

La statistique dans la cité n° 22 - décembre 2020

Lettre bimestrielle du groupe « Statistique et enjeux publics »

Sommaire du n°22 – décembre 2020

Éditorial

Méthodes :

- *La France de plus en plus urbaine*
- *Requiem pour les aires urbaines*
- *La fiabilité des sondages pré-électoraux aux USA est-elle avérée ?*
- *Prévisions électorales : trois métiers*

Outils :

- *R_0 ? R_0 ? vous avez bien dit R_0 ? Comme c'est bizarR₀ ?*
- *L'enquête de recensement de 2021 sera reportée*

Vie des institutions :

- *Suite du feuilleton sur la plateforme des données de santé (PDS)*
- *La saga du recensement de 2020 aux USA (suite et, sans doute, pas fin)*

Notes de lecture :

- *Efficacité d'un vaccin*
- *Covid-19 et minorités ethniques*
- *La statistique publique face à la crise*
- *Comment départager deux candidats à une élection « serrée » ?*
- *Quelles données pour comprendre la propagation du SARS-CoV2 ?*
- *Témoignage à propos du recensement 2020 de la population aux USA*

Annonces :

- *11^e Colloque international francophone sur les sondages*
- *Actualités du groupe Statistique et Enjeux publics*

Tous les numéros de l'infolettre sont téléchargeables sur le site du groupe SEP ([cliquez ICI](#))

Éditorial

C'est la quatrième fois que notre lettre paraît depuis l'apparition de la Covid19 en Europe. Depuis dix mois, nous nous sommes efforcés de la faire paraître aux dates prévues ; maintenir le lien avec les participants à nos traditionnels *Cafés de la Statistique* et avec les lecteurs de *La Statistique dans la cité* a été une priorité. Nous avons repris ce mois de décembre les séances des Cafés sous une forme adaptée aux contraintes que nous impose la situation sanitaire. Nous espérons ainsi continuer à répondre à la question au centre de la devise de notre groupe *Statistique et Enjeux publics* et de nos Cafés : *La statistique éclaire-t-elle les questions de société ?* Cet objectif aura-t-il été atteint dans ce contexte si particulier de la crise sanitaire ?

Dans ce numéro, la rédaction s'est interrogée, à partir de notes de lecture, sur certains aspects de la statistique face à la pandémie, et notamment sur les indicateurs d'efficacité des vaccins qui vont jouer un rôle essentiel au cours du premier semestre de l'année 2021.

Par ailleurs, les deux mois écoulés ont vu à nouveau, à l'occasion des élections du 3 novembre aux États-Unis, les sondages pré-électoraux mis sur la sellette. Deux articles de ce numéro essaient de faire la part des choses et d'évoquer la responsabilité des sondeurs, mais aussi des médias et des utilisateurs ; de nouveaux intervenants (agrégateurs et prévisionnistes) sont venus éclairer un exercice indispensable à la vie démocratique. Mais toujours outre-Atlantique, il faut constater que la saga du recensement américain de la population de 2020 se poursuit et que des pressions politiques

sans précédent dans ce pays continuent d'être exercées sur les statisticiens publics ; cela démontre, s'il en était encore besoin, que la production de statistiques fiables et crédibles n'est pas possible sans le respect de règles éthiques, telles que celles définies par exemple par la *Résolution sur les principes fondamentaux de la statistique officielle* adoptée à l'unanimité par l'Assemblée générale des Nations-Unies. Cette saga prouve que les pays démocratiques ne sont pas à l'abri de la violation de ces règles.

On trouvera aussi dans ce numéro deux articles consacrés aux nouveaux outils d'analyse géographique des villes utilisés par l'Insee. Enfin un point est fait sur les problèmes que pose l'hébergement par un serveur non-européen de la plateforme des données de santé.

Dans l'espoir que 2021 verra le retour à de meilleures conditions de vie et que, en conséquence, les activités du groupe *Statistique et Enjeux publics* pourront reprendre normalement dans le courant de cette année, nous voudrions souhaiter bonne année à tous nos lecteurs. Et surtout n'hésitez pas à nous faire part de vos commentaires et de vos attentes en écrivant à sep@sfds.asso.fr
Prenez soin de vous !

Méthodes

La France de plus en plus urbaine

Le Monde, dans son édition datée du mercredi 28 octobre, a présenté sous le titre « *Une France de plus en plus urbaine* » deux études de l'Insee montrant que les Français sont de plus en plus nombreux à résider dans une zone urbaine ou située sous l'influence d'une ville.

