$$\begin{cases} \mathbb{P}[Y = 1|P = 0, \widehat{Y} = 1] = \mathbb{P}[Y = 1|P = 1, \widehat{Y} = 1] & \text{prédiction positive} \\ \mathbb{P}[Y = 1|P = 0, \widehat{Y} = 0] = \mathbb{P}[Y = 1|P = 1, \widehat{Y} = 0] & \text{prédiction négative,} \end{cases}$$

ou

$$\mathbb{P}[Y = 1|P = 0, \widehat{Y} = \widehat{y}] = \mathbb{P}[Y = 1|P = 1, \widehat{Y} = \widehat{y}],\ \forall \widehat{y} \in \{0, 1\}.$$

Finalement, notons que Kleinberg, Lakkaraju *et al.* 2017 ont introduit une notion d'équilibre par classe ("*balance for positive / negative class*").

$$\begin{cases} \mathbb{E}(S|Y = 1, P = 1) = \mathbb{E}(S|Y = 1, P = 0), & \text{équilibre pour la classe positive} \\ \mathbb{E}(S|Y = 0, P = 1) = \mathbb{E}(S|Y = 0, P = 0), & \text{équilibre pour la classe négative.} \end{cases}$$

4.3.7 Principe de non-reconstruction

Une dernière approche enfin peut être inspirée de Kim 2017, pour qui, une autre façon de définir si une classification est juste, ou pas, est de tester si on peut affirmer, à partir du résultat, si le sujet était membre d'un groupe protégé ou non. En d'autres termes, si le résultat d'un individu ne nous permet pas de prédire les attributs de cet individu mieux qu'en les devinant en l'absence d'informations, nous pouvons dire que ce résultat a été attribué de manière équitable.

Non-reconstruction de l'attribut protégé (Kim 2017)

Si nous ne pouvons pas dire, à partir du résultat ($\boldsymbol{x}$, $s(\boldsymbol{x})$, y et $\widehat{y}$), si le sujet était membre d'un groupe protégé ou non, on parlera d'équité par non-reconstruction de l'attribut protégé

$$\mathbb{P}[P = 0|\boldsymbol{X}, s(\boldsymbol{X}), \widehat{Y}, Y] = \mathbb{P}[P = 1|\boldsymbol{X}, s(\boldsymbol{X}), \widehat{Y}, Y]$$

4.3.8 Comparaison des critères d'équité

Pour résumer, on peut considérer la Figure 4.17. La parité démographique se traduirait par

$$\mathbb{P}(\widehat{Y} = 1|P = p) = \mathbb{P}(\widehat{Y} = 1),\ \forall p$$

soit encore TP + FP doit être identique sur les deux groupes ($p = 0$ et $p = 1$). Ici, c'est le cas, car le taux de positif vaut 50 % dans les deux groupes. Pour la notion d'égalité des chances, signifie

$$\mathbb{P}(\widehat{Y} = 1|P = 0, Y = y) = \mathbb{P}(\widehat{Y} = 1|Y = 1, Y = y) = \mathbb{P}(\widehat{Y} = 1|Y = y),\ \forall y,$$

autrement dit les taux de faux positifs et de faux négatifs doivent être identiques,

$$\frac{\text{FP}}{\text{TN} + \text{FP}} \text{ et } \frac{\text{FN}}{\text{TP} + \text{FN}} \text{ doivent être identiques sur les deux groupes.}$$

C'est ici le cas, car le taux de faux positifs vaut 50 %, que p vaille 0 ou 1 (avec 30/60 contre 20/40), et le taux de faux négatifs vaut aussi 50 % dans les deux groupes. Pour la parité prédictive

$$\mathbb{P}(Y = 1|P = 0, \widehat{Y} = y) = \mathbb{P}(Y = 1|P = 1, \widehat{Y} = y) = \mathbb{P}(Y = 1|\widehat{Y} = y),\ \forall y$$

autrement dit, les valeurs prédictives positives et négatives doivent être identiques,

$$\frac{\text{TP}}{\text{TP} + \text{FP}} \text{ et } \frac{\text{TN}}{\text{TN} + \text{FN}} \text{ doivent être identiques sur les deux groupes.}$$

Or ici les valeurs prédictives positives valent respectivement 60 % et 40 %, suivant la valeur de p (30/50 et 20/50). Pour l'*accuracy* globale,

$$\mathbb{P}(\widehat{Y} = Y|P = p) = \mathbb{P}(\widehat{Y} = Y),\ \forall p$$

soit encore TP + TN doit être identique sur les deux groupes ($p = 0$ et $p = 1$). C'est ici le cas car le taux d'observations bien classées vaut 50 %. Finalement, pour la notion d'égalité de traitement,

$$\mathbb{P}(\widehat{Y} = Y|P = p) = \mathbb{P}(\widehat{Y} = Y),\ \forall p$$

soit encore FP/FN doit être identique sur les deux groupes ($p = 0$ et $p = 1$). Or ici, les taux valent respectivement 3/2 et 2/3, qui ne sont pas égaux.

On peut résumer les différentes notions d'équité dans le tableau 4.1.

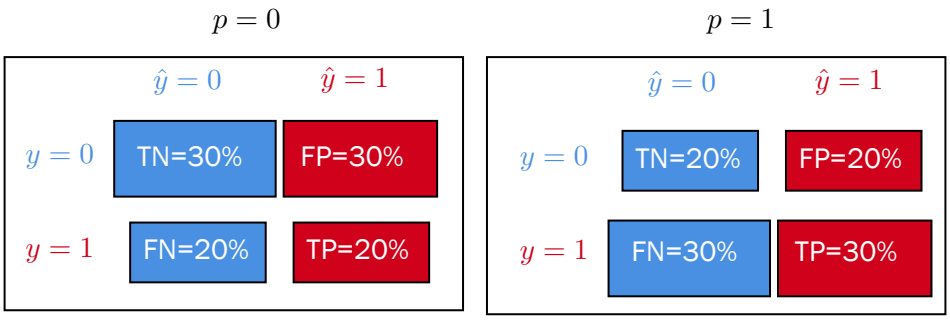


Figure 4.17 – Équité d'un classifieur à partir des matrices de confusion, sur les deux sous-populations, $p = 0$ à gauche et $p = 1$ à droite.

statistical parity	Dwork *et al.* 2012	$\mathbb{P}[\widehat{Y} = 1 \mid P = p] = \text{cst},\ \forall p$	independence
conditional statistical parity	Corbett-Davies *et al.* 2017	$\mathbb{P}[\widehat{Y} = 1 \mid P = p, X = x] = \text{cst}_x,\ \forall p, y$	$\widehat{Y} \perp\!\!\!\perp P$
equalized odds	Hardt *et al.* 2016	$\mathbb{P}[\widehat{Y} = 1 \mid P = p, Y = y] = \text{cst}_y,\ \forall p, y$	separation
equalized opportunity	Hardt *et al.* 2016	$\mathbb{P}[\widehat{Y} = 1 \mid P = p, Y = 1] = \text{cst},\ \forall p$	
predictive equality	Corbett-Davies *et al.* 2017	$\mathbb{P}[\widehat{Y} = 1 \mid P = p, Y = 0] = \text{cst},\ \forall p$	$\widehat{Y} \perp\!\!\!\perp P \mid Y$
balance (positive)	Kleinberg, Lakkaraju *et al.* 2017	$\mathbb{E}[S \mid P = p, Y = 1] = \text{cst},\ \forall p$	$S \perp\!\!\!\perp P \mid Y$
balance (negative)	Kleinberg, Lakkaraju *et al.* 2017	$\mathbb{E}[S \mid P = p, Y = 0] = \text{cst},\ \forall p$	
conditional accuracy equality	Berk, Heidari *et al.* 2017	$\mathbb{P}[Y = y \mid P = p, \widehat{Y} = y] = \text{cst}_y,\ \forall p, y$	sufficiency
predictive parity	Chouldechova 2017	$\mathbb{P}[Y = 1 \mid P = p, \widehat{Y} = 1] = \text{cst},\ \forall p$	
calibration	Chouldechova 2017	$\mathbb{P}[Y = 1 \mid P = p, S = s] = \text{cst}_s,\ \forall p, s$	$Y \perp\!\!\!\perp P \mid \widehat{Y}$
well-calibration	Chouldechova 2017	$\mathbb{P}[Y = 1 \mid P = p, S = s] = s,\ \forall p, s$	
accuracy equality	Berk, Heidari *et al.* 2017	$\mathbb{P}[\widehat{Y} = Y \mid P = p] = \text{cst},\ \forall p$	
treatment equality	Berk, Heidari *et al.* 2017	$\frac{\text{FN}_p}{\text{FP}_p} = \text{cst}_p,\ \forall p$	

Table 4.1 – Définitions de l'équité par groupes.

4.3.9 Relaxation et intervalles de confiance

Seuil exogène et relaxation

Nous avions vu que l'équité démographique se traduit par l'égalité

$$\frac{\mathbb{P}[\widehat{Y} = 1 \mid P = 0]}{\mathbb{P}[\widehat{Y} = 1 \mid P = 1]} = 1 = \frac{\mathbb{P}[\widehat{Y} = 1 \mid P = 1]}{\mathbb{P}[\widehat{Y} = 1 \mid P = 0]}.$$

Si cette approche est intellectuellement intéressante, la réalité statistique fait qu'avoir une égalité parfaite entre deux probabilités (prédictives) est souvent impossible.

Impact disparate : (Feldman *et al.* 2015)

Une fonction de décision $\widehat{Y}$ a un impact disparate si, pour un seuil τ choisi a priori

$$\min\left\{\frac{\mathbb{P}[\widehat{Y}=1|P=0]}{\mathbb{P}[\widehat{Y}=1|P=1]}, \frac{\mathbb{P}[\widehat{Y}=1|P=1]}{\mathbb{P}[\widehat{Y}=1|P=0]}\right\} < \tau \text{ (souvent 80 \%).}$$

Cette règle dite des "quatre cinquièmes", associée au seuil $\tau = 80$ % a été définie à l'origine par le comité consultatif technique sur les tests de la Commission des pratiques équitables en matière d'emploi de l'État de Californie (FEPC, "*Fair Employment Practice Commission*"), qui a publié les directives de l'État de Californie sur les procédures de sélection des employés en octobre 1972, comme le rappellent Feldman *et al.* 2015, Mercat-Bruns 2016 ou encore Biddle 2017. Cette norme s'est ensuite imposée dans les "*Uniform Guidelines on Employee Selection Procedures* de 1978, utilisées par l'*Equal Employment Opportunity Commission* (EEOC), le ministère du travail et le ministère de la justice des États-Unis. Un point important ici est que cette forme de discrimination se produisait même lorsque l'employeur n'avait pas l'intention de discriminer, mais en regardant les statistiques de l'emploi (sur le genre ou un critère racial), il était possible d'observer (et de corriger) un biais discriminatoire.

Par exemple, sur les données de la Figure 4.9,

$$\frac{\mathbb{P}[\widehat{Y}=1|P=\bullet]}{\mathbb{P}[\widehat{Y}=1|P=\bullet]} = \frac{1}{3} \ll 80\ \%.$$

Une autre approche, suggérée pour relaxer l'égalité $\mathbb{P}(\widehat{Y}=1|P=0) = \mathbb{P}(\widehat{Y}=1|P=1)$, consiste à instruire une notion d'ε-équité

$$|\mathbb{P}(\widehat{Y}=1|P=0) - \mathbb{P}(\widehat{Y}=1|P=1)| < \varepsilon.$$

L'écart de gauche étant parfois appelé "différence de parité statistique" (SPD). Žliobaite 2015 suggère de normaliser la différence de parité statistique,

$$NSPD = \frac{SPD}{D_{\max}} \text{ où } D_{\max} = \min\left\{\frac{\mathbb{P}(\widehat{y}=1)}{\mathbb{P}(p=1)}, \frac{\mathbb{P}(\widehat{y}=0)}{\mathbb{P}(p=0)}\right\},$$

de telle sorte que $NSPD = 1$ pour une discrimination maximale.

Seuil endogène et intervalles de confiance

Besse *et al.* 2018 proposent une autre approche, sur la base d'intervalle de confiance pour des critères d'équité. Par exemple, pour l'impact disparate, nous avons vu qu'il fallait calculer

$$T = \frac{\mathbb{P}[Y=1|P=0]}{\mathbb{P}[Y=1|P=1]},$$

dont la version empirique est

$$T_n = \frac{\sum_i y_i \mathbf{1}(p_i=0)}{\sum_i y_i \mathbf{1}(p_i=1)} \cdot \frac{\sum_i \mathbf{1}(p_i=1)}{\sum_i \mathbf{1}(p_i=0)},$$

que l'on peut utiliser pour construire un intervalle de confiance pour T (Besse *et al.* 2018 proposent un test asymptotique, mais des méthodes de rééchantillonage sont envisageables).

4.3.10 Indépendance conditionnelle

Les conditions écrites en termes d'indépendance conditionnelle – par exemple $\widehat{Y} \perp\!\!\!\perp P \mid Y$ pour l'égalité de chances – est aussi appelée "condition de séparation", et sera largement utilisée quand on évoquera les graphes causaux (dans la section suivante). On peut rappeler que cette condition est très proche de la propriété de Markov à l'ordre 1,

$$(X_n) \text{ vérifie la propriété de Markov } \iff X_{n+1} \perp\!\!\!\perp \{X_{n-1}, X_{n-2}, \cdots\} \mid X_n.$$

Car intuitivement, cela signifie que connaître X_n enlève tout intérêt à X_{n-1} pour prédire X_{n+1}. Cette propriété est aussi parfois appelée "propriété de suffisance". On peut en effet montrer qu'on a un théorème de séparabilité de la forme

$$X \perp\!\!\!\perp Y \mid Z \iff \exists g, h : \mathbb{P}[X = x, Y = y, Z = z] = g(x, z) \cdot h(y, z).$$

Notons que si $X \perp\!\!\!\perp Y \mid Z$ et $X \perp\!\!\!\perp Z$ alors on a l'indépendence non-conditionnelle entre nos deux variables, $X \perp\!\!\!\perp Y$. Et si on a à la fois $X \perp\!\!\!\perp Y \mid Z$ et $X \perp\!\!\!\perp Y$ alors soit $X \perp\!\!\!\perp Z$, soit $Y \perp\!\!\!\perp Z$. Comme nous le verrons par la suite

$$X \perp\!\!\!\perp Y \mid Z \iff Y \perp\!\!\!\perp X \mid Z,$$

ce qui posera des soucis quand on essayera de travailler sur des graphes causaux *dirigés*. De plus

$$X \perp\!\!\!\perp Y \mid Z \text{ et } U = h(Y) \implies X \perp\!\!\!\perp U \mid Z.$$

Une notion plus faible est la version au second ordre, $X \perp Y \mid Z$, signifiant simplement que la corrélation partielle est nulle,

$$r_{XY|Z} = \text{cov}[X - \mathbb{E}(X|Z), Y - \mathbb{E}(Y|Z)] = 0.$$

4.3.11 Mise en œuvre et comparaison

Sur l'exemple discret (avec 24 observations) de la Figure 4.7, repris dans le tableau 4.2, on obtient les valeurs du tableau [51] 4.3.

p	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
y	0	0	1	0	1	1	0	0	1	0	1	1	0	0	0	1	1	1	1	0	1	1	1	1
$\widehat{y}$	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 4.2 – Données de la Figure 4.7, ordonnées suivant leur score S (non indiqué ici) avec un seuil identique pour les deux groupes ($p \in \{\bullet, \bullet\}$), et avec $\widehat{y} = \mathbf{1}_{[\text{seuil},1]}(s)$.

Sur l'exemple continu de la Figure 4.5, on obtient les valeurs du tableau 4.4, avec un même seuil de 60 % pour les deux groupes, comme sur la Figure 4.18, ou alors avec deux seuils différents, avec 56 % et 66 % respectivement pour les groupes bleu • (ou $p = 0$) et rouge • (ou $p = 1$), comme sur la Figure 4.19. Sur les Figures 4.18 et 4.19, en haut figurent les densités conditionnelles de $S|Y = 1, P = \bullet$ (trait plein) et $S|Y = 0, P = \bullet$ (en pointillés), à gauche (population supposée favorisée), et

51. Dans les tableaux 4.3, 4.4 et 4.5, la colonne 'diff' donne la différence absolue entre les deux probabilités (exprimées en pourcentage) et la colonne '(%)' donne la différence relative entre les deux probabilités, exprimée en pourcentage.

Nom	Formule probabiliste	•	•	diff	(%)
statistical parity	$\mathbb{P}[\widehat{Y}=1\vert P=\circ]$	25.0 %	75.0 %	50.0	+200.0 %
equalized opportunity	$\mathbb{P}[\widehat{Y}=1\vert P=\circ, Y=1]$	40.0 %	88.9 %	48.9	+122.2 %
predictive equality	$\mathbb{P}[\widehat{Y}=1\vert P=\circ, Y=0]$	0.0 %	57.1 %	57.1	-
conditional accuracy	$\mathbb{P}[Y=0\vert P=\circ, \widehat{Y}=0]$	50.0 %	75.0 %	25.0	+50.0 %
predictive parity	$\mathbb{P}[Y=1\vert P=\circ, \widehat{Y}=1]$	100.0 %	66.7 %	-33.3	-33.3 %
accuracy equality	$\mathbb{P}[\widehat{Y}=Y\vert P=\circ]$	62.6 %	68.8 %	6.2	+10.0 %
treatment equality	$\mathrm{FN}_\circ/\mathrm{FP}_\circ$	-	25.0 %	-	-

Table 4.3 – Données de la Figure 4.7 et de la Table 4.2, avec les différents concepts d'équité, les valeurs des mesures pour les deux groupes, $p \in \{\bullet, \bullet\}$, la différence absolue et la différence relative (en pourcentage).

les densités conditionnelles de $S|Y=1, P=\bullet$ et $S|Y=0, P=\bullet$, à droite (population supposée défavorisée). En bas, figurent les fonctions de survie, $t \mapsto \mathbb{P}(S > t|Y=y, P=\circ)$, avec, sur la Figure 4.18,

$$\widehat{y} = \begin{cases} \mathbf{1}_{[60\ \%, 100\ \%]}(s) \text{ si } p = \bullet \text{ ou } 0 \\ \mathbf{1}_{[60\ \%, 100\ \%]}(s) \text{ si } p = \bullet \text{ ou } 1, \end{cases}$$

et sur la Figure 4.19,

$$\widehat{y} = \begin{cases} \mathbf{1}_{[56\ \%, 100\ \%]}(s) \text{ si } p = \bullet \text{ ou } 0 \\ \mathbf{1}_{[66\ \%, 100\ \%]}(s) \text{ si } p = \bullet \text{ ou } 1. \end{cases}$$

Les probabilités indiquées sur les fonctions de survie sont les probabilités de "vrais" positifs ou négatifs, selon, avec par exemple

$$\mathbb{P}(\widehat{Y}=0|Y=0, P=\bullet) = 90.4\ \% \text{ et } \mathbb{P}(\widehat{Y}=1|Y=1, P=\bullet) = 66.3\ \%$$

alors que les taux de "faux" négatifs ou positifs, sont respectivement

$$\mathbb{P}(\widehat{Y}=1|Y=0, P=\bullet) = 9.6\ \% \text{ et } \mathbb{P}(\widehat{Y}=0|Y=1, P=\bullet) = 33.7\ \%.$$

4.4 Équité au niveau individuel

Dans la section précédente, nous nous intéressions à une notion d'équité "des groupes" (avec des sous-groupes constitués par les valeurs de y, p et / ou $\widehat{y}$). L'équité individuelle est une notion relativement différente. Les trois critères précédents sont tous basés sur le groupe alors que l'équité individuelle, comme son nom l'indique, est basée sur l'individu. Elle a été proposée pour la première fois dans Dwork *et al.* 2012. La notion d'équité individuelle met l'accent sur le fait que des individus similaires (sur les attributs non-protégés) doivent être traités de manière similaire.

4.4.1 Proximité entre individus (et propriété de Lipschitz)

L'idée naturelle, que l'on trouve chez Luong *et al.* 2011, est que deux individus "proches" (au sens des caractéristiques non-protégées $\boldsymbol{x}$) doivent avoir la même prévision. Considérons deux

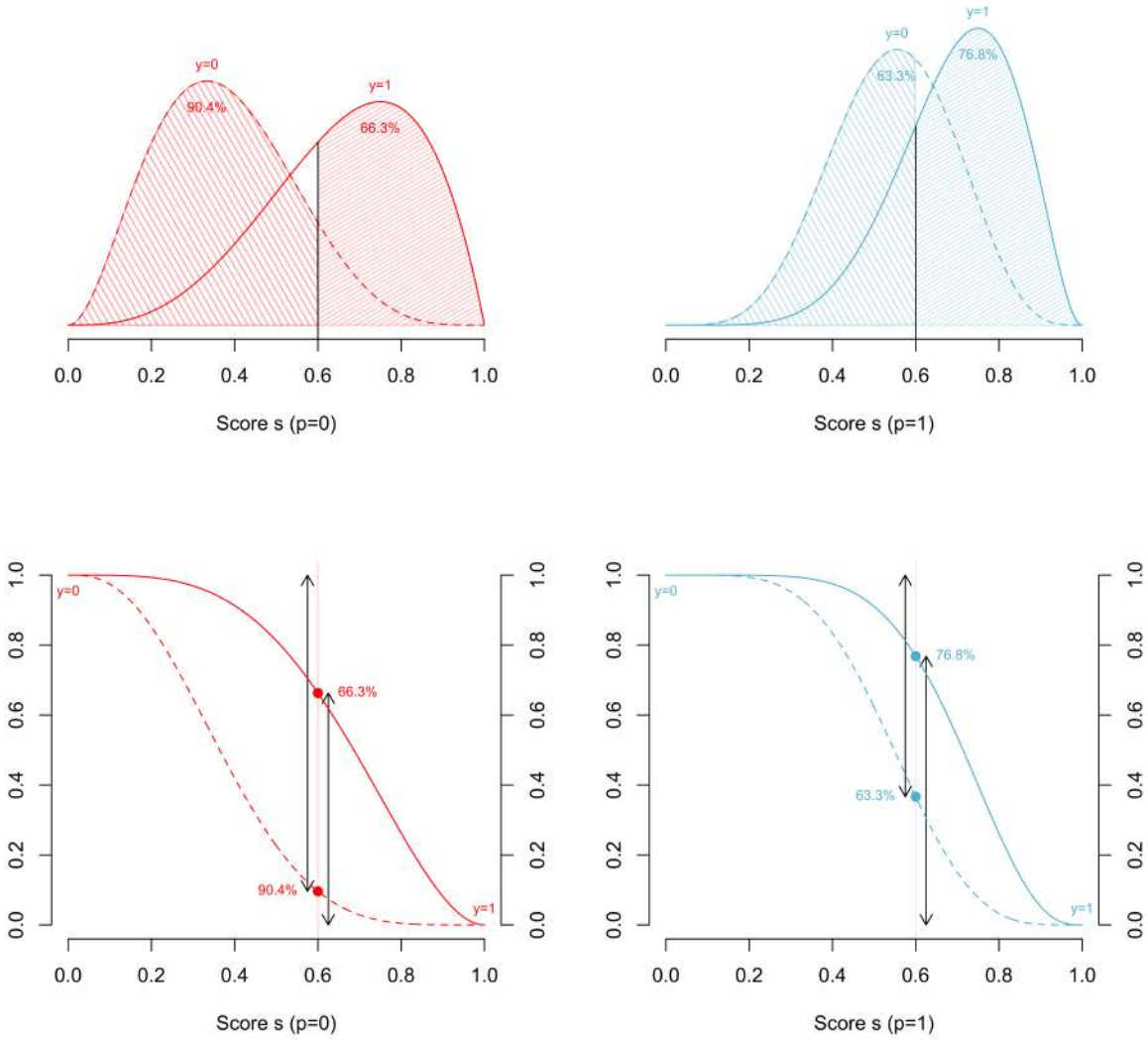


Figure 4.18 – Exemple continu, avec le même seul $(60\ \%)$ dans les deux groupes, $p \in \{\bullet, \bullet\}$ ou $p \in \{0, 1\}$, avec la densité de s en haut, et sa fonction de survie en bas, et $\widehat{y} = \mathbf{1}_{[60\ \%, 100\ \%]}(s)$.

métriques, une sur $[0,1] \times [0,1]$ ou $\{0,1\} \times \{0,1\}$ notée D, et une sur $\mathcal{X} \times \mathcal{X}$ notée d, de telle sorte qu'on aura une équité individuelle sur une base de données de taille n si on a la propriété (dite de Lipschitz) suivante

$$D(\widehat{y}_i, \widehat{y}_j) \text{ ou } D(s_i, s_j) \leq d(\boldsymbol{x}_i, \boldsymbol{x}_j), \ \forall i, j = 1, \cdots, n.$$

Duivesteijn et Feelders 2008 parlaient de classification monotone. Il est difficile de déterminer quelle est la métrique à utiliser pour mesurer la similarité de deux individus (autrement dit entre $\boldsymbol{x}_i$ et $\boldsymbol{x}_j$), comme l'expliquaient Kim *et al.* 2018. Le plus usuel est d'utiliser une distance de type Mahalanobis, pour tenir compte des échelles différentes entre les variables.

Nom	Formule probabiliste	•/0	•/1	diff	(%)
statistical parity	$\mathbb{P}[\widehat{Y}=1\|P=\circ]$	38 %	56.8 %	18.8	49.5 %
equalized opportunity	$\mathbb{P}[\widehat{Y}=1\|P=\circ,Y=1]$	66.3 %	76.8 %	10.5	15.9 %
predictive equality	$\mathbb{P}[\widehat{Y}=1\|P=\circ,Y=0]$	9.6 %	36.7 %	27.1	281.2 %
conditional accuracy	$\mathbb{P}[Y=0\|P=\circ,\widehat{Y}=0]$	72.8 %	73.2 %	0.4	0.6 %
predictive parity	$\mathbb{P}[Y=1\|P=\circ,\widehat{Y}=1]$	87.3 %	67.7 %	-19.6	-22.6 %
accuracy equality	$\mathbb{P}[\widehat{Y}=Y\|P=\circ]$	62 %	56.8 %	-5.3	-8.6 %
treatment equality	$\mathrm{FN}_\circ/\mathrm{FP}_\circ$	350.1	63.2	-286.9	-82 %

Table 4.4 – Différents concepts d'équité sur la base de la Figure 4.18, avec les valeurs des mesures pour les deux groupes, $p \in \{\bullet, \bullet\}$ ou $p \in \{0,1\}$, la différence absolue et la différence relative (en pourcentage), avec le même seuil (60 %) dans les deux groupes pour obtenir $\widehat{y}$ à partir du score s.

Nom	Formule probabiliste	• / 0	•/1	diff	(%)
statistical parity	$\mathbb{P}[\widehat{Y}=1\|P=\circ]$	44.8 %	45.6 %	0.7	+1.6 %
equalized opportunity	$\mathbb{P}[\widehat{Y}=1\|P=\circ,Y=1]$	74.4 %	66.3 %	-8.1	-10.9 %
predictive equality	$\mathbb{P}[\widehat{Y}=1\|P=\circ,Y=0]$	15.3 %	24.9 %	9.6	62.6 %
conditional accuracy	$\mathbb{P}[Y=0\|P=\circ,\widehat{Y}=0]$	76.8 %	69 %	-7.8	-10.1 %
predictive parity	$\mathbb{P}[Y=1\|P=\circ,\widehat{Y}=1]$	82.9 %	72.7 %	-10.2	-12.3 %
accuracy equality	$\mathbb{P}[\widehat{Y}=Y\|P=\circ]$	55.2 %	45.6 %	-9.6	-17.4 %
treatment equality	$\mathrm{FN}_\circ/\mathrm{FP}_\circ$	167.5	135.7	-31.8	-19 %

Table 4.5 – Différents concepts d'équité sur la base de la Figure 4.19, avec les valeurs des mesures pour les deux groupes, $p \in \{\bullet, \bullet\}$, la différence absolue et la différence relative (en pourcentage), avec un seuil différent (56 % si $p = \bullet$ et 66 % si $p = \bullet$) dans les deux groupes pour obtenir $\widehat{y}$ à partir du score s.

4.4.2 Causalité dans un contexte dynamique

Avant de définir la "causalité" dans le contexte des données individuelles, rappelons que, dans un contexte de données temporelles, Granger 1969 a introduit un concept de "causalité" qui prend une forme relativement simple pour des dynamiques décrites par une corrélation

$$\begin{cases} x_{1,t+1} = c_1 + a_{1,1}x_{1,t} + a_{1,2}x_{2,t} + \varepsilon_{1,t} \\ x_{2,t+1} = c_2 + a_{2,1}x_{1,t} + a_{2,2}x_{2,t} + \varepsilon_{2,t} \end{cases}$$

noté aussi $\boldsymbol{x}_{t+1} = \boldsymbol{c} + \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{\varepsilon}_{t+1}$, où les termes hors-diagonale de la matrice d'autorégression $\boldsymbol{A}$ permettent de quantifier la "*lagged causality*", autrement dit un effet causal décalé (entre t et $t+1$) avec respectivement $x_1 \to x_2$ or $x_1 \leftarrow x_2$. Par exemple la Figure 4.20 montre le nuage de points $(x_{1,t}, x_{2,t+1})$ et $(x_{2,t}, x_{1,t+1})$, à gauche et à droite, où respectivement x_1 désigne le nombre de cyclistes à Stockholm, par jour, en 2014 (à une intersection routière donnée) et x_2 désigne la température (moyenne) le même jour. Le graphique de gauche revient à se demander si la température *cause* le nombre de cyclistes (si la température monte, le nombre de cyclistes sur les routes augmente) et le graphique de droite revient à se demander si le nombre de cyclistes

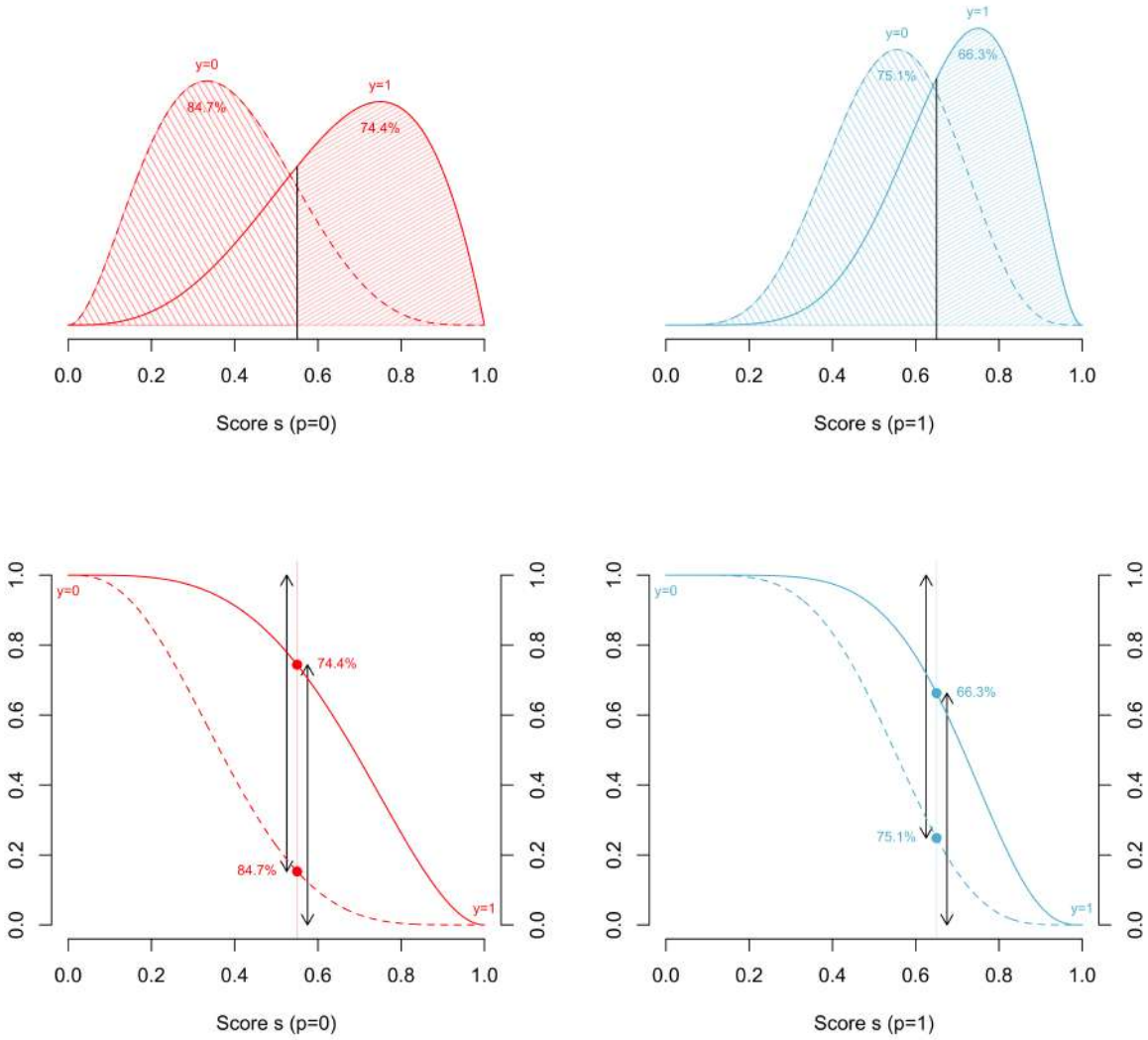


Figure 4.19 – Exemple continu, avec des seuils différents (56 % et 66 %) dans les deux groupes, $p \in \{\bullet, \bullet\}$ ou $p \in \{0, 1\}$, avec la densité de s en haut, et sa fonction de survie en bas, et $\widehat{y} = \mathbf{1}_{[56\ \%, 100\ \%]}(s)$ si $p = \bullet$ et $\widehat{y} = \mathbf{1}_{[66\ \%, 100\ \%]}(s)$ si $p = \bullet$.

cause la température (si le nombre de cyclistes sur les routes augmente, la température monte). Dans les deux cas, si on estime

$$\begin{cases} x_1 \to x_2 : & x_{2,t+1} = \gamma_1 + \alpha_{2,1} x_{1,t} + \eta_{1,t}, \\ x_1 \leftarrow x_2 : & x_{1,t+1} = \gamma_2 + \alpha_{1,2} x_{2,t} + \eta_{2,t}, \end{cases}$$

on observe des régressions significatives (mais ce n'est pas un test causal)

$$\begin{cases} x_1 \to x_2 : & x_{2,t+1} = \underset{(234)}{4320} + \underset{(23)}{757}\, x_{1,t} + \eta_{1,t},\ R^2 = 75.72\,\% \\ x_1 \leftarrow x_2 : & x_{1,t+1} = \underset{(0.04)}{-1.98} + \underset{(0.00003)}{0.001}\, x_{2,t} + \eta_{2,t},\ R^2 = 72.41\,\% \end{cases}$$

On peut utiliser le test de Granger, sur les données de la Figure 4.20, sur les deux hypothèses causales (non pas sur les niveaux mais sur les variations journalières, du nombre de cyclistes, et de la température)

$$\begin{cases} x_1 \to x_2 : & H_0 : a_{2,1} = 0,\ p - \text{value} = 56.66\,\% \\ x_1 \leftarrow x_2 : & H_0 : a_{1,2} = 0,\ p - \text{value} = 0.004\,\% \end{cases}$$

Autrement dit, fort logiquement, on observe que la température a un lien causal sur la présence de cyclistes sur les routes, mais pas l'inverse.

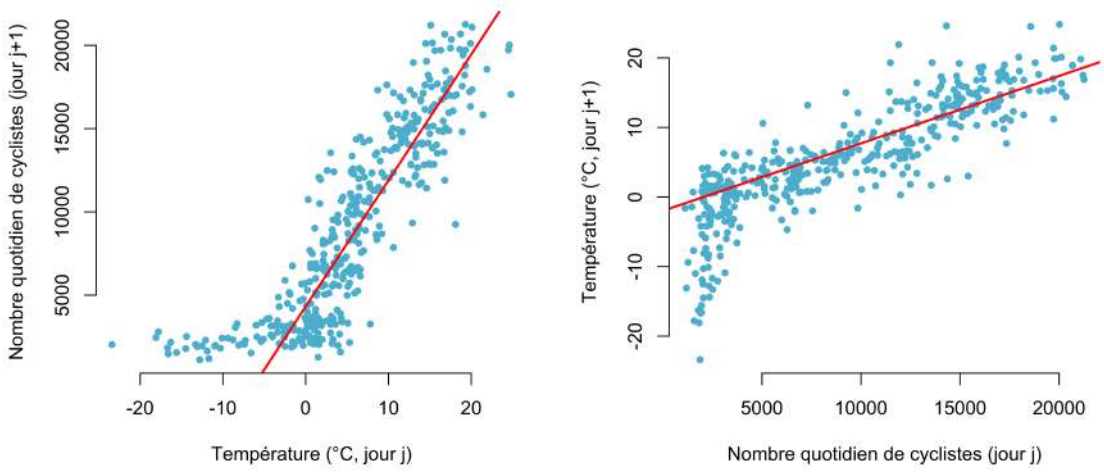


Figure 4.20 – Nombre de cyclistes par jour (x_2), en 2014, à Stockholm, et température quotidienne moyenne (x_1). Les droites de régression sont estimées uniquement sur les jours où la température excédait 0° C.

Dans un contexte non-dynamique, définir la causalité sera un exercice plus périlleux.

4.4.3 Causalité et graphs dirigés

Encore une fois, avoir des variables prédictives causales, en assurance, est particulièrement important. Swedloff 2014 affirmait "*the problem, of course, is that finding such non-causally related correlations means that the policyholder cannot, and likely should not, try to minimize the activity, behavior, or characteristic*". Malheureusement, les corrélations "fortes" (parfois suffisantes en apprentissage machine) ne suffisent pas, et Spirtes *et al.* 1993 notaient qu'il serait difficile de *définir* ce qu'est la causalité, et qu'il serait probablement plus simple de l'*axiomatiser*. En d'autres termes, on va commencer par déterminer les attributs qui seraient considérés comme nécessaires pour qu'une relation soit considérée comme *causale*, on formalise mathématiquement ces propriétés, et on regarde si ces axiomes se traduisent par une caractérisation interprétable d'une relation causale.

Par exemple, il semble légitime que ces relations soient transitives : si x_1 cause x_2 et si x_2 cause x_3, alors il doit également être vrai que x_1 cause x_3. On pourrait alors parler de causalité globale. Mais une version locale semble exister : si x_1 cause x_3 seulement par l'intermédiaire de x_2, alors il est possible de bloquer l'influence de x_1 sur x_3 si on empêche x_2 d'être influencée par x_1. On pourrait aussi demander à ce que la relation causale soit irréflexive, au sens où x_1 ne peut pas se causer elle-même. Le danger de cette propriété est qu'elle tend à chercher une explication causale à toute variable. Enfin, une propriété d'asymétrie de la relation est souvent souhaitée, dans le sens où x_1 cause x_2 implique que x_2 ne peut causer x_1. Comme le notait Wright 1921, bien avant Pearl 1988, l'outil le plus naturel pour décrire visuellement et simplement ces relations causales est probablement celui des graphes dirigés.

Une variable sera ici un nœud du réseau (par exemple x_1 ou x_2), et une relation causale, dans le sens "x_1 cause x_2" se traduira par une flèche dirigée de x_1 vers x_2 (comme nous le faisions sur les séries chronologiques). Comme sur les schémas (a) et (b) de la Figure 4.21. Dans l'exemple (a), Morgan et Winship 2015 parleront de "*mutual dependence*" entre x_2 et x_3, de "*mutual causation*" pour (c), et (b) correspond à un cas de "*mediation*".

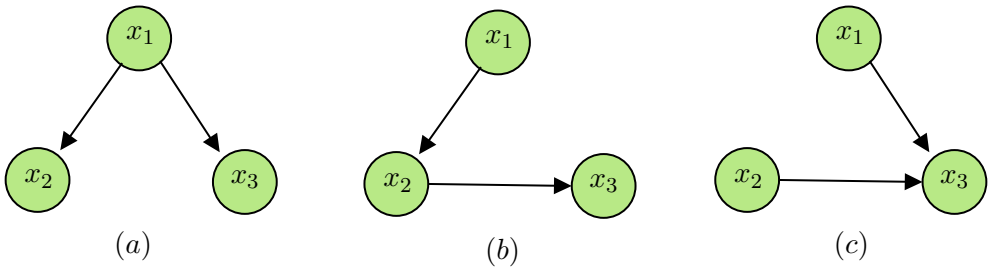


Figure 4.21 – Quelques exemples de graphes dirigés, avec 3 nœuds, et 2 connections. (a) correspond au cas où x_1 est un **facteur de confusion** pour x_2 et x_3, correspondant à un choc commun ou de dépendance mutuelle, (b) correspond au cas où x_2 est un **médiateur** pour x_1 et x_3, et (c) correspond au cas où x_3 est un **collisionneur** (ou "*collider*" en anglais) pour x_1 et x_2, correspondant à un cas de cause mutuelle.

Dans l'exemple (b), on dira que x_2 et x_3 sont causalement dépendantes de x_1, x_2 le sera directement et x_3 indirectement. On dira que x_1 est un parent causal de x_2 (une cause), et inversement que x_2 est un enfant causal de x_1 (une conséquence). Cette relation parent / enfant est associée à l'existence d'un lien entre les deux variables. On dira que x_1 est un ancêtre causal de x_3, et inversement que x_3 est un descendant causal de x_1. Cette relation ancêtre / descendant est associée à l'existence d'un chemin (dirigé) entre les deux variables, c'est-à-dire une succession de liens. S'il n'existe pas de chemin entre deux nœuds, on dira que les deux variables sont causalement indépendantes. Un collisionneur ("*collider*") est une variable qui est la conséquence de deux variables ou plus, comme x_3 sur (c). Ce type de variable est très lié au paradoxe de Berkson. Un non-collisionneur est une variable influencée par une seule, et elle permet de transmettre causalement une conséquence, le long d'un chemin. Les variables causales qui influencent le collisionneur ne sont elles-mêmes pas nécessairement associées. Si c'est le cas, on dit que le collisionneur est blindé ("*shielded*") et la variable est le sommet d'un triangle. Sur la Figure 4.22, x_4 est un descendant de x_1, un enfant de x_2 (et x_4), un parent de x_5 (et x_6) et un ancêtre de x_7. Les variables x_3 et x_5 ne sont pas causalement indépendantes. x_4 est un collisionneur, mais pas x_6. x_4 est un collisionneur non blindé ("*unshielded collider*") car x_2 et x_3 (les deux parents) ne sont pas connectés (ils ne sont pas pour autant indépendants).

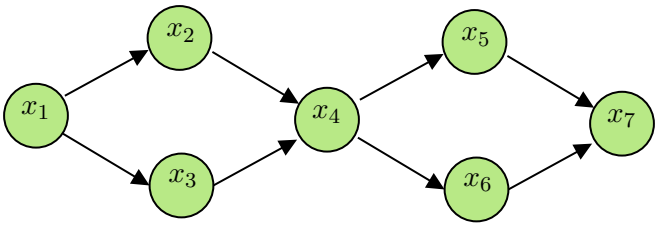


Figure 4.22 – Un exemple de graph dirigé.

Dans la section 2.3, nous avions évoqué le lien fort qui existait entre les proxys et les modèles causaux. Mais ce lien est souvent complexe. Par exemple, dans notre exemple avec la taille des chaussures et les capacités de lecture, l'âge de l'enfant est un facteur de confusion, et si l'âge n'est pas observable, la pointure de chaussure peut être un proxy de l'âge. Il existe d'autres formes de proxys particulièrement utilisés en science économique, comme le rappellent Stahlecker et Trenkler 1993, ou en science du climat, où la largeur des anneaux de croissance des arbres constitue un proxy des conditions environnementales historiques, ou la composition isotopique de l'oxygène dans les carottes de glace polaire (en fonction de la profondeur) constitue un proxy de la température moyenne passée. Dans ces deux cas, le proxy est utilisé car la variable d'intérêt est inobservable.

Considérons un graphe causal $\mathcal{G}$ (c'est-à-dire une collection de variables $\boldsymbol{x}$ et de flèches). Soient x_1 et x_2 deux nœuds, et $\boldsymbol{y}$ un ensemble de nœuds de $\mathcal{G}$ ne contenant ni x_1, ni x_2. On dira que x_1 et x_2 sont séparés (ou d-séparés, "*directed separation*") s'il n'existe pas de chemin non orienté u entre x_1 et x_2 tel que (a) chaque collisionneur (dans u) est soit dans $\boldsymbol{y}$, soit possède un descendant dans $\boldsymbol{y}$ et (b) aucun autre nœud sur u n'est dans $\boldsymbol{y}$. On notera alors $x_1 \perp\!\!\!\perp x_2|\boldsymbol{y}$. On a ici des conditions nécessaires et suffisantes pour que deux sommets d'un graphe causal soient indépendants d'un point de vue probabiliste, après conditionnement sur un autre ensemble de sommets. Car il est possible de traduire cette terminologie causale, présentée ici en terme de graphes, sous une forme probabiliste.

La formule des probabilités composées (ou "*probability chain rule*") permet de calculer la probabilité d'une intersection d'évènements à l'aide de probabilités conditionnelles, puisque

$$\mathbb{P}[A_1 \cap \cdots \cap A_n] = \mathbb{P}[A_1] \times \mathbb{P}_{A_1}(A_2) \times \mathbb{P}_{A_1 \cap A_2}[A_3] \times \cdots \times \mathbb{P}_{A_1 \cap \cdots \cap A_{n-1}}[A_n],$$

ce que l'on pourrait écrire

$$\mathbb{P}[x_1, \cdots, x_n] = \mathbb{P}[x_1] \times \mathbb{P}[x_2|x_1] \times \mathbb{P}[x_3|x_1, x_2] \times \cdots \times \mathbb{P}[x_n|x_1, \cdots, x_{n-1}].$$

Mais cette écriture n'est pas unique, puisqu'on pourrait aussi écrire (par exemple)

$$\mathbb{P}[x_1, \cdots, x_n] = \mathbb{P}[x_n] \times \mathbb{P}[x_{n-1}|x_n] \times \mathbb{P}[x_{n-2}|x_n, x_{n-1}] \times \cdots \times \mathbb{P}[x_1|x_n, \cdots, x_2].$$

Notons que l'on aurait aussi

$$\mathbb{P}[x_1, \cdots, x_n] = \mathbb{P}[x_n, x_{n-1}] \times \mathbb{P}[x_{n-2}|x_n, x_{n-1}] \times \cdots \times]x_1|x_n, \cdots, x_2].$$

puisque $\mathbb{P}[x_n, x_{n-1}] = \mathbb{P}[x_n] \times \mathbb{P}[x_{n-1}|x_n]$ mais aussi $\mathbb{P}[x_{n-1}] \times \mathbb{P}[x_n|x_{n-1}]$.

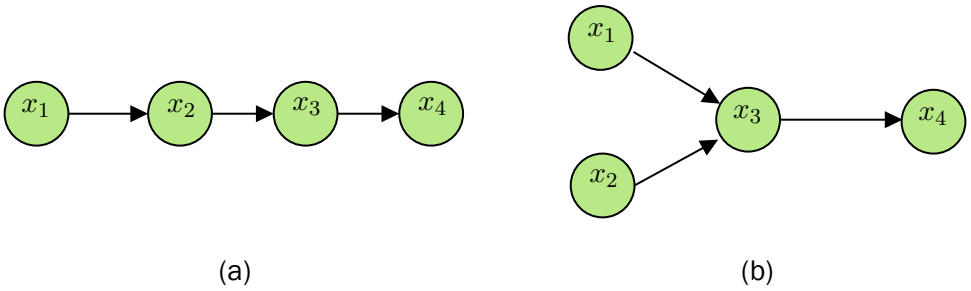


Figure 4.23 – Autres exemples de graphs dirigés.

L'idée ici sera d'écrire des probabilités conditionnelles impliquant uniquement les variables et leurs parents causaux. Par exemple, le graphe (a) de la Figure 4.23 correspondrait à

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

alors que le graphe (b) de la Figure 4.23 serait associé à l'écriture

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

qui s'écrirait aussi

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

car x_1 et x_2 sont supposées indépendantes. Il n'est pas rare de rajouter une hypothèse Markovienne, correspondant au cas où chaque variable est indépendante de tous ses ancêtres conditionnellement à ses parents. Par exemple, sur le graphe (a) de la Figure 4.23, l'hypothèse de Markov permet d'écrire

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

Aussi, le graphe (a) de la Figure 4.23 correspondrait à

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

alors que le graphe (b) de la Figure 4.23 serait associé à l'écriture

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

Pour aller plus loin dans les exemples, sur le schéma (a) de la Figure 4.21

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

de telle sorte que

$$\mathbb{P}[x_2, x_3|x_1] = \frac{\mathbb{P}[x_1, x_2, x_3]}{\mathbb{P}[x_1]} = \mathbb{P}[x_2|x_1] \cdot \mathbb{P}[x_3|x_1],$$

et donc $x_2 \perp\!\!\!\perp x_3$ conditionnellement à x_1. Sur le schéma (b) de la Figure 4.21

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

de telle sorte que

$$\mathbb{P}[x_1, x_3|x_2] = \frac{\mathbb{P}[x_1, x_2, x_3]}{\mathbb{P}[x_2]} = \frac{\mathbb{P}[x_1]\mathbb{P}[x_2|x_1]}{\mathbb{P}[x_2]}\mathbb{P}[x_3|x_2] = \mathbb{P}[x_1|x_2] \cdot \mathbb{P}[x_3|x_2],$$

et donc $x_1 \perp\!\!\!\perp x_3$ conditionnellement à x_2.

Sur le schéma (c) de la Figure 4.21

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

de telle sorte que

$$\mathbb{P}[x_1, x_2] = \mathbb{P}[x_1] \cdot \mathbb{P}[x_2],$$

autrement dit $x_1 \perp\!\!\!\perp x_2$, mais

$$\mathbb{P}[x_1, x_2|x_3] = \frac{\mathbb{P}[x_1, x_2, x_3]}{\mathbb{P}[x_3]} = \frac{\mathbb{P}[x_1] \cdot \mathbb{P}[x_2] \cdot \mathbb{P}[x_3|x_1, x_2]}{\mathbb{P}[x_3]},$$

et donc x_1 n'est pas indépendant de x_2 conditionnellement à x_3.

Compte tenu du lien entre les graphes causaux et ces modèles Markoviens, tester un modèle causal est très lié à des tests d'indépendance, ou plus précisément, des tests d'indépendance conditionnelle. Classiquement, tester l'indépendance entre deux variables se fait à l'aide de la corrélation (ou de la corrélation de rangs), et pour rappel, la corrélation entre x_1 et x_2, conditionnellement à x_3 s'écrit

$$r_{12|3} = \frac{r_{12} - r_{13} \times r_{23}}{\sqrt{(1 - r_{13}^2) \times (1 - r_{23}^2)}}.$$

On notera que le conditionnement se fait ici toujours sur un parent. Dans un réseau causal (appelé aussi DAG), chaque relation enfant / parent est représentée par une relation déterministe $x_2 = f_{1:2}(x_1, \varepsilon)$ où x_1 est le parent de x_2, et où ε est un bruit indépendant.

4.4.4 Introduction à l'inférence causale

Pearl et Mackenzie 2018 notaient que l'inférence causale visait à répondre à la question "*que se serait-il passé si...*" Cette question est centrale en épidémiologie ("*que se serait-il passé si cette personne avait reçu le traitement*") ou dès qu'on cherche à évaluer l'impact d'une politique publique ("*que se serait-il passé si on n'avait pas supprimé cette taxe*"). Mais on note que c'est la question que l'on se pose dès que l'on parle de discrimination ("*que se serait-il passé si cette femme avait été un homme*"). En inférence causale, pour quantifier l'effet d'un médicament, ou d'une mesure de politique publique, on constitue deux groupes, un qui aura le traitement, et un autre qui ne l'aura pas, et servira dès lors de contrefactuel, afin de répondre à la question *que se serait-il passé si la même personne avait accès au traitement*. Quand on analyse les discriminations, on se pose des questions similaires, par exemple *le prix du risque serait-il différent si la même personne avait été un homme et pas une femme*, sauf qu'ici le genre ne relève pas d'un choix, d'une affectation arbitraire à un traitement (aléatoire dans les expériences dite randomisées). C'est d'ailleurs un reproche qui avait été fait initialement sur ce parallèle entre l'analyse des discriminations et l'inférence causale : changer de traitement est possible, alors que changer de sexe est une vue de l'esprit. On peut d'ailleurs penser aux questions sur les liens entre le tabagisme et certains cancers : voir le fait de fumer comme un "traitement" peut avoir du sens mathématiquement, mais éthiquement, on ne pourrait pas forcer quelqu'un à fumer juste pour quantifier la probabilité d'avoir un cancer quelques années plus tard [52](alors que dans une expérience clinique, on pourrait

52. Dans un article humoristique, Smith et Pell 2003 posaient la question d'organiser des expériences randomisées pour prouver le lien causal entre le fait d'avoir un parachute et de survivre à un crash d'avion.

imaginer administrer à un patient des pilules bleues, au lieu des pilules rouges). On entre ici dans la catégorie des approches dites "*quasi-expériementales*", au sens de Cook *et al.* 2002 et DiNardo 2016. Dans l'échelle de causalité présentée par Pearl et Mackenzie 2018 (chapitre 1 et Figure 1.2), le premier niveau correspond à la recherche d'associations statistiques, ou de corrélations, comme le font régulièrement les actuaires. Le second niveau est celui de l'intervention, où des expériences sont organisées (comme pour quantifier l'effet d'un médicament), formellement noté à l'aide de l'opérateur $do(\cdot)$ dans la section 4.4.5, dans le sens où l'on cherche $Y|do(p)$. Dans l'analyse faite ici, on se place à un niveau plus abstrait, où une "intervention" est littéralement impossible (on ne changera pas le sexe d'un assuré, son origine ethnique, ou son âge). Ce troisième niveau est celui du contrefactuel, où on tente de s'imaginer ce qui se serait passé si l'assuré avait eu la caractéristique protégée p, ce que l'on notera $Y^\star_{P\leftarrow p}$ (ou plus simplement $Y^\star_p$).

Dans les données, y (souvent appelé "*outcome*", ou résultat) est la variable que l'on cherche à modéliser et prédire, et qui servira de mesure d'efficacité du traitement. Les résultats potentiels ("*potential outcome*") sont les résultats qui seraient observés sous chaque traitement possible, et on note $y^\star_t$ le résultat qui serait observé si le traitement T avait pris la valeur t. Et les résultats contrefactuels sont ce qui aurait été observé si le traitement avait été différent, autrement dit, pour une personne de type t, son résultat contrefactuel est $y^\star_{1-t}$ (car t prend les valeurs $\{0,1\}$). L'exemple classique est celui d'une personne qui a reçu un vaccin ($t=1$), qui n'est pas tombé malade ($y=0$), dont l'outcome contrefactuel serait $y^\star_0$, parfois noté $y^\star_{t\leftarrow 0}$. Avant de lancer l'étude sur l'efficacité du vaccin, les deux résultats sont potentiels, $y^\star_0$ et $y^\star_1$. Une fois l'étude lancée, le résultat observé sera y et le résultat contrefactuel sera $y^\star_{1-t}$.

Le traitement correspond formellement, dans notre exemple de vaccin, à une intervention [53]. Dans la plupart des expériences, il n'est toutefois pas possible de manipuler la variable dont on veut mesurer l'effet causal. En introduction, nous avions évoqué l'idée que l'indice de masse corporelle (IMC, ou BMI) pouvait avoir un impact sur l'état de santé, mais il est difficile de manipuler l'indice. On peut plutôt manipuler des variables qui auront un impact sur l'indice (en forçant une personne à pratiquer régulièrement le sport, à changer ses habitudes alimentaires, etc.) Ce qui fait qu'on ne mesure pas *stricto sensu* l'effet causal de l'indice de masse corporelle, mais davantage celui des interventions qui influencent l'indice. Et de la même manière, il est impossible d'intervenir sur certaines variables, dites immuables, comme le sexe ou l'origine raciale. Le contrefactuel est alors purement hypothétique. Dawid 2000 était très critique sur l'idée que nous puissions créer (ou observer) un contrefactuel, car "*by definition, we can never observe such [counterfactual] quantities, nor can we assess empirically the validity of any modeling assumption we may make about them, even though our conclusions may be sensitive to these assumption*".

On dira qu'il y a un effet causal (ou "effet causal identifié") de t sur y si $y^\star_0$ et $y^\star_1$ sont sensiblement différentes. Et comme on ne peut pas observer ces variables, au niveau individuel, on va comparer l'effet sur des sous-populations, comme le montrent Rubin 1974, Hernán et Robins 2010 ou Imai 2018. Assez naturellement, on pourrait vouloir mesurer l'effet causal comme la différence entre les $\overline{y}$ dans les deux groupes, celui qui a été traité ($t=1$) et celui qui n'a pas été traité ($t=0$), mais à moins de faire des hypothèses supplémentaires, cette différence ne correspondra pas à l'effet causal moyen (ACE, "*average causal effect*"),

$$\underbrace{\mathbb{E}[Y^\star_{t\leftarrow 1}-Y^\star_{t\leftarrow 0}]}_{\text{effet causal } ACE} \overset{?}{=} \underbrace{\mathbb{E}[Y|T=1]-\mathbb{E}[Y|T=0]}_{\text{mesure d'association}}.$$

53. Pour faire le lien avec ce que nous évoquions auparavant, le traitement t deviendra la variable protégée p quand nous reviendrons à l'analyse de l'équité et de la discrimination.

Le problème ici est que T et Y peuvent être, a priori, corrélées. Par exemple pour les vaccins, les personnes “à risque” peuvent vouloir davantage être vaccinées. Plusieurs conditions seront nécessaires pour identifier un effet causal,

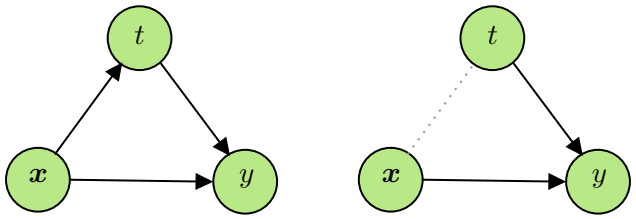
$$\begin{cases} \text{consistance} & : \ y_t^\star = y \text{ si } T = t, \text{ pour tout } t \\ \text{échangeabilité} & : \ Y_t^\star \perp\!\!\!\perp T \mid \boldsymbol{X} = \boldsymbol{x}, \text{ pour tout } t \\ \text{positivité} & : \ \mathbb{P}[T = t | \boldsymbol{X} = \boldsymbol{x}] > 0, \text{ pour tout } t, \end{cases}$$

La motivation de ces conditions est présentée en détails dans Hernán et Robins 2010. La seconde condition, qui dit que l'outcome contrefactuel est indépendant du traitement, signifie que si deux personnes rigoureusement identiques ont respectivement reçu et pas reçu le traitement, l'outcome de l'un sera le contrefactuel de l'autre. Toutefois, on ne demande ici qu'une égalité en loi des outcomes, par une égalité stricte. Avec ces hypothèses,

$$\mathbb{E}[Y_{T\leftarrow t}^\star | \boldsymbol{X} = \boldsymbol{x}] \overset{\text{échg.}}{=} \mathbb{E}[Y_{T\leftarrow t}^\star | T = t, \boldsymbol{X} = \boldsymbol{x}] \overset{\text{const.}}{=} \mathbb{E}[Y | T = t], \ \forall t,$$

où la première égalité est validée par la propriété d'échangeabilité, la seconde par la propriété de consistance, et le dernier terme existant toujours par la propriété de positivité.

Le principal soucis des études observationnelles est que T et $\boldsymbol{X}$ peuvent être non-indépendants (comme sur la Figure 4.24). Gelman et J. Hill 2006 distinguent deux types d'inadéquation (entre les groupes de traitement et de contrôle), avec l'absence de chevauchement complet sur les prédicteurs observés avant le traitement, et le déséquilibre des prédicteurs observés avant le traitement.



étude observationnelle étude randomisée

Figure 4.24 – Étude observationnelle ou étude randomisée.

Stratification

Comme on vient de le montrer, moyennant quelques hypothèses, $\mathbb{E}[Y_{T\leftarrow t}^\star | \boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[Y | T = t]$, on peut mesurer l'effet causal marginal en calculant des moyennes, conditionnement aux $\boldsymbol{x}$. Ainsi,

$$\mathbb{E}[Y_{T\leftarrow t}^\star] = \sum_{\boldsymbol{x}} \mathbb{E}[Y | T = t, \boldsymbol{X} = \boldsymbol{x}] \cdot \mathbb{P}[\boldsymbol{X} = \boldsymbol{x}].$$

Appariement (matching)

L'approche par appariement consiste à apparier chaque individu (à partir de ses covariables $\boldsymbol{x}$) du groupe traité à un individu du groupe non-traité, qui lui ressemble. Ainsi, la distribution des

covariables / facteurs de confusion dans la population contrôle sera la même que dans la population traitée, grâce à cet appariement, et on pourra alors calculer un effet moyen du traitement chez les traités. D'un point de vue pratique, un algorithme glouton peut être utilisé.

- on permute la population des individus traités ($t_i = 1$),
- pour l'individu traité i, on cherche l'individu non-traité dans la base, au sens $j_i^\star = \underset{j:t_j=0}{\text{argmin}}\{d(\boldsymbol{x}_i, \boldsymbol{x}_j)\}$,
- on enlève l'observation non-traitée de la base, et on itère (de manière à apparier tous les individus traités avec une personne non-traitée

On estime ainsi $\mathbb{E}[Y^\star_{T\leftarrow 1}] - \mathbb{E}[Y^\star_{T\leftarrow 0}]$ par

$$ACE = \overline{y}^\star_{T\leftarrow 1} - \overline{y}^\star_{T\leftarrow 0} = \frac{1}{n_1}\sum_{i:t_i=1} y_i - y_{j_i^\star}, \text{ où } j_i^\star = \underset{j:t_j=0}{\text{argmin}}\{d(\boldsymbol{x}_i, \boldsymbol{x}_j)\},$$

et où n_1 est le nombre d'observations traitées. Afin de s'assurer de la robustesse de l'analyse, et il est classique de permuter plusieurs fois les données, mais aussi de retirer des observations s'il n'y a plus personne avec qui apparier (on parle de "caliper" les données à un seuil ϵ si $\min\{d(\boldsymbol{x}_i, \boldsymbol{x}_j)\} > \epsilon$). Une fois les données appariées, on peut classiquement faire un test de Student (apparié) pour mesurer un effet causal si l'outcome y est continu, ou un test du chi-deux si y est binaire (comme suggéré dans McNemar 1947).

Score de propension

Pour l'individu i, compte tenu de ses caractéristiques $\boldsymbol{x}_i$, sa probabilité d'avoir été traité est notée $\pi_i = \mathbb{P}[T = 1|\boldsymbol{X} = \boldsymbol{x}_i]$. Il est possible de faire un appariement sur la base du score de propension, même deux scores identiques ne signifient aucunement que les deux individus sont proches, mais cela permet d'équilibrer les deux groupes, comme le soulignent Colson *et al.* 2016. Le score de propension peut s'estimer par une régression logistique, par exemple, et on a alors $\pi_i \in [0, 1]$.

- on estime tous les scores de propension $\pi_i = \mathbb{P}[T = 1|\boldsymbol{X} = \boldsymbol{x}_i]$, ou les logarithmes de cotes c_i, i.e. (avec un modèle logistique)

$$\log \widehat{c}_i = \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}} \text{ où } \widehat{c}_i = \frac{\widehat{\pi}_i}{1 - \widehat{\pi}_i},$$

- on calcule l'écart-type des $\log \widehat{c}_i$, noté $\widehat{\sigma}$, et on utilise comme "caliper" $\epsilon = k\widehat{\sigma}$ (avec $k = 20\,\%$ par exemple)
- on lance une procédure d'appariement sur la base permutée : pour l'individu traité i, on cherche l'individu non-traité dans la base (au sens $j_i^\star = \underset{j:t_j=0}{\text{argmin}}\{|\log \widehat{c}_i - \log \widehat{c}_j|\}$).

Là encore, on estime $\mathbb{E}[Y^\star_{T\leftarrow 1}] - \mathbb{E}[Y^\star_{T\leftarrow 0}]$ par

$$ACE = \overline{y}^\star_{T\leftarrow 1} - \overline{y}^\star_{T\leftarrow 0} = \frac{1}{n_1}\sum_{i:t_i=1} y_i - y_{j_i^\star}, \text{ où } j_i^\star = \underset{j:t_j=0}{\text{argmin}}\{d(\boldsymbol{x}_i, \boldsymbol{x}_j)\}.$$

Pondération

Une autre approche consiste à utiliser l'inverse de la probabilité de traitement,

$$\begin{cases} \text{traités } (t_i = 1) & : \text{ pondération par l'inverse de la probabilité d'être traité } \pi_i \\ \text{non-traités } (t_j = 0) & : \text{ pondération par l'inverse de la probabilité d'être non-traité } 1 - \pi_j, \end{cases}$$

et on estime ainsi $\mathbb{E}[Y^\star_{T\leftarrow 1}]$ (par exemple) par

$$\overline{y}^\star_{T\leftarrow 1} = \sum_{i:t_i=1} \omega_i y_i \text{ où } \omega_i = \frac{\pi_i^{-1}}{\sum_{j:t_j=1}\pi_j^{-1}}.$$

On reconnaît ici l'estimateur d'Horvitz-Thompson utilisé pour corriger d'un possible biais, introduit par Horvitz et Thompson 1952.

Ces poids permettent d'aller plus loin, par exemple en considérant des modèles de régression (on parlera de "*marginal structural model*") de la forme $y^\star_{T\leftarrow t} = \beta_0 + \beta_1 t + \varepsilon$, de telle sorte que β_1 sera l'effet causal moyen. On peut aussi considérer une régression logistique si y est binaire (ou tout autre modèle linéaire généralisé suivant la distribution possible pour y). On fait une régression sur la base de nos données observées (y_i, t_i), pondérées par π_i^{-1} si $t_i = 1$ et $(1-\pi_i)^{-1}$ si $t_i = 0$.

Contrefactuels et transport optimal

En allant plus loin, ayant observé $(\boldsymbol{x}, p, y)$, on espère créer un contrefactuel en considérant $(\boldsymbol{x}, p^\star, y^\star)$, où $p^\star = 1 - p$ (si on avait observé $p = 0$, on aurait souhaité créer une observation de classe protégée $p = 1$, et inversement). Mais si p change, il y a des chances pour que $\boldsymbol{x}$ change aussi, comme le rappellent Gordaliza *et al.* 2019, Black *et al.* 2020, Torous *et al.* 2021 ou Lara *et al.* 2021. Formellement, on dit simplement que les distributions de $\boldsymbol{x}$, conditionnellement à $p = 0$ ou $p = 1$, ne sont pas identiques, ce qui arrivera forcément en présence d'un proxy de p parmi les variables explicatives $\boldsymbol{x}$. Pour s'en convaincre, supposons que $\boldsymbol{x}$ soit la taille d'une personne et p son genre. Si on observe un homme de 190 cm, le contrefactuel n'est a priori pas une femme de 190cm. Comme on le voit sur la Figure 4.25, l'idée naturelle serait de dire que le contrefactuel d'un homme de 190cm serait probablement une femme de 175cm, car ces deux grandeurs correspondent au même niveau de quantile (de l'ordre de 96 %).

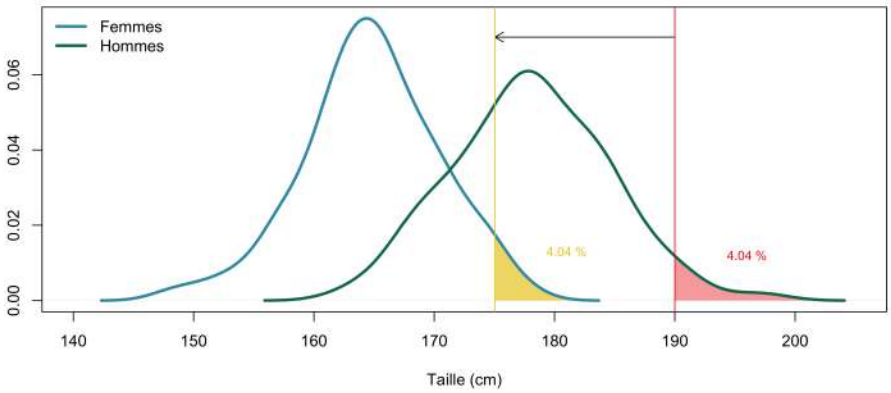


Figure 4.25 – Transport optimal et correspondance des tailles (hommes et femmes), sur la base de données de tailles d'élèves (source : Fox 1997).

L'extension à un ensemble de variables $\boldsymbol{x} = (x_1, \cdots, x_k)$ se fait en utilisant un algorithme de transport optimal (voir Villani 2009 ou Galichon 2016 par exemple). L'idée est de noter que dans les deux groupes, $p = 0$ et $p = 1$, les caractéristiques $\boldsymbol{x}$ ont (potentiellement) deux distributions

différentes, $\mathbb{P}_0$ (ou $\mathbb{P}$) et $\mathbb{P}_1$ (ou $\mathbb{Q}$). Formellement, si $\mathbb{P}$ est une distribution sur $\mathbb{R}^k$, étant donné $T : \mathbb{R}^k \to \mathbb{R}^k$ et on définit la mesure "*push-forward*" associée, $\mathbb{Q}$, définie par

$$\mathbb{Q}(A) = T_\# \mathbb{P}(A) = \mathbb{P}(T^{-1}(A)), \ \forall A \subset \mathbb{R}^k.$$

Un transport optimal $T^\star$ (au sens de Brenier) de $\mathbb{P}$ vers $\mathbb{Q}$ sera une solution du problème de Monge

$$T^\star \in \underset{T:T_\#\mathbb{P}=\mathbb{Q}}{\text{arginf}} \int_{\mathbb{R}^k} \|\boldsymbol{x} - T(\boldsymbol{x})\|^2 d\mathbb{P}(\boldsymbol{x}).$$

Il est possible de montrer (voir Villani 2009 ou Galichon 2016) que $T^\star = \text{grad}(\psi)$ où ψ est une fonction convexe. Si $k = 1$ (comme dans notre exemple précédent avec la taille) T est alors une fonction croissante (en tant que dérivée d'une fonction convexe), et si $F_0(x) = \mathbb{P}_0[X \leq x]$ et $F_1(x) = \mathbb{P}_1[X \leq x]$, alors $T^\star(x) = F_1^{-1} \circ F_0(x)$ vérifie $T^\star_\# \mathbb{P}_0 = \mathbb{P}_1$ (car $F_1(x) = F_0(T^{\star -1}(x))$ et $T^\star$ sera optimale au sens de Brenier.

L'idée naturelle de Gordaliza *et al.* 2019 et Black *et al.* 2020 est de dire qu'un modèle m, construit à partir d'un score s (avec $m(\boldsymbol{x}, p) = \mathbf{1}(s(\boldsymbol{x}, p) > \text{seuil})$) sera non équitable si $m(\boldsymbol{x}, 0) \neq m(T^\star(\boldsymbol{x}), 1)$. Black *et al.* 2020 définissent ainsi l'ensemble "*FlipSet*" comme

$$\mathcal{X}_F(m, T^\star) = \{\boldsymbol{x} \in \mathcal{X} : m(\boldsymbol{x}, 0) \neq m(T^\star(\boldsymbol{x}), 1)\}.$$

On peut aller plus loin, on considérant la partition suivante, en deux sous-ensembles

$$\left\{\mathcal{X}_F^-(m, T^\star) = \{\boldsymbol{x} \in \mathcal{X} : m(\boldsymbol{x}, 0) < m(T^\star(\boldsymbol{x}), 1)\}.\right.$$

et si $T^\star_\# \mathbb{P}_0 = \mathbb{P}_1$, alors l'effet causal moyen vaut

$$ACE = \mathbb{E}[Y^\star_{T \leftarrow 1}] - \mathbb{E}[Y^\star_{T \leftarrow 0}] = \mathbb{P}_0[\mathcal{X}_F^-(m, T^\star)] - \mathbb{P}_0[\mathcal{X}_F^+(m, T^\star)]$$

Exemple de mesure d'effet causal

On s'intéresse à un modèle causal de la forme de celui de la Figure 4.27. On suppose ici que

$$\begin{cases} x_1 = \alpha p + \gamma u_2 \\ y \ = \beta x_1 + \delta u_3 \end{cases}$$

où les variables (u_2, u_3) sont exogènes et indépendantes, non-observées.

On a une observation i, associée à Bob, telle que $p_i = 0$, associée à $x_{1,i}$ et $\widehat{y}_i$. Que se serait-il passé si Bob avait été une femme ($p_i^\star = 1$)? On cherche ici $\widehat{y}_{i,P \leftarrow 1}$. On peut procéder en 3 étapes.

— abduction : on va retrouver les caractéristiques non-observées $(u_{2,i}, u_{3,i})$, en utilisant

$$u_{2,i} = \frac{1}{\gamma}(x_{1,i} - \alpha p_i) \text{ et } u_{3,i} = \frac{1}{\delta}(y_i - \beta x_{1,i})$$

— dans un second temps, on remplace p_i par $p_i^\star$, en supposant $(x_{1,i}, u_{2,i}, u_{3,i})$ inchangé,

$$\begin{cases} x_{1,P \leftarrow 1} = \alpha \cdot 1 + \gamma u_2 = \alpha \cdot 1 + (x_1 - \alpha \cdot 0) = \alpha + x_1 \\ \widehat{y}_{P \leftarrow 1} = \beta x_{1,P \leftarrow 1} + \delta u_3 = \beta(\alpha + x_1) + (y - \beta x_1) \end{cases}$$

soit

$$\widehat{y}_{P \leftarrow 1} = \beta \alpha + y$$

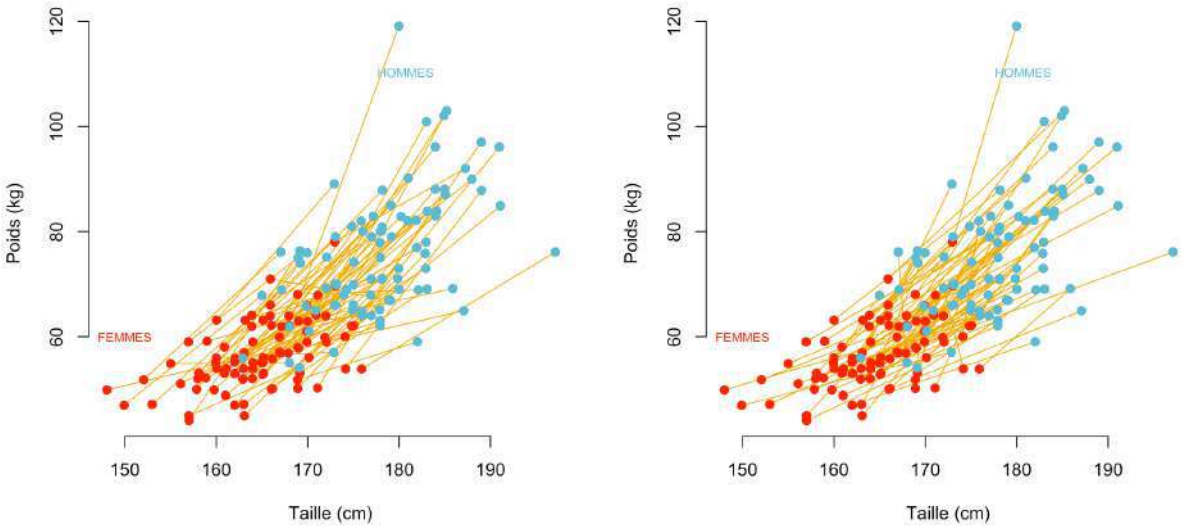


Figure 4.26 – Deux “appariements optimaux” (entre hommes et femmes), basés sur les données de taille et de poids des élèves, sur la base d’un appariement univarié à gauche (sur la taille seulement), et d’un appariement bivarié à droite (source : Fox 1997).

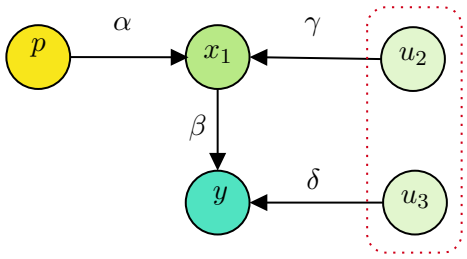


Figure 4.27 – Exemple de réseau causal simple.

4.4.5 Les modèles causaux structurels

Formellement, un modèle causal structurel est un triplet $(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{f})$ (dans Pearl 2010 ou Halpern 2016). Les variables dans $\boldsymbol{U}$ sont appelées variables exogènes, autrement dit, elles sont externes au modèle (nous n’avons pas à expliquer la cause). Les variables dans $\boldsymbol{V}$ sont appelées endogènes. Chaque variable endogène est un descendant d’au moins une variable exogène. Les variables exogènes ne peuvent être descendantes d’aucune autre variable et, en particulier, ne peuvent pas être descendantes d’une variable endogène. Aussi, elles n’ont pas d’ancêtres et sont représentées comme des racines dans les graphes causaux. Enfin, si on connaît la valeur de chaque variable exogène, on peut, à l’aide des fonctions de $\boldsymbol{f}$, déterminer avec une parfaite certitude la valeur de chaque variable endogène. Les graphes causaux que nous avons décrit sont constitués d’un ensemble de nœuds représentant les variables dans $\boldsymbol{U}$ et $\boldsymbol{V}$, et d’un ensemble d’arêtes entre les nœuds représentant les fonctions dans $\boldsymbol{f}$.

Sur le diagramme causal (a) de la Figure 4.28, on dispose de deux variables endogènes, x et y, et de deux variables exogènes, u_x et u_y. Le diagramme (a) est une représentation du monde réel, mais on suppose ici qu’il est possible de faire des interventions, et de changer la valeur de x, en supposant que toutes choses restent égales par ailleurs.

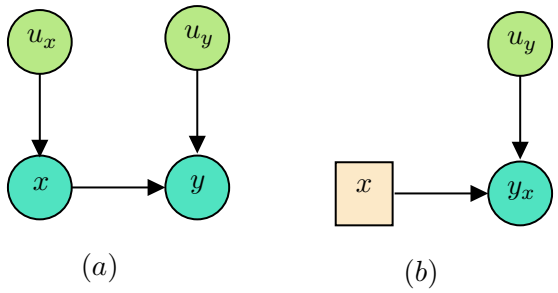


Figure 4.28 – Diagramme causal, avec une intervention sur x, à droite.

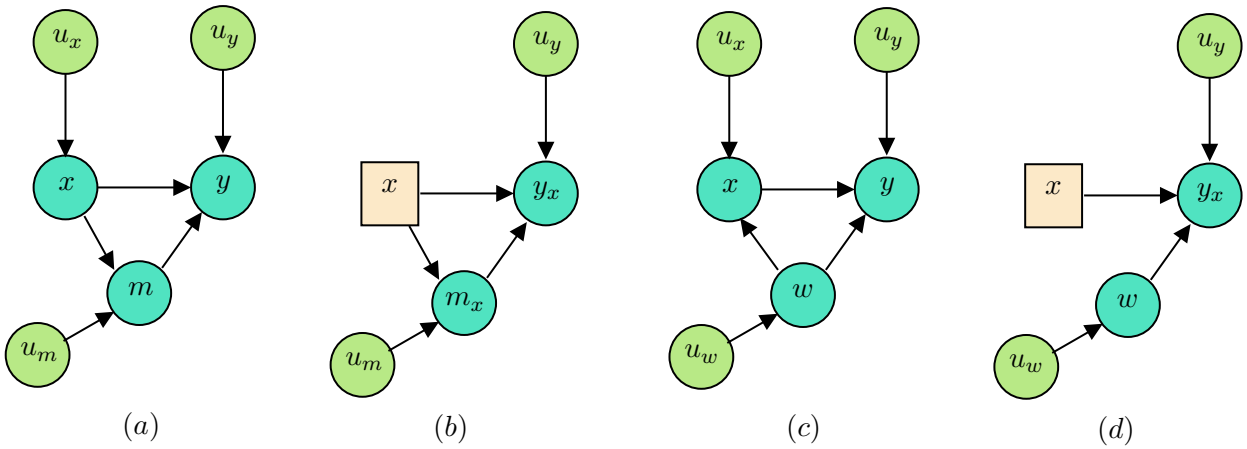


Figure 4.29 – Diagrammes causaux, $x \to y$, avec un un médiateur m à gauche, et avec un facteur de confusion w à droite.

monde réel	avec intervention $(do(x))$
$\begin{cases} X = f_x(U_x) \\ Y = f_y(X, U_y) \end{cases}$	$\begin{cases} X = x \\ Y_x = f_y(x, U_y) \end{cases}$

Dans ce cas, on peut quantifier l'effet causal puisque

$$ACE = \mathbb{P}[Y = 1|do(X = 1)] - \mathbb{P}[Y = 1|do(X = 0)] = \mathbb{P}[Y_1 = 1] - \mathbb{P}[Y_0 = 1],$$

que l'on peut réécrire

$$ACE = \mathbb{P}[Y = 1|X = 1] - \mathbb{P}[Y = 1|X = 0] = \mu_1 - \mu_0,$$

où $\mu_x = \mathbb{P}[Y = 1|X = x]$ s'estime simplement par des fréquences par groupe.

Sur le diagramme causal (a) de la Figure 4.29, on dispose de trois variables endogènes, x, y, et un médiateur m, et de trois variables exogènes, u_x, u_y et u_m. Le diagramme (a) est une représentation du monde réel, mais comme auparavant, on suppose ici qu'il est possible de faire des interventions sur X.

Autrement dit :

en présence d'un médiateur (m)

$$\begin{array}{ll} \text{monde réel} & \text{avec intervention } (do(x)) \\ \begin{cases} X = f_x(U_x) \\ M = f_m(X, U_m) \\ Y = f_y(X, M, U_y) \end{cases} & \begin{cases} X = x \\ M_x = f_m(x, U_m) \\ Y_x = f_y(x, M_x, U_y) \end{cases} \end{array}$$

Sur le diagramme causal (b) de la Figure 4.29, on dispose de trois variables endogènes, x, y, et un facteur de confusion w, et de trois variables exogènes, u_x, u_y et u_w. Le diagramme (c) est une représentation du monde réel, mais comme auparavant, on suppose ici qu'il est possible de faire des interventions sur X. Autrement dit,

en présence d'un facteur de confusion (w)

$$\begin{array}{ll} \text{monde réel} & \text{avec intervention } (do(x)) \\ \begin{cases} X = f_x(W, U_x) \\ W = f_w(U_w) \\ Y = f_y(X, W, U_y) \end{cases} & \begin{cases} X = x \\ W = f_w(U_w) \\ Y_x = f_y(x, W, U_y) \end{cases} \end{array}$$

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

En fait, en présence d'un facteur de confusion

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

Par exemple, on peut supposer que $\mathbb{P}[Y = 1 | W = w, X = x]$ est obtenu à l'aide d'un modèle logistique : si $\mu_x(w) = \mathbb{P}[Y = 1 | W = w, X = x]$

$$\widehat{\mu}_x(w) = \frac{\exp\left[\widehat{\beta}_0 + \widehat{\beta}_x x + \widehat{\beta}_w w\right]}{1 + \exp\left[\widehat{\beta}_0 + \widehat{\beta}_x x + \widehat{\beta}_w w\right]},$$

l'effet causal moyen, $ACE = \mathbb{E}(\mu_1(W) - \mu_0(W))$ sera estimé par

$$\widehat{ACE} = \frac{1}{n} \sum_{i=1}^{n} (\widehat{\mu}_1(w_i) - \widehat{\mu}_0(w_i)).$$

Sur le diagramme causal (a) de la Figure 4.30, $\boldsymbol{U} = \{x, \varepsilon\}$ et $\boldsymbol{V} = \{y\}$, de telle sorte que $y = f(x, \varepsilon)$, ou $y = f_y(x, \varepsilon)$. Ici, x représente le parent de y, et ε les erreurs structurelles dues aux facteurs non-modélisés qui engendreront de l'incertitude. Comme l'explique Pearl 1998, l'équation structurelle $y = f_y(x, \varepsilon)$ représente un mécanisme causal qui spécifie la valeur prise par la variable Y en réponse à chaque couple de valeurs prises par la variable x et les facteurs ε. $y = f_y(x, \varepsilon)$ est calculé sous l'intervention formelle "X est fixé à x", ce que Pearl 1998 note "$do(X = x)$" (ou simplement $do(x)$), puis affecté à Y (historiquement, de Wright 1921 à Holland 1986, en passant par Neyman *et al.* 1923 ou Rubin 1974, diverses notations ont été proposées). Pour reprendre

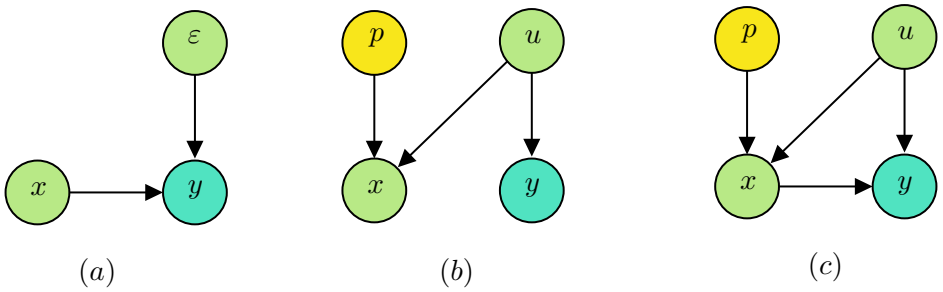


Figure 4.30 – Diagrammes causaux, avec une variable protégée p, une variable explicative x et un outcome y.

la notation et l'interprétation de Pearl 2010, "Y *would be* y *had* X *been* x *in situation* ε" s'écrira $Y_x(\varepsilon) = y$, l'erreur structurelle ε n'étant pas impactée par une intervention sur x.

Dans la terminologie probabiliste, $\mathbb{P}(Y = y|X = x)$ désigne la distribution de population de Y parmi les individus dont la valeur X est x. Ici, $\mathbb{P}(Y = y|do(X = x))$ représente la distribution de Y dans la population si tous les individus de la population avaient leur valeur X fixée à x. Et plus généralement, $\mathbb{P}(Y = y|do(X = x), Z = z)$ va désigner la probabilité conditionnelle que $Y = y$, étant donné que $Z = z$, dans la distribution créée par l'intervention $do(X = x)$. Aussi, dans la littérature, l'effet causal moyen ("*average causal effect*", ACE) correspond à $\mathbb{E}[Y|do(X = 1)] - \mathbb{E}[Y|do(X = 0)]$, ou $\widehat{Y}_1 - \widehat{Y}_0$ si $\widehat{Y}_x = \mathbb{E}[Y|do(X = x)]$ (que l'on notera par la suite aussi $Y^\star_{X\leftarrow 1} - Y^\star_{X\leftarrow 0}$). Pour calculer cette grandeur, étant donné un graphe causal,

$$\mathbb{P}[Y = y|do(X = x)] = \sum_z \mathbb{P}[Y = y|X = x, PA = z] \cdot \mathbb{P}[PA = z],$$

où PA désigne les parents de x, et où z couvre toutes les combinaisons de valeurs que les variables de PA peuvent prendre. Une condition suffisante pour identifier l'effet causal $\mathbb{P}(y|do(x))$ est que chaque chemin entre X et l'un de ses enfants trace au moins une flèche émanant de la variable mesurée (Tian et Pearl 2002).

4.4.6 Équité contrefactuelle

Sur la Figure 4.30, on peut visualiser deux diagrammes causaux, où y est notre variable d'intérêt, p la variable protégée, x une "variable explicative" (potentiellement) et u une caractéristique non-observée. Sur le diagramme (b), p ne cause pas y car il n'existe aucune chemin causal où p causerait y, autrement dit y n'est jamais un parent (même lointain) de p. Sur le diagramme (c), cette fois, p cause (indirectement) y (via le chemin $p \to x \to y$).

Un modèle vérifie la propriété d'équité contrefactuelle ("*counterfactual fairness*") si "*had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same*". Aussi, un classifieur sera contrefactuellement juste si, pour tous les individus, le résultat est égal au résultat de son individu contrefactuel (c'est-à-dire le même individu avec un attribut protégé inversé),

$$\mathbb{P}[Y^\star_{P\leftarrow p} = Y|\boldsymbol{X} = \boldsymbol{x}] = \mathbb{P}[Y^\star_{P\leftarrow q} = Y|\boldsymbol{X} = \boldsymbol{x}], \ \forall p, q,$$

où $Y^\star_{P\leftarrow z}$ est la prévision du classifieur si p prenait la valeur z.

Équité contrefactuelle (Kusner *et al.* 2017)

Si la prédiction dans le monde réel est la même que celle dans le monde contrefactuel où l'individu aurait appartenu à un groupe démographique différent, on a une équité contrefactuelle, autrement dit

$$\mathbb{P}[Y^{\star}_{P\leftarrow p} = y | \boldsymbol{X} = \boldsymbol{x}, P = p] = \mathbb{P}[Y^{\star}_{P\leftarrow p'} = y | \boldsymbol{X} = \boldsymbol{x}, P = p],\ \forall p', \boldsymbol{x}, y.$$

fairness through awareness	Dwork *et al.* 2012	$D(\widehat{y}_i, \widehat{y}_j) \leq d(\boldsymbol{x}_i, \boldsymbol{x}_j),\ \forall i, j$
counterfactual fairness	Kusner *et al.* 2017	$\mathbb{P}[Y^{\star}_{P\leftarrow p} = y \vert \boldsymbol{X} = \boldsymbol{x}, P = p] = \text{cst}_y,\ \forall p$
no proxy discrimination	Kilbertus *et al.* 2017	$\mathbb{P}[\widehat{Y} = y \vert do(P = p)] = \text{cst}_y,\ \forall p$

Table 4.6 – Définitions de l'équité individuelle.

4.5 Corriger une inéquité

4.5.1 Approches de prétraitement (*pre-processing*)

Une approche directe pour éliminer les biais des ensembles de données consisterait à supprimer l'attribut protégé et les autres éléments des données qui sont soupçonnés de contenir des informations connexes. Malheureusement, une telle suppression est rarement suffisante. Il existe souvent des corrélations subtiles dans les données qui signifient que l'attribut protégé peut être reconstruit. Par exemple, nous pouvons supprimer la race, mais conserver les informations relatives à l'adresse du sujet qui peuvent être fortement corrélées à la race.

La mesure dans laquelle il existe des dépendances entre les données $\boldsymbol{x}$ et l'attribut protégé p peut être calculée à l'aide de l'information mutuelle

$$LP = \sum_{\boldsymbol{x},p} \mathbb{P}(\boldsymbol{x}, p) \log \frac{\mathbb{P}(\boldsymbol{x}, p)}{\mathbb{P}(\boldsymbol{x})\mathbb{P}(p)},$$

que Kamishima, Akaho et Sakuma 2011 appellent le "préjudice latent". Plus cette mesure augmente, plus l'attribut protégé devient prévisible à partir des données. En effet, Feldman *et al.* 2015 et Menon et Williamson 2018 ont montré que la prévisibilité de l'attribut protégé met des limites mathématiques à la discrimination potentielle d'un classifieur.

Dans la littérature, quatre approches permettent de supprimer les biais en manipulant l'ensemble des données. Respectivement, ces approches modifient les étiquettes y, les données observées $\boldsymbol{x}$, les paires données / étiquettes $\{\boldsymbol{x}, y\}$ et la pondération de ces paires.

Manipulation des étiquettes

Kamiran et Calders 2009 puis Kamiran et Calders 2012 ont proposé de modifier certaines des étiquettes d'entraînement, ce qu'ils appellent la manipulation des données. Ils calculent un classifieur sur le jeu de données original et trouvent des exemples proches de la surface de décision. Ils échangent ensuite les étiquettes de manière à ce qu'un résultat positif pour le groupe défavorisé soit plus probable et refont la formation. Il s'agit d'une approche heuristique qui améliore empiriquement l'équité au détriment de la précision.

Manipulation des données observées

Feldman *et al.* 2015 ont proposé de manipuler les dimensions individuelles des données $\boldsymbol{x}$ d'une manière qui dépend de l'attribut protégé p. L'idée est d'aligner les distributions cumulatives $F_0[\boldsymbol{x}]$ et $F_1[\boldsymbol{x}]$ pour la caractéristique $\boldsymbol{x}$ lorsque l'attribut protégé p est respectivement 0 et 1, sur une distribution cumulative médiane $F_m[\boldsymbol{x}]$. Cette méthode est similaire à la normalisation des notes d'examen dans différents établissements scolaires, et est appelée "suppression de l'impact disparate". Cette approche présente l'inconvénient de traiter chaque variable d'entrée $\boldsymbol{x} \in \mathcal{X}$ séparément et ignore leurs (possibles) interactions.

Manipulation des étiquettes et des données

Calmon *et al.* 2017 proposent de considérer une transformation ψ qui transformera les paires de données $\{\boldsymbol{x}, y\}$ en nouvelles valeurs de données $\{\boldsymbol{x}', y'\}$ d'une manière qui dépend explicitement de l'attribut protégé p. Calmon *et al.* 2017 formulent ce problème comme un problème d'optimisation dans lequel il faut minimiser le changement d'utilité des données, sous réserve de limites sur le préjudice et la déformation des valeurs originales. Contrairement à la suppression des impacts disparates, cette méthode prend en compte les interactions entre toutes les dimensions des données. Cependant, la transformation randomisée est formulée sous la forme d'une table de probabilité, ce qui ne convient que pour les ensembles de données comportant un petit nombre de variables d'entrée et de sortie discrètes.

Repondération des paires de données

Kamiran et Calders 2012 proposent de repondérer les observations $\{\boldsymbol{x}, y\}$ dans l'ensemble de données d'apprentissage afin que les cas où l'attribut protégé p prédit que le groupe défavorisé obtiendra un résultat positif soient plus fortement pondérés. Ils forment ensuite un classifieur qui utilise ces pondérations dans sa fonction de coût. Ils proposent également de rééchantillonner les données d'apprentissage en fonction de ces pondérations et d'utiliser un classifieur standard.

4.5.2 Algorithmes de retraitement

Dans la section précédente, nous avons introduit la mesure de préjudice latent basée sur l'information mutuelle entre les données $\boldsymbol{x}$ et l'attribut protégé p. De même, nous pouvons mesurer la dépendance entre les étiquettes y et l'attribut protégé p

$$IP = \sum_{y,p} \mathbb{P}(y,p) \log \frac{\mathbb{P}(y,p)}{\mathbb{P}(y)\mathbb{P}(p)}$$

que Kamishima, Akaho et Sakuma 2011 appellent le "préjudice indirect". Intuitivement, s'il n'y a aucun moyen de prédire les étiquettes à partir de l'attribut protégé et vice versa, alors il n'y a pas de possibilité de biais.

Une approche pour éliminer les biais pendant la formation consiste à supprimer explicitement cette dépendance en utilisant l'apprentissage contradictoire. D'autres approches consistent à pénaliser l'information mutuelle en utilisant la régularisation, en ajustant le modèle sous la contrainte qu'il ne soit pas biaisé. Nous allons aborder brièvement chacune de ces approches.

Dé-biaisage "*adversarial*"

Le dé-biaisage "*adversarial*" (présenté dans Beutel, Chen, Zhao *et al.* 2017 ou Zhang, Lemoine *et al.* 2018) réduit l'impact de l'attribut protégé dans les prédictions en essayant de tromper simultanément un deuxième classifieur qui tente de deviner l'attribut protégé p. Beutel, Chen, Zhao *et al.* 2017 forcent les deux classifieurs à utiliser une représentation partagée et ainsi, minimiser la performance du classifieur adverse signifie supprimer toute information sur l'attribut protégé de cette représentation. Beutel, Chen, Zhao *et al.* 2017 proposent une représentation pour la classification qui était également utilisée pour prédire l'attribut protégé. Le système a été formé de manière "*adversariale*", encourageant les bonnes performances du système mais punissant la classification correcte de l'attribut protégé. De cette façon, une représentation qui ne contient pas d'informations sur l'attribut protégé est apprise.

Suppression des préjugés par régularisation

Kamishima, Akaho et Sakuma 2011 ont proposé d'ajouter une condition de régularisation supplémentaire à la sortie du classifieur de régression logistique qui tente de minimiser l'information mutuelle entre l'attribut protégé et la prédiction $\widehat{y}$. Ils ont d'abord réorganisé l'expression du préjudice indirect en utilisant la définition de la probabilité conditionnelle pour quantifier le potentiel de discrimination,

$$PI(\boldsymbol{x}) = \sum_{y,p} \mathbb{P}(Y = y|\boldsymbol{X} = \boldsymbol{x}, P = p) \log \frac{\mathbb{P}(y,p)}{\mathbb{P}(y)\mathbb{P}(p)} = \sum_{y,p} \mathbb{P}(y|\boldsymbol{x}, p) \log \frac{\mathbb{P}(y|p)}{\mathbb{P}(y)}$$

Ensuite, ils formulent une perte de régularisation basée sur l'espérance de celle-ci sur l'ensemble des données :

$$L_{\text{reg}} = \sum_{i=1}^{n} \sum_{\widehat{y},p} \mathbb{P}(\widehat{y}_i|\boldsymbol{x}_i, p_i) \log \frac{\mathbb{P}(\widehat{y}_i|p_i)}{\mathbb{P}(\widehat{y}_i)}$$

où la somme se fait sur l'ensemble des données qu'ils ajoutent à la perte d'apprentissage principale.

Références

Aaronson, D., D. A. Hartley et B. Mazumder (2017). « The Effects of the 1930s HOLC “Redlining” Maps ». In : *Federal Reserve Bank of Chicago Working Paper* 2017-12.

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

Adams, S. J. (2004). « Age discrimination legislation and the employment of older workers ». In : *Labour Economics* 11.2, p. 219-241.

Adomavicius, G. et A. Tuzhilin (2005). « Personalization technologies : a process-oriented perspective ». In : *Communications of the ACM* 48.10, p. 83-90.

Agarwal, S. et S. Mishra (2021). *Responsible AI : Implementing Ethical and Unbiased Algorithms*. Springer.

Agarwal, S. (2021). « Trade-Offs between Fairness and Interpretability in Machine Learning ». In : *IJCAI 2021 Workshop on AI for Social Good*.

Ahmed, A. M., L. Andersson et M. Hammarstedt (2013). « Are gay men and lesbians discriminated against in the hiring process ? » In : *Southern Economic Journal* 79.3, p. 565-585.

Aigner, D. J. et G. G. Cain (1977). « Statistical theories of discrimination in labor markets ». In : *Industrial and Labor Relations Review* 30.2, p. 175-187.

Akerlof, G. A. (août 1970). « The Market for ”Lemons” : Quality Uncertainty and the Market Mechanism ». In : *The Quarterly Journal of Economics* 84.3.

Al Ramiah, A., M. Hewstone, J. F. Dovidio et L. A. Penner (2010). « The social psychology of discrimination : Theory, measurement and consequences ». In : *Making Equality Count*. The Liffey Press, p. 84-112.

Alabi, D., N. Immorlica et A. Kalai (2018). « Unleashing linear optimizers for group-fair learning and optimization ». In : *Conference On Learning Theory*. PMLR, p. 2043-2066.

Alexander, L. (1992). « What makes wrongful discrimination wrong ? Biases, preferences, stereotypes, and proxies ». In : *University of Pennsylvania Law Review* 141.1, p. 149-219.

Alipourfard, N., P. G. Fennell et K. Lerman (2018). « Can you Trust the Trend ? Discovering Simpson’s Paradoxes in Social Data ». In : *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, p. 19-27.

Amadieu, J.-F. (2008). « Vraies et fausses solutions aux discriminations ». In : *Formation emploi. Revue française de sciences sociales* 101, p. 89-104.

Amossé, T. et G. De Peretti (2011). « Hommes et femmes en ménage statistique : une valse à trois temps ». In : *Travail, genre et sociétés* 2, p. 23-46.

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

Andreyeva, T., R. M. Puhl et K. D. Brownell (2008). « Changes in perceived weight discrimination among Americans, 1995–1996 through 2004–2006 ». In : *Obesity* 16.5, p. 1129-1134.

Anguraj, K. et S. Padma (2012). « Analysis of facial paralysis disease using image processing technique ». In : *International Journal of Computer Applications* 54.11.
Angwin, J., J. Larson, S. Mattu et L. Kirchner (2016). « Machine Bias : There's Software Used across the Country to Predict Future Criminals and It's Biased against Blacks ». In : May 23.
Aran, X. F., J. M. Such et N. Criado (2019). « Attesting biases and discrimination using language semantics ». In : *arXiv* 1909.04386.
Arnal, P. et R. Durand (2011). « Comment le sexe devint genre, et comment le genre devint code : le sort cruel d'une variable explicative ». In : *Risques* 87, p. 31-35.
Arya, S., C. Eckel et C. Wichman (2013). « Anatomy of the credit score ». In : *Journal of Economic Behavior & Organization* 95, p. 175-185.
Austin, R. (1983). « The insurance classification controversy ». In : *University of Pennsylvania Law Review* 131.3, p. 517-583.
Automobile Insurance Rate Board (2022). « Technical Guidance : Change in Rates and Rating Programs ». In : *Albera AIRB*.
Avraham, R. (2017). « Discrimination and insurance ». In : *Handbook of the Ethics of Discrimination*. Sous la dir. de K. Lippert-Rasmussen. Routledge, p. 335-347.
Avraham, R., K. D. Logue et D. Schwarcz (2014). « Towards a Universal Framework for Insurance Anti-Discrimination Laws ». In : *Connecticut Insurance Law Journal* 21, p. 1.
Awad, E., S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon et I. Rahwan (2018). « The moral machine experiment ». In : *Nature* 563.7729, p. 59-64.
Ayuso, M., M. Guillen et A. M. Pérez-Marín (2016). « Telematics and gender discrimination : some usage-based evidence on whether men's risk of accidents differs from women's ». In : *Risks* 4.2, p. 10.
Ayuso, M., R. Sánchez et M. Santolino (2020). « Does longevity impact the severity of traffic crashes ? A comparative study of young-older and old-older drivers ». In : *Journal of safety research* 73, p. 37-46.
Babic, B., S. Gerke, T. Evgeniou et I. G. Cohen (2021). « Beware explanations from AI in health care ». In : *Science* 373.6552, p. 284-286.
Badain, D. I. (1980). « Insurance redlining and the future of the urban core ». In : *Colum. JL & Soc. Probs.* 16, p. 1.
Bagdasaryan, E., O. Poursaeed et V. Shmatikov (2019). « Differential privacy has disparate impact on model accuracy ». In : *Advances in Neural Information Processing Systems* 32, p. 15479-15488.
Bailey, R. A. et L. J. Simon (1960). « Two studies in automobile insurance ratemaking ». In : *ASTIN Bulletin : The Journal of the IAA* 1.4, p. 192-217.
Baird, I. M. (1994). « Obesity and Insurance Risk ». In : *Pharmacoeconomics* 5.1, p. 62-65.
Baker, T. et J. Simon (2002). *Embracing Risk : The Changing Culture of Insurance and Responsibility. Chicago : Univ*. Chicago Press.
Baker, T. (2003). « Containing the promise of insurance : Adverse selection and risk classification ». In : *Connecticut Insurance Law Journal* 9, p. 371-396.
Baker, T. (2011). « Health insurance, risk, and responsibility after the Patient Protection and Affordable Care Act ». In : *University of Pennsylvania Law Review*, p. 1577-1622.
Ball, K. K., O. J. Clay, J. D. Edwards, B. A. Fausto, K. M. Wheeler, C. Felix et L. A. Ross (2021). « Indicators of crash risk in older adults : a longitudinal analysis from the ACTIVE Study ». In : *Journal of aging and health*.
Bamman, D., J. Eisenstein et T. Schnoebelen (2014). « Gender identity and lexical variation in social media ». In : *Journal of Sociolinguistics* 18.2, p. 135-160.

Barnard, C. et B. Hepple (1999). « Indirect Discrimination : Interpreting Seymour-Smith ». In : *The Cambridge Law Journal* 58.2, p. 399-412.
Barocas, S., M. Hardt et A. Narayanan (2017). « Fairness in machine learning ». In : *Nips tutorial* 1, p. 2017.
Barocas, S., M. Hardt et A. Narayanan (2019). *Fairness and Machine Learning*. fairmlbook.org.
Barocas, S. et A. D. Selbst (2016). « Big data's disparate impact ». In : *California Law Review* 104, p. 671-732.
Barry, L. (2020). « Insurance, Big Data and Changing Conceptions of Fairness ». In : *European Journal of Sociology* 61, p. 159-184.
Barry, L. et A. Charpentier (2020). « Personalization as a promise : Can Big Data change the practice of insurance ? » In : *Big Data & Society* 7.1, p. 2053951720935143.
Bartik, A. et S. Nelson (2016). « Deleting a signal : Evidence from pre-employment credit checks ». In : *SSRN* 2759560.
Bartlett, R., A. Morse, R. Stanton et N. Wallace (2018). « Consumer-lending discrimination in the era of fintech ». In : *University of California, Berkeley, Working Paper*.
Bartlett, R., A. Morse, R. Stanton et N. Wallace (2021). « Consumer-lending discrimination in the FinTech era ». In : *Journal of Financial Economics*.
Baumeister, R. F. (2017). « Addiction, cigarette smoking, and voluntary control of action : Do cigarette smokers lose their free will ? » In : *Addictive Behaviors Reports* 5, p. 67-84.
Bayer, P. B. (1986). « Mutable Characteristics and the Definition of Discrimination Under Title VII ». In : *UC Davis Law Review* 20, p. 769.
Becker, G. S. (2005). « Is Ethnic and other Profiling Discrimination ? » In : *The Becker-Posner Blog* 01-23-2005.
Becker, G. S. (1957). *The economics of discrimination*. University of Chicago press.
Beckert, J. (2013). « Imagined futures : fictional expectations in the economy ». In : *Theory and society* 42.3, p. 219-240.
Beider, P. (1987). « Sex Discrimination in Insurance ». In : *Journal of Applied Philosophy* 4, p. 65-75.
Bénéplanc, G., A. Charpentier et P. Thourot (2022). *Manuel d'assurance*. Presses Universitaires de France.
Beniger, J. (2009). *The control revolution : Technological and economic origins of the information society*. Harvard university press.
Benjamin, B. et R. Michaelson (1988). « Mortality differences between smokers and non-smokers ». In : *Journal of the Institute of Actuaries* 115.3, p. 519-525.
Bera, M. (2001). « a modélisation prédictive à très grand nombre de variables ». In : *Risques* 45, p. 32-37.
Bergkamp, L. (1989). « Life insurance and HIV testing : insurance theory, discrimination and solutions ». In : *Medecine & Law* 8, p. 567.
Bergstrom, C. T. et J. D. West (2021). *Calling bullshit : the art of skepticism in a data-driven world*. Random House Trade Paperbacks.
Berk, R., H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel et A. Roth (2017). « A convex framework for fair regression ». In : *arXiv* 1706.02409.
Berk, R., H. Heidari, S. Jabbari, M. Kearns et A. Roth (2021). « Fairness in criminal justice risk assessments : The state of the art ». In : *Sociological Methods & Research* 50.1, p. 3-44.
Berke, A., M. Bakker, P. Vepakomma, K. Larson et A. '. Pentland (2020). « Assessing Disease Exposure Risk with Location Data : A Proposal for Cryptographic Preservation of Privacy ». In : *arXiv* 2003.14412.
Bernstein, A. (2013). « What's wrong with stereotyping ». In : *Arizona Law Review* 55, p. 655.

Bertail, P., D. Bounie, S. Clémençon et P. Waelbroeck (2019). « Algorithmes : biais, discrimination et équité ». In.

Bertillon, A. et A. Chervin (1909). *Anthropologie métrique : conseils pratiques aux missionnaires scientifiques sur la manière de mesurer, de photographier et de décrire des sujets vivants et des pièces anatomiques*. Imprimerie nationale.

Bertrand, M. et S. Mullainathan (2004). « Are Emily and Greg more employable than Lakisha and Jamal ? A field experiment on labor market discrimination ». In : *American economic review* 94.4, p. 991-1013.

Besnard, P. et C. Grange (1993). « La fin de la diffusion verticale des gouts ?(Prénoms de l'élite et du vulgum) ». In : *L'Année sociologique*, p. 269-294.

Besse, P., E. del Barrio, P. Gordaliza et J.-M. Loubes (2018). « Confidence intervals for testing disparate impact in fair learning ». In : *arXiv* 1807.06362.

Beutel, A., J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi *et al.* (2019). « Fairness in recommendation ranking through pairwise comparisons ». In : *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, p. 2212-2220.

Beutel, A., J. Chen, Z. Zhao et E. H. Chi (2017). « Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations ». In : *arXiv* 1707.00075.

Bhattacharya, A. (2015). « Facebook patent : Your friends could help you get a loan - or not ». In : *CNN Business* 2015/08/04.

Bian, Y., C. Yang, J. L. Zhao et L. Liang (2018). « Good drivers pay less : A study of usage-based vehicle insurance models ». In : *Transportation research part A : policy and practice* 107, p. 20-34.

Bickel, P. J., E. A. Hammel et J. W. O'Connell (1975). « Sex bias in graduate admissions : Data from Berkeley ». In : *Science* 187.4175, p. 398-404.

Bidadanure, J. (2017). « Discrimination and age ». In : *Handbook of the Ethics of Discrimination*. Sous la dir. de K. Lippert-Rasmussen. Routledge, p. 243-253.

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

Bigot, R. et A. Cayol (2020). *Le droit des assurances en tableaux*. Ellipses.

Bigot, R., D. Cocteau-Senn et C. Arthur (2019). « La protection des données personnelles en assurance : dialogue du juriste avec l'actuaire ». In : *Regards sur le nouveau droit des données personnelles*. Sous la dir. d'E. Netter. CEPRISCA, collection Colloques.

Billings, P. R., M. A. Kohn, M. De Cuevas, J. Beckwith, J. S. Alper et M. R. Natowicz (1992). « Discrimination as a consequence of genetic testing. » In : *American journal of human genetics* 50.3, p. 476.

Birnbaum, B. (2020). « Insurance Consumer Protection Issues Resulting From, Or Heightened ByCOVID-19 ». In : *Center for Economic Justice Report*.

Black, D. R., J. P. Sciacca et D. C. Coster (1994). « Extremes in body mass index : probability of healthcare expenditures ». In : *Preventive medicine* 23.3, p. 385-393.

Black, E., S. Yeom et M. Fredrikson (2020). « FlipTest : Fairness Testing via Optimal Transport ». In : *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain, p. 111-121.

Blanchet, P. (2017). *Discriminations : combattre la glottophobie*. Éditions Textuel.

Blank, R. M., M. Dabady et C. F. Citro (2004). *Measuring racial discrimination*. National Academies Press.

Bloch, M. (1932). « Noms de personne et histoire sociale ». In : *Annales d'histoire économique et sociales* 4.13, p. 67-69.

Blodgett, S. L. et B. O'Connor (2017). « Racial disparity in natural language processing : A case study of social media african-american english ». In : *arXiv* 1707.00061.

Boczar, D., F. R. Avila, R. E. Carter, P. A. Moore, D. Giardi, G. Guliyeva, C. J. Bruce, C. J. McLeod et A. J. Forte (2021). « Using facial recognition tools for health assessment ». In : *Plastic Surgical Nursing* 41.2, p. 112-116.

Bohren, J. A., K. Haggag, A. Imas et D. G. Pope (2019). *Inaccurate statistical discrimination : An identification problem*. Rapp. tech. National Bureau of Economic Research.

Bolukbasi, T., K.-W. Chang, J. Zou, V. Saligrama et A. Kalai (2016). « Man is to Computer Programmer as Woman is to Homemaker ? Debiasing Word Embeddings ». In : *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Curran Associates Inc., p. 4356-4364.

Bond, E. W. et K. J. Crocker (1991). « Smoking, skydiving, and knitting : The endogenous categorization of risks in insurance markets with asymmetric information ». In : *Journal of Political Economy* 99.1, p. 177-200.

Bonnefon, J.-F. (2019). *La voiture qui en savait trop. L'intelligence artificielle a-t-elle une morale ?* Humensciences Editions.

Boonekamp, C. et D. Donaldson (1979). « Certain Alternatives for Price Uncertainty ». In : *The Canadian Journal of Economics / Revue canadienne d'Economique* 12.4, p. 718-728.

Borch, K. (1962). « Application of Game Theory to Some Problems in Automobile Insurance* ». In : *ASTIN Bulletin : The Journal of the IAA* 2.2, p. 208-221.

Borges, J. L. (1946). « Del rigor en la ciencia ». In : *Los Anales de Buenos Aires*.

Borgesius, F. Z., H. Schraffenberger et M. van Bekkum (2021). « Insurance, Algorithmic Decision-Making, and Discrimination ». In : *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*.

Borgesius, F. Z. (2020). « Price Discrimination, Algorithmic Decision-Making, and European Non-Discrimination Law ». In : *European Business Law Review* 31.3.

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

Born, P. (2019). « Genetic Testing in Underwriting : Implications for Life Insurance Markets. » In : *Journal of Insurance Regulation* 38.5.

Bornstein, S. (2018). « Antidiscriminatory algorithms ». In : *Alabama Law Review* 70, p. 519.

Bosmajian, H. A. (1974). *The language of oppression*. T. 10. Public Affairs Press.

Bouk, D. (2015). *How Our Days Became Numbered : Risk and the Rise of the Statistical Individual*. The University of Chicago Press.

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

Boyd, D., K. Levy et A. Marwick (2014). « The networked nature of algorithmic discrimination ». In : *Data and Discrimination : Collected Essays. Open Technology Institute*.

Brams, S. J., S. J. Brams et A. D. Taylor (1996). *Fair Division : From cake-cutting to dispute resolution*. Cambridge University Press.

Brandt, A. M. (2007). *cigarette century : the rise, fall and deadly persistence of the product that defined America*. Basic Books.

Breiman, L. (1984). *Classification and regression trees*. Wadsworth & Brooks.

Brilmayer, L., D. Laycock et T. A. Sullivan (1983). « The Efficient Use of Group Averages as Nondiscrimination : A Rejoinder to Professor Benston ». In : *The University of Chicago Law Review* 50.1, p. 222-249.

Brockett, P. L. et L. L. Golden (2007). « Biological and psychobehavioral correlates of credit scores and automobile insurance losses : Toward an explication of why credit scoring works ». In : *Journal of Risk and Insurance* 74.1, p. 23-63.
Brosnan, S. F. (2006). « Nonhuman species' reactions to inequity and their implications for fairness ». In : *Social Justice Research* 19.2, p. 153-185.
Brown, I. et C. T. Marsden (2013). *Regulating code : Good governance and better regulation in the information age*. MIT Press.
Browne, K. R. (1993). « Statistical proof of discrimination : beyond damned lies ». In : *Washington Law Review* 68, p. 477.
Brudno, B. (1976). *Poverty, Inequality, and the Law*. West Publishing Company.
Bruner, J. S. *et al.* (1957). « Going beyond the information given ». In : *Contemporary approaches to cognition*. Sous la dir. de J. Bruner, E. Brunswik, L. Festinger, F. Heider, K. Muenzinger, C. Osgood et D. Rapaport. Harvard University Press, p. 119-160.
Buchez, P. J. B. (1846). *Histoire parlementaire de la Révolution française*. T. 1. Hetzel.
Budd, L. P., R. A. Moorthi, H. Botha, A. C. Wicks et J. Mead (2021). « Automated Hiring at Amazon ». In : *Universiteit van Amsterdam* E-0470.
Bulmer, M. (2003). *Francis Galton : pioneer of heredity and biometry*. Johns Hopkins University Press.
Bunel, M., Y. L'Horty et P. Petit (2016). « Discrimination based on place of residence and access to employment ». In : *Urban Studies* 53.2, p. 267-286.
Buolamwini, J. et T. Gebru (2018). « Gender shades : Intersectional accuracy disparities in commercial gender classification ». In : *Conference on fairness, accountability and transparency*. PMLR, p. 77-91.
Bureau d'Assurance du Canada (2021). « Facts of the Property and Casualty Insurance Industry in Canada ». In : *Insurance Bureau of Canada*.
Burgdorf, M. P. et R. Burgdorf Jr (1974). « A history of unequal treatment : The qualifications of handicapped persons as a Suspect Class under the Equal Protection Clause ». In : *Santa Clara Lawyer* 15, p. 855.
Butler (1969). « Age-ism : Another form of bigotry ». In : *The gerontologist* 9.4_Part_1, p. 243-246.
Butler, P. et T. Butler (1989). « Driver record : A political red herring that reveals the basic flaw in automobile insurance pricing ». In : *Journal of Insurance Regulation* 8.2, p. 200-234.
Calders, T. et S. Verwer (2010). « Three naive bayes approaches for discrimination-free classification ». In : *Data mining and knowledge discovery* 21.2, p. 277-292.
Calders, T. et I. Žliobaitė (2013). « Why unbiased computational processes can lead to discriminative decision procedures ». In : *Discrimination and privacy in the information society*. Springer, p. 43-57.
Caliskan, A., J. J. Bryson et A. Narayanan (2017). « Semantics derived automatically from language corpora contain human-like biases ». In : *Science* 356.6334, p. 183-186.
Calmon, F. P., D. Wei, K. N. Ramamurthy et K. R. Varshney (2017). « Optimized Data Pre-Processing for Discrimination Prevention ». In : *arXiv* 1704.03354.
Cao, D., C. Chen, M. Piccirilli, D. Adjeroh, T. Bourlai et A. Ross (2011). « Can facial metrology predict gender ? » In : *2011 International Joint Conference on Biometrics (IJCB)*. IEEE, p. 1-8.
Cardon, D. (2019). *Culture numérique*. Presses de Sciences Po.
Carnis, L. et S. Lassarre (2019). « Politique et management de la sécurité routière ». In : *La sécurité routière en France, Quand la recherche fait son bilan et trace des perspectives*. Sous la dir. de C. Laurent, G. Catherine et G. Marie-Line. L'Harmattan.

Carpusor, A. G. et W. E. Loges (2006). « Rental discrimination and ethnicity in names ». In : *Journal of applied social psychology* 36.4, p. 934-952.
Carrasco, V. (2007). « Le pacte civil de solidarité : une forme d'union qui se banalise ». In : *Infostat Justice* 97.4.
Castelvecchi, D. (2016). « Can we open the black box of AI ? » In : *Nature News* 538.7623, p. 20.
Castro, C. (2019). « What's wrong with machine bias ». In : *Ergo, an Open Access Journal of Philosophy* 6.
Caton, S. et C. Haas (2020). « Fairness in machine learning : A survey ». In : *arXiv* 2010.04053.
Chakraborty, S., K. R. Raghavan, M. P. Johnson et M. B. Srivastava (2013). « A Framework for Context-Aware Privacy of Sensor Data on Mobile Systems ». In : *Proceedings of the 14th Workshop on Mobile Computing Systems and Applications*. HotMobile '13. Jekyll Island, Georgia : Association for Computing Machinery. isbn : 9781450314213.
Chardenon, A. (2019). « Voici Maxime, le chatbot juridique d'Axa, fruit d'une démarche collaborative ». In : *L'usine digitale* 12 février.
Charles, K. K. et J. Guryan (2011). « Studying discrimination : Fundamental challenges and recent progress ». In : *Annual Review of Economics* 3.1, p. 479-511.
Charpentier, A. (2014). *Computational Actuarial Science*. The R series. CRC Press.
Charpentier, A. et M. Denuit (2004). *Mathématiques de l'assurance non-vie - Tarification et provisionnement (Tome 1)*. Economica.
Charpentier, A. et M. Denuit (2005). *Mathématiques de l'assurance non-vie - Principes fondamentaux de théorie du risque (Tome 2)*. Economica.
Charpentier, A., R. Élie et C. Remlinger (2020). « Reinforcement Learning in Economics and Finance ». In : *arXiv* 2003.10014.
Charpentier, A., E. Flachaire et A. Ly (2018). « Econometrics and machine learning ». In : *Economie et Statistique* 505.1, p. 147-169.
Chassagnon, A. (1996). « Sélection adverse : modèle générique et applications ». Thèse de doct. Paris, EHESS.
Chaufton, A. (1886). *Les assurances, leur passé, leur présent, leur avenir, au point de vue rationnel, technique et pratique, moral, économique et social, financier et administratif, légal, législatif et contractuel, en France et à l'étranger*. Chevalier-Marescq.
Cheung, I. et A. T. McCartt (2011). « Declines in fatal crashes of older drivers : Changes in crash risk and survivability ». In : *Accident Analysis & Prevention* 43.3, p. 666-674.
Chouldechova, A. (2017). « Fair prediction with disparate impact : A study of bias in recidivism prediction instruments ». In : *Big data* 5.2, p. 153-163.
Chun, W. H. K. (2021). *Discriminating Data : Correlation, Neighborhoods, and the New Politics of Recognition*. MIT Press.
Clarke, D. D., P. Ward, C. Bartle et W. Truman (2010). « Older drivers' road traffic crashes in the UK ». In : *Accident Analysis & Prevention* 42.4, p. 1018-1024.
Cohen, A. et P. Siegelman (2010). « Testing for adverse selection in insurance markets ». In : *Journal of Risk and insurance* 77.1, p. 39-84.
Cohen, J. E. (1986). « An uncertainty principle in demography and the unisex issue ». In : *The American Statistician* 40.1, p. 32-39.
Coldman, A. J., T. Braun et R. P. Gallagher (1988). « The classification of ethnic status using name information. » In : *Journal of Epidemiology & Community Health* 42.4, p. 390-395.
Collins, E. (2018). « Punishing Risk ». In : *Georgetown Law Journal* 107, p. 57.

Colson, K. E., K. E. Rudolph, S. C. Zimmerman, D. E. Goin, E. A. Stuart, M. Van Der Laan et J. Ahern (2016). « Optimizing matching and analysis combinations for estimating causal effects ». In : *Scientific reports* 6.1, p. 1-11.
Com-Ruelle, L. et S. Dumesnil (1999). « Concentration des dépenses et grands consommateurs de soins médicaux ». In : *Bulletin d'information en économie de la santé* 20, p. 1-4.
Conseil de l'Union Européenne (2004). « Directive 2004/113/CE du Conseil du 13 décembre 2004 mettant en œuvre le principe de l'égalité de traitement entre les femmes et les hommes dans l'accès ' des biens et services et la fourniture de biens et services ». In : *JO L* 373, p. 37-43.
Cook, R. D. (1980). « Smoking and lung cancer ». In : *RA Fisher : An Appreciation*. Springer, p. 182-191.
Cook, T. D., D. T. Campbell et W. Shadish (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, MA.
Cooper, P. J. (1990). « Differences in accident characteristics among elderly drivers and between elderly and middle-aged drivers ». In : *Accident analysis & prevention* 22.5, p. 499-508.
Cooper, P. F., R. J. Manski et J. V. Pepper (2012). « The effect of dental insurance on dental care use and selection bias ». In : *Medical Care*, p. 757-763.
Corbett-Davies, S., E. Pierson, A. Feller, S. Goel et A. Huq (2017). « Algorithmic decision making and the cost of fairness ». In : *arXiv* 1701.08230.
Corlier, F. (1998). « Segmentation : le point de vue de l'assureur ». In : *Compétitivité, éthique et assurance*. Sous la dir. de H. Cousy, H. Classens et C. Van Schoubroeck. Academia Bruylant.
Cornu, G. (2016). *Vocabulaire juridique*. T. 6. Presses Universitaires de France.
Correll, J., C. M. Judd, B. Park et B. Wittenbrink (2010). « Measuring prejudice, stereotypes and discrimination ». In : *The SAGE handbook of prejudice, stereotyping and discrimination*, p. 45-62.
Coulmont, B. (2011). *Sociologie des prénoms*. la Découverte.
Coulmont, B. et P. Simon (2019). « Quels prénoms les immigrés donnent-ils à leurs enfants en France ? » In : *Population Societes* 4, p. 1-4.
Coutts, S. (1984). « Motor insurance rating, an actuarial approach ». In : *Journal of the Institute of Actuaries* 111.1, p. 87-148.
Cowell, M. J. et B. L. Hirst (1979). *Mortality differences between smokers and non-smokers*. State Mutual Life Assurance Company of America Worcester, Mass.
Cowgill, B. et C. E. Tucker (2019). « Economics, fairness and algorithmic bias ». In : *Journal of Economic Perspectives*.
Crawford, K. et R. Calo (2016). « There is a blind spot in AI research ». In : *Nature* 538, p. 311-313.
Creswell, J. (2010). « Speedy new traders make waves far from Wall Street ». In : *New York Times* May 17.
Crizzle, A. M., S. Classen et E. Y. Uc (2012). « Parkinson disease and driving : an evidence-based review ». In : *Neurology* 79.20, p. 2067-2074.
Crocker, K. J. et A. Snow (2013). « The theory of risk classification ». In : *Handbook of insurance*. Sous la dir. de H. Loubergé et G. Dionne. Springer, p. 281-313.
Cuddeback, G., E. Wilson, J. G. Orme et T. Combs-Orme (2004). « Detecting and statistically correcting sample selection bias ». In : *Journal of Social Service Research* 30.3, p. 19-33.
Cummins, J. D., B. D. Smith, R. N. Vance et J. Vanderhel (2013). *Risk classification in life insurance*. T. 1. Springer Science & Business Media.
Czerniawski, A. M. (2007). « From Average to Ideal : The Evolution of the Height and Weight Table in the United States, 1836-1943 ». In : *Social Science History* 31.2, p. 273-296.

Dahlby, B. G. (1983). « Adverse selection and statistical discrimination : An analysis of Canadian automobile insurance ». In : *Journal of Public Economics* 20.1, p. 121-130.
Dalenius, T. (1977). « Towards a methodology for statistical disclosure control ». In : *statistik Tidskrift* 15.429-444, p. 2-1.
Dalziel, J. R. et R. S. Job (1997). « Motor vehicle accidents, fatigue and optimism bias in taxi drivers ». In : *Accident Analysis & Prevention* 29.4, p. 489-494.
Dane, S. M. (2006). « The potential for racial discrimination by homeowners insurers through the use of geographic rating territories ». In : *Journal of Insurance Regulation* 24.4, p. 21.
Daniel, J. E. et J. L. Daniel (1998). « Preschool children's selection of race-related personal names ». In : *Journal of Black Studies* 28.4, p. 471-490.
Daniels, G. S. (1952). *The" Average Man"?* Rapp. tech. Air Force Aerospace Medical Reseach Lab, Wright-Patterson AFB OH.
Daniels, N. (1990). « Insurability and the HIV epidemic : ethical issues in underwriting ». In : *The Milbank Quarterly*, p. 497-525.
Daniels, N. (2004). « The functions of insurance and the fairness of genetic underwriting ». In : *Genetics and life insurance : Medical underwriting and social policy*, p. 119-145.
Dares (2020). « Temps partiel : des conditions d'emploi contrastées ». In : *Dars Analyses*.
Daston, L. (1987). « The domestication of risk : Mathematical probability and insurance, 1650-1830 ». In : *The probabilistic revolution*. MIT, p. 237-260.
Daston, L. (1992). « Objectivity and the Escape from Perspective ». In : *Social studies of science* 22.4, p. 597-618.
Datta, A., A. Datta, J. Makagon, D. K. Mulligan et M. C. Tschantz (2018). « Discrimination in online advertising : A multidisciplinary inquiry ». In : *Conference on Fairness, Accountability and Transparency*. PMLR, p. 20-34.
Davenport, T. (2006). « Competing on analytics ». In : *harvard Business Review,* 84, p. 1-10.
David, H. (2015). « Why are there still so many jobs? The history and future of workplace automation ». In : *Journal of economic perspectives* 29.3, p. 3-30.
Davidson, R. et J. G. MacKinnon (2004). *Econometric theory and methods*. T. 5. Oxford University Press New York.
Davis, G. A. (2004). « Possible aggregation biases in road safety research and a mechanism approach to accident modeling ». In : *Accident Analysis & Prevention* 36.6, p. 1119-1127.
Dawid, A. P. (2000). « Causal inference without counterfactuals ». In : *Journal of the American statistical Association* 95.450, p. 407-424.
De Pril, N. et J. Dhaene (1996). « Segmentering in verzekeringen ». In : *DTEW Research Report 9648*, p. 1-56.
De Wit, G. et J. Van Eeghen (1984). « Rate making and society's sense of fairness ». In : *ASTIN Bulletin : The Journal of the IAA* 14.2, p. 151-163.
De Witt, J. (1671). « Value of life annuities in proportion to redeemable annuities ». In : *Originally in Dutch. Translated in* Hendriks *1853*, p. 232-49.
Debet, A. (2007). « Mesure de la diversité et protection des données personnelles ». In : *Commission Nationale de l'Informatique et des Libertés* 16/05/2007 08:40 DECO / IRC.
Défenseur des droits (2020). « Algorithmes : prévenir l'automatisation des discriminations ». In.
Demuijnck, G. (2009). « Non-discrimination in human resources management as a moral obligation ». In : *Journal of Business Ethics* 88.1, p. 83-101.
Denuit, M. et A. Charpentier (2004). *Mathématiques de l'assurance non-vie : Tome I Principes fondamentaux de théorie du risque*. Economica.

Denuit, M., X. Maréchal, S. Pitrebois et J.-F. Walhin (2007). *Actuarial modelling of claim counts : Risk classification, credibility and bonus-malus systems*. John Wiley & Sons.
Depoid, P. (1967). *Applications de la statistique aux assurances accidents et dommages : cours professé à l'Institut de statistique de l'Université de Paris. 2e édition revue et augmentée...* Berger-Levrault.
Deschamps, J.-C. et B. Personnaz (1979). « Etudes entre groupes 'dominants' et 'dominés' : Importance de la présence du hors-groupe dans les discriminations évaluatives et comportementales ». In : *Social Science Information* 18.2, p. 269-305.
Desrosières, A. (2016). *La politique des grands nombres : histoire de la raison statistique*. La découverte.
Devine, P. G. (1989). « Stereotypes and prejudice : Their automatic and controlled components. » In : *Journal of personality and social psychology* 56.1, p. 5.
Dilley, S. et G. Greenwood (2017). « Abandoned 999 calls to police more than double ». In : *BBC* 19 September 2017.
DiNardo, J. (2016). « Natural Experiments and Quasi-Natural Experiments ». In : *The New Palgrave Dictionary of Economics*. London : Palgrave Macmillan UK, p. 1-12.
Dinur, R., B. Beit-Hallahmi et J. E. Hofman (1996). « First names as identity stereotypes ». In : *The Journal of social psychology* 136.2, p. 191-200.
Doll, R. et A. B. Hill (1964). « Mortality in relation to smoking : ten years' observations of British doctors ». In : *British medical journal* 1.5395, p. 1399.
Doll, R., R. Peto, J. Boreham et I. Sutherland (2004). « Mortality in relation to smoking : 50 years' observations on male British doctors ». In : *British Medical Journal* 328.7455, p. 1519.
Dorlin, E. (2005). « Sexe, genre et intersexualité : la crise comme régime théorique ». In : *Raisons politiques* 2, p. 117-137.
Dorn, H. F. (1958). « The mortality of smokers and nonsmokers ». In : *Proceedings of the social statistics section*. T. 1, p. 34-71.
Dostie, G. (1974). « Entrevue de Michèle Lalonde ». In : *Le Journal* 1er juin 1974.
Dressel, J. et H. Farid (jan. 2018). « The accuracy, fairness, and limits of predicting recidivism ». In : *Science Advances* 4.1, eaao5580.
Du Bois, W. (1896). « Review of Race Traits and Tendencies of the American Negro ». In : *Annals of the American Academy*, p. 127-33.
Dubet, F. (2014). *La Préférence pour l'inégalité. Comprendre la crise des solidarités : Comprendre la crise des solidarités*. Seuil - La République des idées.
Dubet, F. (2016). *Ce qui nous unit : Discriminations, égalité et reconnaissance*. Seuil - La République des idées.
Dubourg, É. et V. Gautron (2014). « La rationalisation des méthodes d'évaluation des risques de récidive. Entre promotion institutionnelle, réticences professionnelles et prudence interprétative ». In : *Champ pénal/Penal field* 11.
Duhigg, C. (2012). « How companies learn your secrets ». In : *The New York Times* 02-16-2019.
Duivesteijn, W. et A. Feelders (2008). « Nearest neighbour classification with monotonicity constraints ». In : *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, p. 301-316.
Dulisse, B. (1997). « Older drivers and risk to other road users ». In : *Accident Analysis & Prevention* 29.5, p. 573-582.
Duncan, C. et W. Loretto (2004). « Never the right age? Gender and age-based discrimination in employment ». In : *Gender, Work & Organization* 11.1, p. 95-115.
Durkheim, É. (1897). *Le suicide : étude sociologique*. Félix Alcan Editeur.

Durry, G. (2001). « La sélection de la clientèle par l'assureur : aspects juridiques ». In : *Risques* 45, p. 65-71.
Dwork, C., M. Hardt, T. Pitassi, O. Reingold et R. Zemel (2012). « Fairness through awareness ». In : *Proceedings of the 3rd innovations in theoretical computer science conference*, p. 214-226.
Dwoskin, E. (2018). « Facebook is rating the trustworthiness of its users on a scale from zero to one ». In : *Washington Post* 21-08.
Eco, U. (1992). *Comment voyager avec un saumon*. Grasset.
Edelman, B., M. Luca et D. Svirsky (2017). « Racial discrimination in the sharing economy : Evidence from a field experiment ». In : *American economic journal : applied economics* 9.2, p. 1-22.
Edgeworth, F. Y. (1922). « Equal pay to men and women for equal work ». In : *The Economic Journal* 32.128, p. 431-457.
Ekeland, I. (1995). *Le chaos*. Flammarion.
Ellis, E. et P. Watson (2012). *EU anti-discrimination law*. Oxford University Press.
Epstein, L. et G. King (2002). « The rules of inference ». In : *The University of Chicago Law Review*, p. 1-133.
Ericson, R. V., A. Doyle, D. Barry et D. Ericson (2003). *Insurance as governance*. University of Toronto Press.
Erwin, P. G. (1995). « A review of the effects of personal name stereotypes. » In : *Representative Research in Social Psychology*.
Escafré-Dublet, A., L. Kesztenbaum et P. Simon (2020). « Quand le recensement comptait les Français musulmans ». In : *Population & Sociétés* 11, p. 1-4.
Espeland, W. N. et M. L. Stevens (1998). « Commensuration as a social process ». In : *Annual review of sociology* 24.1, p. 313-343.
Eubanks, V. (2018). *Automating inequality : How high-tech tools profile, police, and punish the poor*. St. Martin's Press.
Ewald, F. (1986). *Histoire de l'Etat providence : les origines de la solidarité*. Grasset.
Fagyal, Z. (2010). « Accents de banlieue ». In : *Aspects prosodiques du français populaire en contact avec les langues de l'immigration, L'Harmattan*.
Farkas, L. (2011). *How to present a discrimination claim : handbook on seeking remedies under the EU non-discrimination directives*. Publications Office of the European Union.
Favaretto, M., E. De Clercq et B. S. Elger (2019). « Big Data and discrimination : perils, promises and solutions. A systematic review ». In : *Journal of Big Data* 6.1, p. 1-27.
Fazelpour, S. et Z. C. Lipton (2020). « Algorithmic fairness from a non-ideal perspective ». In : *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, p. 57-63.
Feeley, M. et J. Simon (1994). « Actuarial justice : The emerging new criminal law ». In : *The futures of criminology* 173, p. 174.
Feine, J., U. Gnewuch, S. Morana et A. Maedche (2019). « Gender bias in chatbot design ». In : *International Workshop on Chatbot Research and Design*. Springer, p. 79-93.
Feiring, E. (2009). « reassessing insurers' access to genetic information : genetic privacy, ignorance, and injustice ». In : *Bioethics* 23.5, p. 300-310.
Feldman, M., S. A. Friedler, J. Moeller, C. Scheidegger et S. Venkatasubramanian (2015). « Certifying and removing disparate impact ». In : *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, p. 259-268.
Feller, W. (1957). *An introduction to probability theory and its applications*. Wiley.
Ferrence, R. G. (1990). *Deadly fashion : The rise and fall of cigarette smoking in North America*. Garland.
Finkelstein, A. (2014). *Moral hazard in health insurance*. Columbia University Press.

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, K. Baicker et O. H. S. Group (2012). « The Oregon health insurance experiment : evidence from the first year ». In : *The Quarterly journal of economics* 127.3, p. 1057-1106.

Finkelstein, E. A., D. S. Brown, L. A. Wrage, B. T. Allaire et T. J. Hoerger (2010). « Individual and aggregate years-of-life-lost associated with overweight and obesity ». In : *Obesity* 18.2, p. 333-339.

Fiscella, K. et A. M. Fremont (2006). « Use of geocoding and surname analysis to estimate race and ethnicity ». In : *Health services research* 41.4p1, p. 1482-1500.

Fischer, M. J. et D. S. Massey (2004). « The ecology of racial discrimination ». In : *City & Community* 3.3, p. 221-241.

Fisher, R. (1958). « Cancer and smoking ». In : *Nature* 182.4635, p. 596-596.

Fisher, R. (1959). *Smoking : the cancer controversy : some attempts to assess the evidence*. Oliver et Boyd Edinburgh.

Fiske, T. S. (1917). « Emory McClintock ». In : *Bulletin of the American Mathematical Society* 23.8, p. 353-357.

Fix, M. et M. A. Turner (1998). *A National Report Card on Discrimination in America : The Role of Testing*. ERIC : Proceedings of the Urban Institute Conference (Washington, DC, March 1998).

Fleurbaey, M. (1996). *Théories économiques de la justice*. Economica.

Fontaine, H. (2003). « Driver age and road traffic accidents : what is the risk for seniors ? » In : *Recherche-transports-sécurité*.

Fontaine, K. R., D. T. Redden, C. Wang, A. O. Westfall et D. B. Allison (2003). « Years of life lost due to obesity ». In : *Journal of the American Medical Association* 289.2, p. 187-193.

Foot, P. (1967). « The problem of abortion and the doctrine of the double effect ». In : *Oxford review* 5.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Sage Publications, Inc.

France Info (2019). « Intelligence artificielle : comment faire créer son modèle mathématique ? » In : *France Info* 26-11-2019.

Freedman, D. A. (1999). « Ecological inference and the ecological fallacy ». In : *International Encyclopedia of the social & Behavioral sciences* 6.4027-4030, p. 1-7.

Frees, E. W. (2009). *Regression modeling with actuarial and financial applications*. Cambridge University Press.

Frezal, S. et L. Barry (2019). « Fairness in Uncertainty : Some Limits and Misinterpretations of Actuarial Fairness ». In : *Journal of Business Ethics*. issn : 1573-0697.

Fricker, M. (2007). *Epistemic injustice : Power and the ethics of knowing*. Oxford University Press.

Friedler, S. A., C. Scheidegger et S. Venkatasubramanian (2016a). « On the (im) possibility of fairness ». In : *arXiv* 1609.07236.

Friedler, S. A., C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton et D. Roth (2019). « A comparative study of fairness-enhancing interventions in machine learning ». In : *Proceedings of the conference on fairness, accountability, and transparency*, p. 329-338.

Friedler, S. A., C. Scheidegger et S. Venkatasubramanian (2016b). « On the (im)possibility of fairness ». In : *arXiv* 1609.07236.

Friedman, J. H. (2001). « Greedy function approximation : a gradient boosting machine ». In : *Annals of statistics*, p. 1189-1232.

Friedman, S. et M. Canaan (2014). « Overcoming speed bumps on the road to telematics ». In : *Challenges and opportunities facing auto insurers with and without usage-based programs*.

Frisch, R. et F. V. Waugh (1933). « Partial time regressions as compared with individual trends ». In : *Econometrica*, p. 387-401.
Froot, K. A., M. Kim et K. S. Rogoff (1995). « The law of one price over 700 years ». In : *National Bureau of Economic Research (NBER)* 5132.
Fukuchi, K., T. Kamishima et J. Sakuma (2015). « Prediction with model-based neutrality ». In : *IEICE TRANSACTIONS on Information and Systems* 98.8, p. 1503-1516.
Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai et A. Walther (2020). « Predictably unequal ? the effects of machine learning on credit markets ». In : *The Effects of Machine Learning on Credit Markets*.
Gadet, F. (2007). *La variation sociale en français*. Editions Ophrys.
Gajane, P. et M. Pechenizkiy (2017). « On formalizing fairness in prediction with machine learning ». In : *arXiv preprint arXiv:1710.03184*.
Galhotra, S., Y. Brun et A. Meliou (2017). « Fairness testing : testing software for discrimination ». In : *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, p. 498-510.
Galichon, A. (2016). *Optimal transport methods in economics*. Princeton University Press.
Gambs, S., M.-O. Killijian et M. N. del Prado Cortez (2010). « Show Me How You Move and I Will Tell You Who You Are ». In : *Proceedings of the 3rd ACM International Workshop on Security and Privacy in GIS and LBS*.
Gandy, O. H. (2016). *Coming to terms with chance : Engaging rational discrimination and cumulative disadvantage*. Routledge.
Garg, N., L. Schiebinger, D. Jurafsky et J. Zou (2018). « Word embeddings quantify 100 years of gender and ethnic stereotypes ». In : *Proceedings of the National Academy of Sciences* 115.16, E3635-E3644.
Gautron, V. et É. Dubourg (2015). « La rationalisation des outils et méthodes d'évaluation : de l'approche clinique au jugement actuariel ». In : *Criminocorpus. Revue d'Histoire de la justice, des crimes et des peines*.
Gebru, T., J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden et L. Fei-Fei (2017). « Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States ». In : *Proceedings of the National Academy of Sciences* 114.50, p. 13108-13113.
Geller, L. N., J. S. Alper, P. R. Billings, C. I. Barash, J. Beckwith et M. R. Natowicz (1996). « Individual, family, and societal dimensions of genetic discrimination : a case study analysis ». In : *Science and Engineering Ethics* 2.1, p. 71-88.
Gelman, A. (2009). *Red state, blue state, rich state, poor state : Why Americans vote the way they do-expanded edition*. Princeton University Press.
Gelman, A. et J. Hill (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.
Ghili, S., E. Kazemi et A. Karbasi (2019). « Eliminating latent discrimination : Train then mask ». In : *Proceedings of the AAAI Conference on Artificial Intelligence*. T. 33. 01, p. 3672-3680.
Gibbs, W. W. (1995). « Treatment that tightens the belt : is insurance part of America's obesity problem ? » In : *Scientific American* 272.3, p. 34-35.
Giles, C. (2020). « Goodhart's Law Comes back to Haunt the UK's Covid Strategy ». In : *Financial Times* 14-5.
Gillis, T. B. et J. L. Spiess (2019). « Big data and discrimination ». In : *The University of Chicago Law Review* 86.2, p. 459-488.

Gilpin, L. H., D. Bau, B. Z. Yuan, A. Bajwa, M. Specter et L. Kagal (2018). « Explaining explanations : An overview of interpretability of machine learning ». In : *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, p. 80-89.
Gino, F. et L. Pierce (2010). « Robin Hood under the hood : Wealth-based discrimination in illicit customer help ». In : *Organization Science* 21.6, p. 1176-1194.
Glenn, B. J. (2000). « The shifting rhetoric of insurance denial ». In : *Law and Society Review*, p. 779-808.
Glenn, B. J. (2003). « Postmodernism : the basis of insurance ». In : *Risk Management and Insurance Review* 6.2, p. 131-143.
Gollier, C. (2002). « La solidarite sous l'angle economique' ». In : *Revue Générale du Droit des Assurances*, p. 824-830.
Gordaliza, P., E. Del Barrio, G. Fabrice et J.-M. Loubes (2019). « Obtaining fairness using optimal transport theory ». In : *International Conference on Machine Learning*. PMLR, p. 2357-2365.
Gosseries, A. (2014). « What makes age discrimination special : A philosophical look at the ECJ case law ». In : *Netherlands Journal of Legal Philosophy* 43, p. 59-80.
Gottlieb, S. (2011). « Medicaid is worse than no coverage at all ». In : *Wall Street Journal* 10/03.
Gouriéroux, C. (1999a). « The econometrics of risk classification in insurance ». In : *The Geneva Papers on Risk and Insurance Theory* 24.2, p. 119-137.
Gouriéroux, C. (1999b). *Statistique de l'assurance*. Economica.
Gouriéroux, C. et J. Jasiak (2011). *The econometrics of individual risk*. Princeton university press.
Granger, C. W. (1969). « Investigating causal relations by econometric models and cross-spectral methods ». In : *Econometrica : journal of the Econometric Society*, p. 424-438.
Grari, V., A. Charpentier, S. Lamprier et M. Detyniecki (2022). « A fair pricing model via adversarial learning ». In : *arXiv* 2202.12008.
Greenland, S. (2002). « Causality theory for policy uses ». In : *Summary measures of population*. Sous la dir. de C. Murray. Harvard University Press, p. 291-302.
Grobon, S. et L. Mourlot (2014). « Le genre dans la statistique publique en France ». In : *Regards croisés sur l'économie* 2, p. 73-79.
Groupement des Assureurs Automobiles (2021). « Plan statistique automobile, Résultats généraux, Voitures de tourisme ». In : *GAA*.
Gschlössl, S., P. Schoenmaekers et M. Denuit (2011). « Risk classification in life insurance : methodology and case study ». In : *European Actuarial Journal* 1.1, p. 23-41.
Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti et D. Pedreschi (août 2018). « A Survey of Methods for Explaining Black Box Models ». In : *ACM Comput. Surv.* 51.5.
Gurcan, O. (2018). « Genetic Discrimination, Life Insurance, and Justice as Fairness ». In : *Canadian Society for Study of Practical Ethics*.
Guseva, A. et A. Rona-Tas (2001). « Uncertainty, risk, and trust : Russian and American credit card markets compared ». In : *American sociological review*, p. 623-646.
Hager, W. D. et L. Zimpleman (1982). « The Norris Decision, Its Implications and Application ». In : *Drake Law Review* 32, p. 913.
Hajian, S., F. Bonchi et C. Castillo (2016). « Algorithmic bias : From discrimination discovery to fairness-aware data mining ». In : *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, p. 2125-2126.
Hakamies-Blomqvist, L. E. (1993). « Fatal accidents of older drivers ». In : *Accident Analysis & Prevention* 25.1, p. 19-27.
Hale, K. (2021). « A.I. Bias Caused 80% Of Black Mortgage Applicants To Be Denied ». In : *Forbes* 09/2021.

Hall, M. A. et S. S. Rich (2000). « Laws restricting health insurers' use of genetic information : impact on genetic discrimination ». In : *The American Journal of Human Genetics* 66.1, p. 293-307.
Halpern, J. Y. (2016). *Actual causality*. MiT Press.
Hamilton, D. L. et R. K. Gifford (1976). « Illusory correlation in interpersonal perception : A cognitive basis of stereotypic judgments ». In : *Journal of Experimental Social Psychology* 12.4, p. 392-407.
Hand, D. J. (2020). *Dark Data : Why What You Don't Know Matters*. Princeton University Press.
Hara, K., J. Sun, R. Moore, D. W. Jacobs et J. Froehlich (2014). « Tohme : detecting curb ramps in google street view using crowdsourcing, computer vision, and machine learning ». In : *Proceedings of the 27th annual ACM symposium on User interface software and technology*.
Hara, K., J. Sun, J. Chazan, D. Jacobs et J. E. Froehlich (2013). « An initial study of automatic curb ramp detection with crowdsourced verification using google street view images ». In : *First AAAI Conference on Human Computation and Crowdsourcing*.
Harari, Y. N. (2018). *21 Lessons for the 21st Century*. Random House.
Harcourt, B. E. (2008). *Against prediction*. University of Chicago Press.
Harcourt, B. E. (jan. 2011). « Surveiller et punir à l'âge actuariel ». In : *Déviance et Société* 35, p. 163.
Hardt, M., E. Price et N. Srebro (2016). « Equality of opportunity in supervised learning ». In : *Advances in neural information processing systems* 29, p. 3315-3323.
Hargreaves, D. J., A. M. Colman et W. Sluckin (1983). « The attractiveness of names ». In : *Human Relations* 36.4, p. 393-401.
Harrington, S. E. et G. Niehaus (1998). « Race, redlining, and automobile insurance prices ». In : *The Journal of Business* 71.3, p. 439-469.
Harris, M. I. (1999). « Racial and ethnic differences in health insurance coverage for adults with diabetes. » In : *Diabetes Care* 22.10, p. 1679-1682.
Harwayne, F. (1959). « Merit Rating in Private Passenger Automobile Liability Insurance and the California Driver Record Study ». In : *Automobile Insurance Rate Making (New York : Casualty Actuarial Society, 1961)*, p. 205-206.
Haugeland, J. (1989). *Artificial intelligence : The very idea*. MIT press.
Heckman, J. J. (1992). « Randomization and social policy evaluation ». In : *Evaluating welfare and training programs* 1, p. 201-230.
Hedges, B. A. (1977). « Gender discrimination in pension plans : Comment ». In : *The Journal of Risk and Insurance* 44.1, p. 141-144.
Heen, M. L. (2009). « Ending Jim Crow life insurance rates ». In : *northwestern Jounral of Law & Social Policy* 4, p. 360.
Heller, D. (2015). *High price of mandatory auto insurance in predominantly African American Communities*. Rapp. tech. Consumer Federation of America.
Hellman, D. S. (1998). « Two Types of Discrimination : The Familiar and the Forgotten ». In : *California Law Review* 86, p. 315.
Hendriks, F. (1853). « Contributions to the history of insurance, and of the theory of life contingencies ». In : *Journal of the Institute of Actuaries* 3.2, p. 93-120.
Henriet, D. et J.-C. Rochet (1987). « Some reflections on insurance pricing ». In : *European Economic Review* 31.4, p. 863-885.
Héran, F. (2010). *Inégalités et discriminations : Pour un usage critique et responsable de l'outil statistique*. Commissaire à la Diversité et à l'Égalité des Chances.
Heras, A. J., P.-C. Pradier et D. Teira (2020). « What was fair in actuarial fairness ? » In : *History of the Human Sciences* 33.2, p. 91-114.
Hernán, M. A. et J. M. Robins (2010). *Causal inference*. CRC Press.

Hildebrandt, M. et S. Gutwirth (2008). *Profiling the European citizen*. Springer.
Hoffman, F. L. (1931). « Cancer and smoking habits ». In : *Annals of surgery* 93.1, p. 50.
Hoffman, F. L. (1896). *Race traits and tendencies of the American Negro*. T. 11. 1-3. American Economic Association.
Hoffman, K. M., S. Trawalter, J. R. Axt et M. N. Oliver (2016). « Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites ». In : *Proceedings of the National Academy of Sciences* 113.16, p. 4296-4301.
Hoffmann, A. L. (2019). « Where fairness fails : data, algorithms, and the limits of antidiscrimination discourse ». In : *Information, Communication & Society* 22.7, p. 900-915.
Hofstede, G. (1995). « Insurance as a product of national values ». In : *The Geneva Papers on Risk and Insurance-Issues and Practice* 20.4, p. 423-429.
Holland, P. W. (1986). « Statistics and causal inference ». In : *Journal of the American statistical Association* 81.396, p. 945-960.
Hong, D., Y.-Y. Zheng, Y. Xin, L. Sun, H. Yang, M.-Y. Lin, C. Liu, B.-N. Li, Z.-W. Zhang, J. Zhuang *et al.* (2021). « Genetic syndromes screening by facial recognition technology : VGG-16 screening model construction and evaluation ». In : *Orphanet Journal of Rare Diseases* 16.1, p. 1-8.
Hooker, S., N. Moorosi, G. Clark, S. Bengio et E. Denton (2020). « Characterising bias in compressed models ». In : 2010.03058.
Horvitz, D. G. et D. J. Thompson (1952). « A generalization of sampling without replacement from a finite universe ». In : *Journal of the American statistical Association* 47.260, p. 663-685.
Hu, P. S., D. A. Trumble, D. J. Foley, J. W. Eberhard et R. B. Wallace (1998). « Crash risks of older drivers : a panel data analysis ». In : *Accident Analysis & Prevention* 30.5, p. 569-581.
Hume, D. (1739). *A Treatise of Human Nature*. Cambridge University Press Archive.
Hunt, E. (2016). « Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter ». In : *The Guardian* 24.3, p. 2016.
Ilic, L., M. Sawada et A. Zarzelli (2019). « Deep mapping gentrification in a large Canadian city using deep learning and Google Street View ». In : *PloS one* 14.3, e0212814.
Imai, K. (2018). *Quantitative social science : an introduction*. Princeton University Press.
Ingold, D. et S. Soper (2106). « Amazon Doesn't Consider the Race of Its Customers. Should It ? » In : *Bloomberg* April 21st.
Ito, J. (2021). « Supposedly 'Fair' Algorithms Can Perpetuate Discrimination ». In : *Wired* 02.05.2019.
Jacobs, D. B. et B. D. Sommers (2015). « Using drugs to discriminate—adverse selection in the insurance marketplace ». In : *New England Journal of Medicine*.
Jarvis, B., R. F. Pearlman, S. M. Walsh, D. A. Schantz, S. Gertz et A. M. Hale-Pletka (2019). « Insurance rate optimization through driver behavior monitoring ». In : *Google Patents* 10,169,822.
Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell et S. Ermon (2016). « Combining satellite imagery and machine learning to predict poverty ». In : *Science* 353.6301, p. 790-794.
Johfre, S. S. (2020). « What Age Is in a Name ? » In : *Sociological Science* 7, p. 367-390.
Johnston, L. (1945). « Effects of tobacco smoking on health ». In : *British Medical Journal* 2.4411, p. 98.
Joly, Y., M. Braker et M. Le Huynh (2010). « Genetic discrimination in private insurance : global perspectives ». In : *New genetics and society* 29.4, p. 351-368.
Joly, Y., I. N. Feze et J. Simard (2013). « Genetic discrimination and life insurance : a systematic review of the evidence ». In : *BMC medicine* 11.1, p. 1-15.
Jung, C., S. Kannan, C. Lee, M. M. Pai, A. Roth et R. Vohra (2020). « Fair Prediction with Endogenous Behavior ». In : *arXiv* 2002.07147.

Kachur, A., E. Osin, D. Davydov, K. Shutilov et A. Novokshonov (2020). « Assessing the Big Five personality traits using real-life static facial images ». In : *Scientific reports* 10.1, p. 1-11.
Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus et Giroux.
Kamiran, F. et T. Calders (2009). « Classifying without discriminating ». In : *2009 2nd international conference on computer, control and communication*. IEEE, p. 1-6.
Kamiran, F. et T. Calders (2012). « Data preprocessing techniques for classification without discrimination ». In : *Knowledge and Information Systems* 33.1, p. 1-33.
Kamishima, T., S. Akaho, H. Asoh et J. Sakuma (2012). « Fairness-aware classifier with prejudice remover regularizer ». In : *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, p. 35-50.
Kamishima, T., S. Akaho et J. Sakuma (2011). « Fairness-aware Learning through Regularization Approach ». In : *2011 IEEE 11th International Conference on Data Mining Workshops*, p. 643-650.
Kanngiesser, P. et F. Warneken (2012). « Young children consider merit when sharing resources with others ». In : *PLOS ONE* 8.8.
Karras, T., S. Laine, M. Aittala, J. Hellsten, J. Lehtinen et T. Aila (2020). « Analyzing and improving the image quality of stylegan ». In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 8110-8119.
Karter, A. J., A. Ferrara, J. Y. Liu, H. H. Moffet, L. M. Ackerson et J. V. Selby (2002). « Ethnic disparities in diabetic complications in an insured population ». In : *Journal of the American Medical Association* 287.19, p. 2519-2527.
Kearns, M. et A. Roth (2019). *The ethical algorithm : The science of socially aware algorithm design*. Oxford University Press.
Kelly, H. (2021). « A priest's phone location data outed his private life. It could happen to anyone. » In : *The Wshinghton Post* 22-07-2021.
Kenber, B., P. Morgan-Bentley et L. Goddard (2018). « Drug prices : NHS wastes £30m a year paying too much for unlicensed drugs ». In : *Times* 26 May 2018.
Keyfitz, K., W. Flieger *et al.* (1968). *World population : an analysis of vital data.* The University of Chicago Press.
Khaitan, T. (2017). « Indirect discrimination ». In : *Handbook of the Ethics of Discrimination*. Sous la dir. de K. Lippert-Rasmussen. Routledge, p. 30-41.
Kilbertus, N., M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing et B. Schölkopf (2017). « Avoiding discrimination through causal reasoning ». In : *arXiv* 1706.02744.
Kim, M. P., O. Reingold et G. N. Rothblum (2018). « Fairness through computationally-bounded awareness ». In : *arXiv preprint arXiv:1803.03239*.
Kim, P. (2017). « Auditing algorithms for discrimination ». In : *University of Pennsylvania Law Review* 166, p. 189.
Kimball, S. L. (1979). « Reverse sex discrimination : Manhart ». In : *American Bar Foundation Research Journal* 4.1, p. 83-139.
King, G., M. A. Tanner et O. Rosen (2004). *Ecological inference : New methodological strategies*. Cambridge University Press.
Kita, K. et Ł. Kidziński (2019). « Google street view image of a house predicts car accident risk of its resident ». In : *arXiv* 1904.05270.
Kitchin, R. (2017). « Thinking critically about and researching algorithms ». In : *Information, communication & society* 20.1, p. 14-29.
Kiviat, B. (2019). « The moral limits of predictive practices : The case of credit-based insurance scores ». In : *American Sociological Review* 84.6, p. 1134-1158.

Klein, B. D. (1997). « How do actuaries use data containing errors ?: models of error detection and error correction ». In : *Information Resources Management Journal (IRMJ)* 10.4, p. 27-36.
Klein, R. (2021). *Matching rate to risk : Analysis of the availability and affordability of private passenger automobile insurance*. Rapp. tech. Insurance Information Institute.
Klein, R. W. et M. F. Grace (2001). « Urban homeowners insurance markets in Texas : A search for redlining ». In : *Journal of Risk and Insurance*, p. 581-613.
Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig et S. Mullainathan (août 2017). « Human Decisions and Machine Predictions ». In : *The Quarterly Journal of Economics* 133.1, p. 237-293.
Kleinberg, J. et S. Mullainathan (2019). « Simplicity creates inequity : implications for fairness, stereotypes, and interpretability ». In : *Proceedings of the 2019 ACM Conference on Economics and Computation*, p. 807-808.
Kleinberg, J., S. Mullainathan et M. Raghavan (2016). « Inherent trade-offs in the fair determination of risk scores ». In : *arXiv* 1609.05807.
Kluger, R. (1999). *Ashes to Ashes : America's Hundred-Year Cigarette War, the Public Health, and the Unabashed Triumph of Philip Morris.*
Koetter, F., M. Blohm, J. Drawehn, M. Kochanowski, J. Goetzer, D. Graziotin et S. Wagner (2019). « Conversational agents for insurance companies : from theory to practice ». In : *International Conference on Agents and Artificial Intelligence*. Springer, p. 338-362.
Korzybski, A. (1958). *Science and sanity : An introduction to non-Aristotelian systems and general semantics*. Institute of GS.
Kosinski, M. (2021). « Facial recognition technology can expose political orientation from naturalistic facial images ». In : *Scientific reports* 11.1, p. 1-7.
Kotter-Grühn, D., A. E. Kornadt et Y. Stephan (2016). « Looking beyond chronological age : Current knowledge and future directions in the study of subjective age ». In : *Gerontology* 62.1, p. 86-93.
Kraemer, H. C., E. Stice, A. Kazdin, D. Offord et D. Kupfer (2001). « How do risk factors work together ? Mediators, moderators, and independent, overlapping, and proxy risk factors ». In : *American journal of psychiatry* 158.6, p. 848-856.
Kranzberg, M. (1986). « Technology and History:" Kranzberg's Laws" ». In : *Technology and culture* 27.3, p. 544-560.
Krenn, K. (2017). « Introduction : Markets and Classifications-Constructing Market Orders in the Digital Age. An Introduction ». In : *Historical Social Research*, p. 7-22.
Kroll, J. A., J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson et H. Yu (2017). « Accountable algorithms ». In : *University of Pennsylvania Law Review* 165, p. 633-705.
Krumm, J. (2007). « Inference Attacks on Location Tracks ». In : *Fifth International Conference on Pervasive Computing (Pervasive 2007), May 13- 16, Toronto, Ontario, Canada*.
Kugelgen, J. von, L. Gresele et B. Scholkopf (fév. 2021). « Simpson's Paradox in COVID-19 Case Fatality Rates : A Mediation Analysis of Age-Related Causal Effects ». In : *IEEE Transactions on Artificial Intelligence* 2.1, p. 18-27.
Kuhlmann, J. (2011). « La discrimination fondée sur le sexe en assurance : obsessions et fantasmes ». In : *Risques* 87, p. 17-24.
Kusner, M. J., J. Loftus, C. Russell et R. Silva (2017). « Counterfactual Fairness ». In : *Advances in Neural Information Processing Systems 30*. Sous la dir. d'I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan et R. Garnett. NIPS, p. 4066-4076.
l'Horty, Y., M. Bunel, S. Mbaye, P. Petit et L. Du Parquet (2019). « Discriminations dans l'accès à la banque et à l'assurance : Les enseignements de trois testings ». In : *Revue d'économie politique* 129.1, p. 49-78.

Ladd, H. F. (1998). « Evidence on discrimination in mortgage lending ». In : *Journal of Economic Perspectives* 12.2, p. 41-62.
Lakkaraju, H., E. Kamar, R. Caruana et J. Leskovec (2019). « Faithful and customizable explanations of black box models ». In : *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, p. 131-138.
Lancaster, R. et R. Ward (2002). *The contribution of individual factors to driving behaviour : Implications for managing work-related road safety*. HM Stationery Office.
Landes, X. (2015). « How fair is actuarial fairness ? » In : *Journal of Business Ethics* 128.3, p. 519-533.
Lang, K. et A. K.-L. Spitzer (2020). « Race Discrimination : An Economic Perspective ». In : *The Journal of Economic Perspectives* 34.2, p. 68-89.
LaPar, D. J., C. M. Bhamidipati, C. M. Mery, G. J. Stukenborg, D. R. Jones, B. D. Schirmer, I. L. Kron et G. Ailawadi (2010). « Primary payer status affects mortality for major surgical operations ». In : *Annals of surgery* 252.3, p. 544.
Lara, L. de, A. González-Sanz, N. Asher et J.-M. Loubes (2021). « Transport-based Counterfactual Models ». In : *arXiv* 2108.13025.
Larson, J., J. Angwin, L. Kirchner et S. Mattu (2017). « How we examined racial discrimination in auto insurance prices ». In : *ProPublica, April* 5.
Lasry, J. M. (2015). « La rencontre choc de l'assurance et du big data ». In : *Risques* 103, p. 19-24.
Latner, J. D. et A. J. Stunkard (2003). « Getting worse : the stigmatization of obese children ». In : *Obesity research* 11.3, p. 452-456.
Lauer, J. (2017). *Creditworthy : A History of Consumer Surveillance and Financial Identity in America*. Columbia University Press.
Laulom, S. (2012). « Égalité des sexes et primes d'assurances ». In : *Semaine sociale Lamy* 1531, p. 44-49.
Law, S., B. Paige et C. Russell (2019). « Take a Look Around : Using Street View and Satellite Images to Estimate House Prices ». In : *ACM Transactions on Intelligent Systems and Technology* 10.5.
Le Monde (2021a). « Comment une machine joue-t-elle au Cluedo ? » In : *Le Monde* 09-07-2021.
Le Monde (2021b). « La discrimination par l'accent bientôt réprimée ? Une proposition de loi adoptée jeudi à l'Assemblée ». In : *Le Monde* 26-11-2020.
Leben, D. (2020). « Normative principles for evaluating fairness in machine learning ». In : *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, p. 86-92.
LeCun, Y., Y. Bengio et G. Hinton (2015). « Deep learning ». In : *nature* 521.7553, p. 436-444.
Léon, P. R. (1993). *Précis de phonostylistique : parole et expressivité/Pierre R. Léon,..* Nathan.
Leuner, J. (2019). « A Replication Study : Machine Learning Models Are Capable of Predicting Sexual Orientation From Facial Images ». In : *arXiv* 1902.10739.
Lew, E. A. et L. Garfinkel (1979). « Variations in mortality by weight among 750,000 men and women ». In : *Journal of chronic diseases* 32.8, p. 563-576.
Li, G., E. R. Braver et L.-H. Chen (2003). « Fragility versus excessive crash involvement as determinants of high death rates per vehicle-mile of travel among older drivers ». In : *Accident Analysis & Prevention* 35.2, p. 227-235.
Liebler, C. A., S. R. Porter, L. E. Fernandez, J. M. Noon et S. R. Ennis (2017). « America's churning races : Race and ethnicity response changes between census 2000 and the 2010 census ». In : *Demography* 54.1, p. 259-284.
Liisa, H.-B. (1994). « Aging and fatal accidents in male and female drivers ». In : *Journal of Gerontology* 49.6, S286-S290.

Lindholm, M., R. Richman, A. Tsanakas et M. V. Wuthrich (2021). « Discrimination-Free Insurance Pricing ». In : *ASTIN Bulletin* 52.
Lippert-Rasmussen, K. (2007). « Nothing personal : On statistical discrimination ». In : *Journal of Political Philosophy* 15.4, p. 385-403.
Lippi-Green, R. (2012). *English with an accent : Language, ideology, and discrimination in the United States*. Routledge.
Lippmann, W. (1922). *Public opinion*. Routledge.
Lipton, Z. C. (2018). « The Mythos of Model Interpretability : In machine learning, the concept of interpretability is both important and slippery. » In : *Queue* 16.3, p. 31-57.
Litman, T. (2005). « Pay-As-You-Drive Pricing and Insurance Regulatory Objectives. » In : *Journal of Insurance Regulation* 23.3.
Löffler, M., B. Münstermann, T. Schumacher, C. Mokwa et S. Behm (2016). « Insurers need to plug into the Internet of Things–or risk falling behind ». In : *European Insurance*.
Loi, M. et M. Christen (2021). « Choosing how to discriminate : navigating ethical trade-offs in fair algorithmic design for the insurance sector ». In : *Philosophy & Technology*, p. 1-26.
Lombroso, C. (1876). *L'uomo delinquente*. Hoepli.
Lovejoy, B. (2021). « LinkedIn breach reportedly exposes data of 92% of users, including inferred salaries ». In : *9to5mac* 06/29.
Lovell, M. C. (1963). « Seasonal adjustment of economic time series and multiple regression analysis ». In : *Journal of the American Statistical Association* 58.304, p. 993-1010.
Low, L., S. King et T. Wilkie (1998). « Genetic discrimination in life insurance : empirical evidence from a cross sectional survey of genetic support groups in the United Kingdom ». In : *British Medical Journal* 317.7173, p. 1632-1635.
Lowe, R. (1991). « Genetic Testing and Insurance : Apocalypse Now ». In : *Drake Law Review* 40, p. 507.
Luong, B. T., S. Ruggieri et F. Turini (2011). « k-NN as an implementation of situation testing for discrimination discovery and prevention ». In : *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 502-510.
Lury, C. et S. Day (2019). « Algorithmic personalization as a mode of individuation ». In : *Theory, Culture & Society* 36.2, p. 17-37.
Lutton, L., A. Fan et A. Loury (2020). « Where Banks Don't Lend ». In : *WBEZ*.
MacIntyre, A. C. (1969). « Hume on 'is' and 'ought' ». In : *The is-ought question*. Springer, p. 35-50.
Macnicol, J. (2006). *Age discrimination : An historical and contemporary analysis*. Cambridge University Press.
Madras, D., E. Creager, T. Pitassi et R. Zemel (2019). « Fairness through causal awareness : Learning causal latent-variable models for biased data ». In : *Proceedings of the conference on fairness, accountability, and transparency*, p. 349-358.
Maedche, A. (2020). « Gender Bias in Chatbot Design ». In : *Chatbot Research and Design*, p. 79.
Majumder, M. A. et M. A. Rothstein (2001). « What is genetic discrimination and when and how can it be prevented ? » In : *Genetics in Medicine* 3.5, p. 354-358.
Mangel, M. et F. J. Samaniego (1984). « Abraham Wald's Work on Aircraft Survivability ». In : *Journal of the American Statistical Association* 79.386, p. 259-267.
Manning, K. J., P. P. Barco et M. T. Schultheis (2019). « Driving evaluation in older adults ». In : *Handbook on the Neuropsychology of Aging and Dementia*. Springer, p. 231-251.
Martani, A., D. Shaw et B. S. Elger (2019). « Stay fit or get bit-ethical issues in sharing health data with insurers' apps ». In : *Swiss medical weekly* 149.2526.

Martin, G. D. (1977). « Gender Discrimination in Pension Plans : Author's Reply ». In : *The Journal of Risk and Insurance* 44.1, p. 145-149.
Martin, K., R. Ricciardelli et I. Dror (2020). « How forensic mental health nurses' perspectives of their patients can bias healthcare : A qualitative review of nursing documentation ». In : *Journal of clinical nursing* 29.13-14, p. 2482-2494.
Mas, L. (2020). « A Confederate flag spotted in the window of police barracks in Paris ». In : *France 24* 10/07.
Mayberry, R. M., F. Mili et E. Ofili (2000). « Racial and ethnic differences in access to medical care ». In : *Medical Care Research and Review* 57.1_suppl, p. 108-145.
Mayer, J., P. Mutchler et J. C. Mitchell (2016). « Evaluating the privacy properties of telephone metadata ». In : *Proceedings of the National Academy of Sciences* 113.20, p. 5536-5541.
Maynard, A. (1979). « Pricing, insurance and the National Health Service ». In : *Journal of Social Policy* 8.2, p. 157-176.
Mayson, S. G. (2018). « Bias in, bias out ». In : *Yale Law Journal* 128, p. 2218.
Mazieres, A. et C. Roth (2018). « Large-scale diversity estimation through surname origin inference ». In : *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 139.1, p. 59-73.
McConlogue, N. (2021). « Discrimination on Wheels : How Big Data Uses License Plate Surveillance to Put the Breaks on Disadvantaged Drivers ». In : *WVU College of Law Research Paper (forthcoming)., Stanford Journal of Civil Rights and Civil Liberties* 18.
McDonnell, M. et D. Baxter (2019). « Chatbots and gender stereotyping ». In : *Interacting with Computers* 31.2, p. 116-121.
McFall, L. (2019). « Personalizing solidarity? The role of self-tracking in health insurance pricing ». In : *Economy and society* 48.1, p. 52-76.
McFall, L., G. Meyers et I. V. Hoyweghen (2020). « Editorial : The personalisation of insurance : Data, behaviour and innovation ». In : *Big Data & Society* 7.2.
McFall, L. et L. Moor (2018). « Who, or what, is insurtech personalizing?: Persons, prices and the historical classifications of risk ». In : *Distinktion : journal of social theory* 19.2, p. 193-213.
McKinsey (2017). « Technology, jobs and the future of work ». In : *McKinsey Global Institute*.
McNemar, Q. (1947). « Note on the sampling error of the difference between correlated proportions or percentages ». In : *Psychometrika* 12.2, p. 153-157.
McPherson, M., L. Smith-Lovin et J. M. Cook (2001). « Birds of a feather : Homophily in social networks ». In : *Annual Review of Sociology* 27.1, p. 415-444.
Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman et A. Galstyan (2021). « A survey on bias and fairness in machine learning ». In : *ACM Computing Surveys (CSUR)* 54.6, p. 1-35.
Mellin, W. D. (1957). « Work with new electronic 'brains' opens field for army math experts ». In : *The Hammond Times* 10, p. 66.
Mendoza, J. J. (2017). « Discrimination and immigration ». In : *Handbook of the Ethics of Discrimination*. Sous la dir. de K. Lippert-Rasmussen. Routledge, p. 254-263.
Menezes, C. F. et D. L. Hanson (1970). « On the theory of risk aversion ». In : *International Economic Review*, p. 481-487.
Menon, A. K. et R. C. Williamson (2018). « The cost of fairness in binary classification ». In : *Conference on Fairness, Accountability and Transparency*. PMLR, p. 107-118.
Mercat-Bruns, M. (2016). *Discrimination at Work*. University of California Press.
Mercat-Bruns, M. (2020). « Les rapports entre vieillissement et discrimination en droit : une fertilisation croisée utile sur le plan individuel et collectif ». In : *La Revue des Droits de l'Homme* 17.

Meuleners, L. B., A. Harding, A. H. Lee et M. Legge (2006). « Fragility and crash over-representation among older drivers in Western Australia ». In : *Accident Analysis & Prevention* 38.5, p. 1006-1010.
Meyer, R. B. et M. A. Rothstein (2004). « The insurer perspective ». In : *Genetics and Life Insurance*. Sous la dir. de M. A. Rothstein. MIT Cambridge Massachusets, p. 27-47.
Meyers, G. et I. Van Hoyweghen (2018). « Enacting actuarial fairness in insurance : From fair discrimination to behaviour-based fairness ». In : *Science as Culture* 27.4, p. 413-438.
Milanković, M. (1920). *Théorie mathématique des phénomènes thermiques produits par la radiation solaire*. Gauthier-Villars.
Miller, G. et D. R. Gerstein (1983). « The life expectancy of nonsmoking men and women. » In : *Public Health Reports* 98.4, p. 343.
Miller, J. M. (1988). « Genetic Testing and Insurance Classification : National Action Can Prevent Discrimination Based on the Luck of the Genetic Draw ». In : *Dickinson Law Review* 93, p. 729.
Miller, M. J., R. A. Smith et K. N. Southwood (2003). « The relationship of credit-based insurance scores to private passenger automobile insurance loss propensity ». In : *Actuarial Study, Epic Actuaries*.
Minty, D. (2016). « Price optimisation for insurance optimising price ; destroying value ». In : *Thinkpiece Chartered Insurance Institute*.
Miracle, J. M. (2016). « De-Anonymization Attack Anatomy and Analysis of Ohio Nursing Workforce Data Anonymization ». Thèse de doct. Wright State University.
Mittelstadt, B. D., P. Allo, M. Taddeo, S. Wachter et L. Floridi (2016). « The ethics of algorithms : Mapping the debate ». In : *Big Data & Society* 3.2, p. 2053951716679679.
Mittra, J. (2007). « Predictive genetic information and access to life assurance : The poverty of 'genetic exceptionalism' ». In : *BioSocieties* 2.3, p. 349-373.
Monnet, J. (2017). « Discrimination et assurance ». In : *Journal de Droit de la Santé et de l'Assurance Maladie* 16, p. 13-19.
Moor, L. et C. Lury (2018). « Price and the person : Markets, discrimination, and personhood ». In : *Journal of Cultural Economy* 11.6, p. 501-513.
Morgan, S. L. et C. Winship (2015). *Counterfactuals and causal inference*. Cambridge University Press.
Morris, D. S., D. Schwarcz et J. C. Teitelbaum (2017). « Do Credit-Based Insurance Scores Proxy for Income in Predicting Auto Claim Risk ? » In : *Journal of Empirical Legal Studies* 14.2, p. 397-423.
Morris, J. (2021). « Israeli data : How can efficacy vs. severe disease be strong when 60% of hospitalized are vaccinated ? » In : *Covid-19 Data Science* 08/22.
Morrison, E. J. (1996). « Insurance Discrimination Against Battered Women : Proposed Legislative Protections ». In : *Ind. LJ* 72, p. 259.
Morrison, P. J. (2005). « Insurance, unfair discrimination, and genetic testing ». In : *The Lancet* 366.9489, p. 877-880.
Motulsky, A. G., N. A. Holtzman, J. E. Fullarton, L. B. Andrews *et al.* (1994). *Assessing genetic risks : implications for health and social policy*. T. 1. National Academies Press.
Moulin, H. (2004). *Fair division and collective welfare*. MIT press.
Muir, J. M. (1957). « Principles and Practices in Connection with Classification Ratings Systems For Liability Insurance as applied to Private Passenger Automobiles ». In : *Proceedings of the Casualty Actuarial Society*.
Mullins, M., C. P. Holland et M. Cunneen (2021). « Creating ethics guidelines for artificial intelligence and big data analytics customers : The case of the consumer European insurance market ». In : *Patterns* 2.10, p. 100362.

Mundubeltz-Gendron, S. (2019). « Comment l'intelligence artificielle va bouleverser le monde du travail dans l'assurance ». In : *L'Argus de l'Assurance* 10/04.
Must, A., J. Spadano, E. H. Coakley, A. E. Field, G. Colditz et W. H. Dietz (1999). « The disease burden associated with overweight and obesity ». In : *Journal of the American Medical Association* 282.16, p. 1523-1529.
Myers, R. J. (1977). « Gender discrimination in pension plans : Further comment ». In : *The Journal of Risk and Insurance* 44.1, p. 144-145.
Nelson, A. (2002). « Unequal treatment : confronting racial and ethnic disparities in health care. » In : *Journal of the national medical association* 94.8, p. 666.
Neyman, J., D. M. Dabrowska et T. Speed (1923). « On the application of probability theory to agricultural experiments. Essay on principles. Section 9. » In : *Statistical Science*, p. 465-472.
Nielsen, A. (2020). *Practical Fairness*. O'Reilly Media.
Noguéro, D. (2010). « Sélection des risques. Discrimination, assurance et protection des personnes vulnérables ». In : *Revue générale du droit des assurances* 3, p. 633-663.
Nordholm, L. A. (1980). « Beautiful patients are good patients : evidence for the physical attractiveness stereotype in first impressions of patients ». In : *Social Science & Medicine. Part A : Medical Psychology & Medical Sociology* 14.1, p. 81-83.
Norman, P. (2003). « Statistical discrimination and efficiency ». In : *The Review of Economic Studies* 70.3, p. 615-627.
Nuruzzaman, M. et O. K. Hussain (2020). « IntelliBot : A Dialogue-based chatbot for the insurance industry ». In : *Knowledge-Based Systems* 196, p. 105810.
O'Neil, C. (2016). *Weapons of math destruction : How big data increases inequality and threatens democracy*. Crown.
Obermeyer, Z., B. Powers, C. Vogeli et S. Mullainathan (2019). « Dissecting racial bias in an algorithm used to manage the health of populations ». In : *Science* 366.6464, p. 447-453.
Ohm, P. et S. Peppet (2016). « What If Everything Reveals Everything? » In : *Big Data Is Not a Monolith*. Sous la dir. de C. R. Sugimoto, H. R. Ekbia et M. Mattioli. MIT Press.
Olhede, S. C. et P. J. Wolfe (2018). « The growing ubiquity of algorithms in society : implications, impacts and innovations ». In : *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences* 376.2128, p. 20170364.
Oliver, M. et T. Shapiro (2013). *Black wealth/white wealth : A new perspective on racial inequality*. Routledge.
Ong, P. M. et M. A. Stoll (2007). « Redlining or risk? A spatial analysis of auto insurance rates in Los Angeles ». In : *Journal of Policy Analysis and Management* 26.4, p. 811-830.
Onuoha, M. (2018). « Algorithmic Violence ». In : *github. com/MimiOnuoha/*.
Orwat, C. (2020). « Risks of Discrimination through the Use of Algorithms ». In : *Institute for Technology Assessment and Systems Analysis*.
Outreville, J. F. (1990). « The economic significance of insurance markets in developing countries ». In : *Journal of Risk and Insurance*, p. 487-498.
Outreville, J. F. (1996). « Life insurance markets in developing countries ». In : *Journal of Risk and Insurance*, p. 263-278.
Owsley, C. et G. McGwin Jr (2010). « Vision and driving ». In : *Vision research* 50.23, p. 2348-2361.
Oza, D., D. Padhiyar, V. Doshi et S. Patil (2020). « Insurance claim processing using RPA along with chatbot ». In : *Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST)*.
Palmer, D. E. (2007). « Insurance, risk assessment and fairness : An ethical analysis ». In : *Insurance ethics for a more ethical world*. Emerald Group Publishing Limited.

Parléani, G. (2012). « Commentaire des lignes directrices de la Commission européenne sur les suites de l'arrêt « test Achats » ». In : 3, p. 563.
Parquet, L. du et P. Petit (2021). « Discrimination à l'embauche : retour sur deux décennies de testings en France ». In : *Revue Française d'Économie*.
Pasquale, F. (2015a). *The black box society : the secret algorithms that control money and information*. Harvard University Press.
Pasquale, F. (2015b). *The black box society : the secret algorithms that control money and information*. Harvard University Press.
Patterson, J. T. (1987). *The dread disease*. Harvard University Press.
Paugam, S., B. Cousin, C. Giorgetti et J. Naudet (2017). *Ce que les riches pensent des pauvres*. Seuil.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan kaufmann.
Pearl, J. (1998). « Graphs, causality, and structural equation models ». In : *Sociological Methods & Research* 27.2, p. 226-284.
Pearl, J. *et al.* (2000). *Models, reasoning and inference*. Cambridge University Press.
Pearl, J. (2010). « An Introduction to Causal Inference ». In : *The International Journal of Biostatistics* 6.2.
Pearl, J. et D. Mackenzie (2018). *The book of why : the new science of cause and effect*. Basic books.
Pedreshi, D., S. Ruggieri et F. Turini (2008). « Discrimination-aware data mining ». In : *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 560-568.
Peppet, S. R. (2014). « Regulating the internet of things : first steps toward managing discrimination, privacy, security and consent ». In : *Texas Law Review* 93, p. 85.
Petauton, P. (1998). « L'opération d'assurance : définitions et principes ». In : *Encyclopédie de l'assurance*. Sous la dir. de F. Ewald et J.-H. Lorenzi. Economica.
Petauton, P. (2011). « Éthique, statistique et tarification ». In : *Risques* 87, p. 25-30.
Petit, P., E. Duguet et Y. L'Horty (2015). « Discrimination résidentielle et origine ethnique : une étude expérimentale sur les serveurs en Île-de-France ». In : *Economie prevision* 1, p. 55-69.
Peyre, É. et J. Wiels (2015). *Mon corps a-t-il un sexe ?. Sur le genre, dialogues entre biologies et sciences sociales*. La découverte.
Phelps, E. S. (1972). « The statistical theory of racism and sexism ». In : *The american economic review* 62.4, p. 659-661.
Picard, P. (2003). « Les frontières de l'assurabilité ». In : *Risques* 54, p. 65-66.
Pichard, M. (2006). *Les droits à : Étude de législation française*. Economica.
Pleiss, G., M. Raghavan, F. Wu, J. Kleinberg et K. Q. Weinberger (2017). « On fairness and calibration ». In : *arXiv preprint arXiv:1709.02012*.
Poku, M. (2016). « Campbell's Law : implications for health care ». In : *Journal of health services research & policy* 21.2, p. 137-139.
Pope, D. G. et J. R. Sydnor (2011). « Implementing anti-discrimination policies in statistical profiling models ». In : *American Economic Journal : Economic Policy* 3.3, p. 206-31.
Porter, T. M. (2020). *Trust in numbers*. Princeton University Press.
Power, M. J., P. Neville, E. Devereux, A. Haynes et C. Barnes (2013). « 'Why bother seeing the world for real ?' : Google Street View and the representation of a stigmatised neighbourhood ». In : *New Media & Society* 15.7, p. 1022-1040.
Pradier, P.-C. (2011). « (Petite) histoire de la discrimination (dans les assurances) ». In : *Risques* 87, p. 51-57.

Pradier, P.-C. (2012). « Les bénéfices terrestres de la charité. Les rentes viagères des Hôpitaux parisiens, 1660-1690 ». In : *Histoire & Mesure* 26.XXVI-2, p. 31-76.
Prainsack, B. et I. Van Hoyweghen (2020). « Shifting solidarities : Personalisation in insurance and medicine ». In : *Shifting Solidarities*. Springer, p. 127-151.
Price, W. N. et I. G. Cohen (2019). « Privacy in the age of medical big data ». In : *Nature medicine* 25.1, p. 37-43.
Prince, A. E. et D. Schwarcz (2019). « Proxy discrimination in the age of artificial intelligence and big data ». In : *Iowa Law Review* 105, p. 1257.
Puddifoot, K. (2021). *How stereotypes deceive us*. Oxford University Press.
Puhl, R. et K. Brownell (2001). « Bias, discrimination, and obesity ». In : *Obesity research* 9.12, p. 788-805.
Puhl, R. M., T. Andreyeva et K. D. Brownell (2008). « Perceptions of weight discrimination : prevalence and comparison to race and gender discrimination in America ». In : *International journal of obesity* 32.6, p. 992-1000.
Pujol, D., R. McKenna, S. Kuppam, M. Hay, A. Machanavajjhala et G. Miklau (2020). « Fair decision making using privacy-protected data ». In : *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, p. 189-199.
Quesenberry, C. P., B. Caan et A. Jacobson (1998). « Obesity, health services use, and health care costs among members of a health maintenance organization ». In : *Archives of internal medicine* 158.5, p. 466-472.
Quetelet, A. (1846). *Lettres sur la théorie des probabilités, appliquée aux sciences morales et politiques*. Hayez.
Rajkomar, A., M. Hardt, M. D. Howell, G. Corrado et M. H. Chin (2018). « Ensuring Fairness in Machine Learning to Advance Health Equity ». In : *Annals of Internal Medicine* 169.12, p. 866-872.
Rambachan, A., J. Kleinberg, J. Ludwig et S. Mullainathan (2020). « An economic perspective on algorithmic fairness ». In : *AEA Papers and Proceedings*. T. 110, p. 91-95.
Rattani, A., N. Reddy et R. Derakhshani (2017). « Gender prediction from mobile ocular images : A feasibility study ». In : *IEEE International Symposium on Technologies for Homeland Security (HST)*. IEEE, p. 1-6.
Rattani, A., N. Reddy et R. Derakhshani (2018). « Convolutional neural networks for gender prediction from smartphone-based ocular images ». In : *IET Biometrics* 7.5, p. 423-430.
Ravid, I. et A. Haim (2021). « Progressive Algorithms ». In.
Rawls, J. (1999). *A theory of justice : Revised edition*. Harvard University Press.
Rebert, L. et I. Van Hoyweghen (2015). « The right to underwrite gender : The goods & services directive and the politics of insurance pricing ». In : *Tijdschrift Voor Genderstudies* 18.4, p. 413-431.
Reijns, T., R. Weurding et J. Schaffers (2021). « Ethical artificial intelligence – The Dutch insurance industry makes it a mandate ». In : *KPMG Insights* 03/2021.
Rescher, N. (2013). « How wide is the gap between facts and values ? » In : *Studies in Value Theory*. De Gruyter, p. 25-52.
Rhodes, N. G. et P. J. Savill (1984). « Smoker v. Non-Smoker ». In : *Journal of the Staple Inn Actuarial Society* 27.1, p. 1-29.
Riach, P. A. et J. Rich (1991). « Measuring Discrimination by Direct Experimental Methods : Seeking Gunsmoke ». In : *Journal of Post Keynesian Economics* 14.2, p. 143-150.
Ribeiro, M. T., S. Singh et C. Guestrin (2016). « "Why Should I Trust You ?" : Explaining the Predictions of Any Classifier ». In : *arXiv* 1602.04938.

Ribeiro, M. T., S. Singh et C. Guestrin (2017). « Model-agnostic interpretability of machine learning ». In : *arXiv* 1606.05386.
Rice, S. A. (1926). « " Stereotypes" : a source of error in judging human character. » In : *Journal of Personnel Research*.
Ricœur, P. (1991). « Le juste entre le légal et le bon ». In : *Esprit* 174, p. 5-21.
Robinson, W. S. (1950). « Ecological correlations and the behavior of individuals ». In : *American Sociological Review* 15.3, p. 351-357.
Rodríguez Cardona, D., A. Janssen, N. Guhr, M. H. Breitner et J. Milde (2021). « A Matter of Trust ? Examination of Chatbot Usage in Insurance Business ». In : *Proceedings of the 54th Hawaii International Conference on System Sciences*, p. 556.
Roemer, J. E. (1996). *Theories of distributive justice*. Harvard University Press.
Roemer, J. E. (1998). *Equality of opportunity*. Harvard University Press.
Roemer, J. E. (2002). « Egalitarianism against the veil of ignorance ». In : *The Journal of philosophy* 99.4, p. 167-184.
Rose, T. (2016). *The end of average : How to succeed in a world that values sameness*. Penguin UK.
Rosenbaum, P. (2018). *Observation and experiment*. Harvard University Press.
Ross, S. M. (2014). *Introduction to probability models*. Academic press.
Ross, S. L. et J. Yinger (1999). « Does Discrimination in Mortgage Lending Exist ? The Boston Fed Study and Its Critics ». In : *Mortgage lending discrimination : A review of existing evidence*. Urban Institute, p. 43-84.
Rothschild-Elyassi, G., J. Koehler et J. Simon (2018). « Actuarial Justice ». In : *The Handbook of Social Control*. John Wiley & Sons, Ltd. Chap. 14, p. 194-206. isbn : 9781119372394.
Rothstein, M. A. (1992). « Discrimination based on genetic information ». In : *Jurimetrics* 33.1, p. 13-18.
Royal, A. et M. Walls (2019). « Flood risk perceptions and insurance choice : Do decisions in the floodplain reflect overoptimism ? » In : *Risk Analysis* 39.5, p. 1088-1104.
Rubin, D. B. (1974). « Estimating causal effects of treatments in randomized and nonrandomized studies. » In : *Journal of educational Psychology* 66.5, p. 688.
Rubinstein, A. (2012). *Economic fables*. Open book publishers.
Rubinstein, Y. et D. Brenner (2014). « Pride and prejudice : Using ethnic-sounding names and inter-ethnic marriages to identify labour market discrimination ». In : *Review of Economic Studies* 81.1, p. 389-425.
Rudin, C. (2019). « Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead ». In : *Nature Machine Intelligence* 1.5, p. 206-215.
Rudin, C. et J. Radin (2019). « Why Are We Using Black Box Models in AI When We Don't Need To ? A Lesson From An Explainable AI Competition ». In : *Harvard Data Science Review* 1.2. https://hdsr.mitpress.mit.edu/pub/f9kuryi8.
Ruillier, J. (2004). *Quatre petits coins de rien du tout*. Bilboquet.
Rule, N. O. et N. Ambady (2010). « Democrats and Republicans can be differentiated from their faces ». In : *PloS one* 5.1, e8733.
Rundle, A. G., M. D. Bader, C. A. Richards, K. M. Neckerman et J. O. Teitler (2011). « Using Google Street View to audit neighborhood environments ». In : *American journal of preventive medicine* 40.1, p. 94-100.
Schauer, F. (2006). *Profiles, probabilities, and stereotypes*. Harvard University Press.
Schauer, F. (2017). « Statistical (and non-statistical) discrimination ». In : *Handbook of the Ethics of Discrimination*. Sous la dir. de K. Lippert-Rasmussen. Routledge, p. 42-53.

Schlesinger, A., K. P. O'Hara et A. S. Taylor (2018). « Let's talk about race : Identity, chatbots, and AI ». In : *Proceedings of the 2018 chi conference on human factors in computing systems*, p. 1-14.
Schmeiser, H., T. Störmer et J. Wagner (2014). « Unisex insurance pricing : consumers' perception and market implications ». In : *The Geneva Papers on Risk and Insurance-Issues and Practice* 39.2, p. 322-350.
Schouten, G. (2017). « Discrimination and gender ». In : *Handbook of the Ethics of Discrimination*. Sous la dir. de K. Lippert-Rasmussen. Routledge, p. 185-195.
Schweik, S. M. (2009). *The ugly laws*. New York University Press.
Seelye, K. Q. (1994). « Insurability for Battered Women ». In : *New York Times* May 12.
Seicshnaydre, S. E. (2007). « Is the Road to Disparate Impact Paved with Good Intentions : Stuck on State of Mind in Antidiscrimination Law ». In : *Wake Forest L. Rev.* 42, p. 1141.
Selbst, A. D. et S. Barocas (2018). « The intuitive appeal of explainable machines ». In : *Fordham Law Review* 87, p. 1085.
Seligman, D. (1983). « Insurance and the Price of Sex ». In : *Fortune* February 21st.
Seresinhe, C. I., T. Preis et H. S. Moat (2017). « Using deep learning to quantify the beauty of outdoor places ». In : *Royal Society open science* 4.7, p. 170170.
Shapo, N. et M. S. Masar III (2020). « Modern Regulatory Frameworks for the Use of Genetic and Epigenetic Underwriting Technology in Life Insurance. » In : *Journal of Insurance Regulation* 39.10.
Shikhare, S. (2021). « Next Generation LTC - Life insurance Underwriting using Facial Score Model ». In : *Insurance Data Science conference*.
Shin, P. (2017). « Discrimination and race ». In : *Handbook of the Ethics of Discrimination*. Sous la dir. de K. Lippert-Rasmussen. Routledge, p. 196-206.
Siegelman, P. (2003). « Adverse selection in insurance markets : an exaggerated threat ». In : *Yale Law Journal* 113, p. 1223.
Singer, P. (2011). *Practical ethics*. Cambridge university press.
Slovic, P. (1987). « Perception of risk ». In : *Science* 236.4799, p. 280-285.
Smith, D. J. (1977). *Racial disadvantage in Britain : the PEP report*. Harmondsworth : Penguin.
Smith, G. C. et J. P. Pell (2003). « Parachute use to prevent death and major trauma related to gravitational challenge : systematic review of randomised controlled trials ». In : *Bmj* 327.7429, p. 1459-1461.
Smith, R. et M. DeLair (1999). « New Evidence from Lender Testing : Discrimination at the Pre-Application Stage ». In : *Mortgage lending discrimination : A review of existing evidence*. Urban Institute, p. 23-42.
Sokolova, M., N. Japkowicz et S. Szpakowicz (2006). « Beyond accuracy, F-score and ROC : a family of discriminant measures for performance evaluation ». In : *Australasian joint conference on artificial intelligence*. Springer, p. 1015-1021.
Spirtes, P., C. Glymour et R. Scheines (1993). « Discovery algorithms for causally sufficient structures ». In : *Causation, prediction, and search*. Springer, p. 103-162.
Squires, G. D. (2003). « Racial profiling, insurance style : Insurance redlining and the uneven development of metropolitan areas ». In : *Journal of Urban Affairs* 25.4, p. 391-410.
Squires, G. D. et J. Chadwick (2006). « Linguistic profiling : A continuing tradition of discrimination in the home insurance industry ? » In : *Urban Affairs Review* 41.3, p. 400-415.
Stahlecker, P. et G. Trenkler (1993). « Some further results on the use of proxy variables in prediction ». In : *The Review of Economics and Statistics*, p. 707-711.

Steensma, C., L. Loukine, H. Orpana, E. Lo, B. Choi, C. Waters et S. Martel (2013). « Comparing life expectancy and health-adjusted life expectancy by body mass index category in adult Canadians : a descriptive study ». In : *Population health metrics* 11.1, p. 1-12.
Stein, A. (1994). « Will Health Care Reform Protect Victims of Abuse-Treating Domestic Violence as a Public health Issue ». In : *Human Rights* 21, p. 16.
Stenholm, S., J. Head, V. Aalto, M. Kivimäki, I. Kawachi, M. Zins, M. Goldberg, L. G. Platts, P. Zaninotto, L. M. Hanson *et al.* (2017). « Body mass index as a predictor of healthy and disease-free life expectancy between ages 50 and 75 : a multicohort study ». In : *International journal of obesity* 41.5, p. 769-775.
Stephan, Y., A. R. Sutin et A. Terracciano (2015). « How old do you feel ? The role of age discrimination and biological aging in subjective age ». In : *PloS one* 10.3, e0119293.
Stevenson, M. (2018). « Assessing risk assessment in action ». In : *Minnesota Law Review* 103, p. 303.
Stiglitz, J. E. et A. Weiss (1981). « Credit rationing in markets with imperfect information ». In : *The American economic review* 71.3, p. 393-410.
Stolley, P. D. (1991). « When genius errs : RA Fisher and the lung cancer controversy ». In : *American Journal of Epidemiology* 133.5, p. 416-425.
Stone, D. A. (1993). « The struggle for the soul of health insurance ». In : *Journal of Health Politics, Policy and Law* 18.2, p. 287-317.
Strandburg, K. J. (2019). « Rulemaking and inscrutable automated decision tools ». In : *Columbia Law Review* 119.7, p. 1851-1886.
Struyck, N. (1740). « Inleiding tot de Algemeene Geographie ». In : *Tirion* 1740, p. 231.
Struyck, N. (1912). *Les oeuvres de Nicolas Struyck (1687-1769) : qui se rapportent au calcul des chances, à la statistique général, à la statistique des décès et aux rentes viagèter*. Société générale néerlandaise d'assurances sur la vie et de rentes viagères.
Sunstein, C. R. (2019). « Algorithms, correcting biases ». In : *Social Research : An International Quarterly* 86.2, p. 499-511.
Suresh, H. et J. V. Guttag (2019). « A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle ». In : *arXiv* 1901.10002.
Sutin, A. R. et A. Terracciano (2013). « Perceived weight discrimination and obesity ». In : *PloS one* 8.7, e70048.
Swedloff, R. (2014). « Risk classification's big data (R)evolution ». In : *Connecticut Insurance Law Journal* 21, p. 339.
Sweeney, L. (2013). « Discrimination in online ad delivery : Google ads, black names and white names, racial discrimination, and click advertising ». In : *Queue* 11.3, p. 10-29.
Taskesen, B., J. Blanchet, D. Kuhn et V. A. Nguyen (2021). « A statistical test for probabilistic fairness ». In : *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, p. 648-665.
Tendil, C. (1998). « La tarification de l'offre : techniques et problèmes ». In : *Encyclopédie de l'assuranc*e. Sous la dir. de F. Ewald et J.-H. Lorenzi. Economica.
Tene, O. et J. Polonetsky (2017). « Taming the Golem : Challenges of ethical algorithmic decision-making ». In : *North Carolina Journal of Law & Technology.* 19, p. 125.
Thiery, Y. et C. Van Schoubroeck (2006). « Fairness and equality in insurance classification ». In : *The Geneva Papers on Risk and Insurance-Issues and Practice* 31.2, p. 190-211.
Thomas, G. (2017). *Loss coverage : Why insurance works better with some adverse selection*. Cambridge University Press.

Thomas, R. G. (2007). « Some novel perspectives on risk classification ». In : *The Geneva Papers on Risk and Insurance-Issues and Practice* 32.1, p. 105-132.
Thomsen, F. K. (2017). « Direct discrimination ». In : *Handbook of the Ethics of Discrimination*. Sous la dir. de K. Lippert-Rasmussen. Routledge, p. 19-29.
Thomson, J. J. (1976). « Killing, letting die, and the trolley problem ». In : *The Monist* 59.2, p. 204-217.
Thornton, S. M., S. Pan, S. M. Erlien et J. C. Gerdes (2016). « Incorporating ethical considerations into automated vehicle control ». In : *IEEE Transactions on Intelligent Transportation Systems* 18.6, p. 1429-1439.
Tian, J. et J. Pearl (2002). « A general identification condition for causal effects ». In : *Proceedings of the Eighteenth National Conference on Artificial Intelligence*. MIT Press, p. 567-573.
Tobler, C. (2008). *Limits and potential of the concept of indirect discrimination*. Office for Official Publications of the European Communities.
Torous, W., F. Gunsilius et P. Rigollet (2021). « An Optimal Transport Approach to Causal Inference ». In : *arXiv* 2108.05858.
Tribalat, M. (2016). *Statistiques ethniques, Une querelle bien française*. L'Artilleur.
Tsamados, A., N. Aggarwal, J. Cowls, J. Morley, H. Roberts, M. Taddeo et L. Floridi (2021). « The ethics of algorithms : key problems and solutions ». In : *AI & Society*, p. 1-16.
Tuppat, J. et J. Gerhards (2021). « Immigrants' First Names and Perceived Discrimination : A Contribution to Understanding the Integration Paradox ». In : *European Sociological Review* 37.1, p. 121-135.
Turner, M. A. et F. Skidmore (1999). « Introduction, summary and recommendations ». In : *Mortgage lending discrimination : A review of existing evidence*, p. 1-22.
Tzioumis, K. (2018). « Demographic aspects of first names ». In : *Scientific data* 5.1, p. 1-9.
Uotinen, V., T. Rantanen et T. Suutama (2005). « Perceived age as a predictor of old age mortality : a 13-year prospective study ». In : *Age and Ageing* 34.4, p. 368-372.
Upton, G. et I. Cook (2014). *A dictionary of statistics 3e*. Oxford university press.
Valfort, M.-A. (2017). « La religion, facteur de discrimination à l'embauche en France ? » In : *Revue économique* 68.5, p. 895-907.
Van Hoyweghen, I., K. Horstman et R. Schepers (2007). « Genetic 'risk carriers' and lifestyle 'risk takers'. Which risks deserve our legal protection in insurance ? » In : *Health Care Analysis* 15.3, p. 179-193.
Van Lancker, W. (2020). « Automating the Welfare State : Consequences and Challenges for the Organisation of Solidarity ». In : *Shifting Solidarities*. Springer, p. 153-173.
Van Parijs, P. (2002). « Linguistic justice ». In : *Politics, Philosophy & Economics* 1.1, p. 59-74.
Vandenbroucke, J. P. et G. W. Comstock (1989). « Those who were wrong ». In : *American journal of epidemiology* 130.1, p. 3-5.
Verma, S. et J. Rubin (2018). « Fairness definitions explained ». In : *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, p. 1-7.
Villani, C. (2009). *Optimal transport : old and new*. T. 338. Springer.
Villazor, R. C. (2008). « Blood quantum land laws and the race versus political identity dilemma ». In : *Californial Law Review* 96, p. 801.
Vogel, R., A. Bellet, S. Clémen *et al.* (2021). « Learning Fair Scoring Functions : Bipartite Ranking under ROC-based Fairness Constraints ». In : *International Conference on Artificial Intelligence and Statistics*. PMLR, p. 784-792.
Voicu, I. (2018). « Using first name information to improve race and ethnicity classification ». In : *Statistics and Public Policy* 5.1, p. 1-13.

Volpp, K. G., A. B. Troxel, M. V. Pauly, H. A. Glick, A. Puig, D. A. Asch, R. Galvin, J. Zhu, F. Wan, J. DeGuzman *et al.* (2009). « A randomized, controlled trial of financial incentives for smoking cessation ». In : *New England Journal of Medecine* 360, p. 699-709.

Wachter, S., B. Mittelstadt et C. Russell (2020). « Why Fairness Cannot Be Automated : Bridging the Gap Between EU Non-Discrimination Law and AI ». In : *ArXiv* 2005.05906.

Wagner, C., M. Strohmaier, A. Olteanu, E. Kıcıman, N. Contractor et T. Eliassi-Rad (2021). « Measuring algorithmically infused societies ». In : *Nature*, p. 1-6.

Wang, Y. et M. Kosinski (2018). « Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. » In : *Journal of personality and social psychology* 114.2, p. 246.

Weir, J. M. et J. E. Dunn Jr (1970). « Smoking and mortality : a prospective study ». In : *Cancer* 25.1, p. 105-112.

Weld, G., E. Jang, A. Li, A. Zeng, K. Heimerl et J. E. Froehlich (2019). « Deep learning for automatically detecting sidewalk accessibility problems using streetscape imagery ». In : *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, p. 196-209.

Westreich, D. (2012). « Berkson's bias, selection bias, and missing data ». In : *Epidemiology* 23.1, p. 159.

Wiehl, D. G. (1960). *Build and Blood Pressure*. Society of Actuaries.

Wiggins, B. (2013). *Managing Risk, Managing Race : Racialized Actuarial Science in the United States, 1881–1948*. University of Minnesota PhD thesis.

Wiggins, B. (2020). *Calculating race : Racial discrimination in risk assessment*. Oxford University Press.

Williams, W. E. (2013). *Race & economics : How much can be blamed on discrimination ?* Hoover Press.

Winlaw, M., S. H. Steiner, R. J. MacKay et A. R. Hilal (2019). « Using telematics data to find risky driver behaviour ». In : *Accident Analysis & Prevention* 131, p. 131-136.

Wissoker, D., W. Zimmermann, G. Galster, K. Hartnett, R. Smith, C. Herbig, M. McDonough et J. Silver (1998). *Testing for discrimination in home insurance*. Urban Institute Washington.

Wolff, M. J. (2006). « The myth of the actuary : life insurance and Frederick L. Hoffman's race traits and tendencies of the American negro ». In : *Public Health Reports* 121.1, p. 84-91.

Wolffhechel, K., J. Fagertun, U. P. Jacobsen, W. Majewski, A. S. Hemmingsen, C. L. Larsen, S. K. Lorentzen et H. Jarmer (2014). « Interpretation of appearance : The effect of facial features on first impressions and personality ». In : *PloS one* 9.9, e107721.

Works, R. (1977). « Whatever's FAIR-Adequacy, Equity, and the Underwriting Prerogative in Property Insurance Markets ». In : *Nebraska Law Review* 56, p. 445.

Wortham, L. (1985). « Insurance classification : too important to be left to the actuaries ». In : *University of Michigan Journal Law Reform* 19, p. 349.

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

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

Yamamoto, T., J. Hashiji et V. N. Shankar (2008). « Underreporting in traffic accident data, bias in parameters and the structure of injury severity models ». In : *Accident Analysis & Prevention* 40.4, p. 1320-1329.

Yong, E. (2018). « A popular algorithm is no better at predicting crimes than random people ». In : *The Atlantic* 17-01.

Young, R. K., A. H. Kennedy, A. Newhouse, P. Browne et D. Thiessen (1993). « The effects of names on perception of intelligence, popularity, and competence ». In : *Journal of Applied Social Psychology* 23.21, p. 1770-1788.

Yusuf, S., S. Hawken, S. Ounpuu, L. Bautista, M. G. Franzosi, P. Commerford, C. C. Lang, Z. Rumboldt, C. L. Onen, L. Lisheng *et al.* (2005). « Obesity and the risk of myocardial infarction in 27 000 participants from 52 countries : a case-control study ». In : *The Lancet* 366.9497, p. 1640-1649.

Zafar, M. B., I. Valera, M. G. Rodriguez et K. P. Gummadi (2017). « Fairness Constraints : Mechanisms for Fair Classification ». In : *arXiv* 1507.05259.

Zarsky, T. Z. (2014). « Understanding discrimination in the scored society ». In : *Washington Law Review* 89, p. 1375.

Zelizer, V. A. R. (2018). *Morals and markets*. Columbia University Press.

Zhang, B. H., B. Lemoine et M. Mitchell (2018). « Mitigating Unwanted Biases with Adversarial Learning ». In : *arXiv* 1801.07593.

Zhang, J. et E. Bareinboim (2018). « Fairness in decision-making—the causal explanation formula ». In : *Thirty-Second AAAI Conference on Artificial Intelligence*.

Žliobaite, I. (2015). « On the relation between accuracy and fairness in binary classification ». In : *arXiv* 1505.05723.

Žliobaite, I. et B. Custers (2016). « Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models ». In : *Artificial Intelligence and Law* 24.2, p. 183-201.

Žliobaitė, I. (2017). « Measuring discrimination in algorithmic decision making ». In : *Data Mining and Knowledge Discovery* 31.4, p. 1060-1089.

Zuboff, S. (2019). *The age of surveillance capitalism : The fight for a human future at the new frontier of power*. Public Affairs.

Zweig, J. (2010). « High Trading Is Bad News For Investors ». In : *Wall Street Journal* February 13.

Institut Louis Bachelier
Palais Brongniart
28, place de la Bourse
75002 Paris
Tél. : +33 (0)1 73 01 93 40
Fax : +33 (0)1 73 01 93 28
contact@institutlouisbachelier.org

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE DE L'ÉCONOMIE
DE L'INDUSTRIE ET DE L'EMPLOI

LABEX
Louis Bachelier