D'une part, l'Insee a diffusé des données actualisées ⁽¹⁾ sur les unités urbaines, zones dont la définition ⁽²⁾ n'a pas changé, mais qui ont été mises à jour depuis le dernier découpage de 2010. D'autre part, en s'appuyant sur le nouvel outil « Zonage en aires d'attraction des villes », l'Institut a publié simultanément des résultats au niveau national ⁽³⁾ et des déclinaisons au niveau régional.

Ce nouveau dispositif n'est pas une surprise pour nos lecteurs ; *La statistique dans la cité*, dans ses numéros 18 et 19 de février et avril 2020, a consacré deux articles à la présentation des nouveaux outils d'analyse géographique des villes. En 2020 le zonage en aires d'attraction des villes ⁽⁴⁾ se substitue au zonage en aires urbaines (ZAU) utilisé antérieurement. Le nouveau zonage repose sur une méthode harmonisée avec celle utilisée par Eurostat et l'OCDE, ce qui facilite les comparaisons internationales et permet de visualiser l'influence en France des grandes villes étrangères proches du territoire français. Cette approche fonctionnelle de la ville aide à analyser les disparités territoriales selon la taille de l'aire et la distinction entre centre et périphérie. On compte ainsi en 2017 52,9 millions d'habitants dans les unités urbaines, soit 2,8 millions de résidents de plus que dix ans auparavant : cette augmentation résulte d'une part de la croissance de la population des communes appartenant à une unité urbaine et d'autre part de l'intégration de nouvelles communes au fort dynamisme démographique.

Par ailleurs, l'Insee a délimité sur le territoire français 699 aires d'attraction d'une ville, dans lesquelles réside 93 % de la population. Près d'un habitant sur cinq vit dans la seule aire d'attraction de Paris qui compte 13 millions d'habitants. Près de 20 % de la population également réside dans les treize autres aires comptant plus de 700 000 habitants : Lyon, Marseille-Aix-en-Provence, Lille, Toulouse, Bordeaux, Genève-Annemasse, Nantes, Luxembourg, Strasbourg, Sarrebruck, Montpellier, Rennes et Grenoble ; on notera que pour trois d'entre elles la commune-centre se trouve à l'étranger.

Le nombre important des naissances explique l'essentiel de la croissance démographique de ces très grandes aires, au sein desquelles on constate des migrations importantes de la commune-centre vers les communes périphériques. La croissance démographique s'avère nettement plus modérée dans les aires intermédiaires (entre 200 000 et 700 000 habitants), alors que les plus petites (moins de 50 000 habitants) se dépeuplent dans certains cas.

Ces résultats reposent sur des données antérieures à la Covid-19. Il sera intéressant de suivre l'évolution de ces phénomènes migratoires en lien avec un développement potentiellement durable du télétravail.

(1) *Insee Focus n° 210, le 21 octobre 2020*

(2) *Une unité urbaine regroupe une ou plusieurs communes partageant une même zone de bâti continu comptant au moins 2 000 habitants*

(3) *Insee Focus n° 211, le 21 octobre 2020*

(4) *L'aire d'attraction d'une ville illustre l'étendue de son influence sur les communes environnantes. Une aire est composée d'un pôle, défini à partir de critères de population et d'emploi, et d'une couronne, constituée des communes dont au moins 15 % des actifs travaillent dans le pôle. Au sein du pôle, la commune la plus peuplée est appelée commune-centre*

Requiem pour les aires urbaines

Le zonage en aires urbaines est né en 1995. Auparavant existaient depuis 1962 les « Zones de peuplement industriel ou urbain » : mais ce zonage a paru inutilisable après le recensement de 1990 car les ZPIU avaient fini par concerner plus de 96 % de la population. En 2020, les aires urbaines disparaissent à leur tour, remplacées par les « aires d'attraction des villes » (cf. l'article ci-dessus). Elles auront donc vécu 25 ans.

Leur nom de baptême était un peu trompeur, car ces aires contenaient beaucoup de petites communes verdoyantes ayant conservé leur aspect rural. Mais d'un autre côté, c'était bien là le point ! Même à la campagne, on est désormais aussi un peu en ville quand on peut se rendre dans un centre en quelques minutes pour travailler, faire des courses, utiliser un équipement public,... La nouvelle terminologie dit cela mieux, et réserve la qualification de « urbaines » aux zones denses de bâti continu, ce qui est plus conforme à la pratique courante.

Curieusement, les nouvelles « aires d'attraction des villes » s'exposent d'emblée à être critiquées comme les anciennes ZPIU, puisqu'elles concernent déjà 93 % de la population. Mais les temps ont changé...

Adieu les aires urbaines !

La fiabilité des sondages pré-électoraux aux USA est-elle avérée ?

Comme il y a quatre ans, à l'occasion des élections du 3 novembre dernier aux États-Unis, la fiabilité des instituts de sondage a été mise en doute. Dans un article du n° 2 de décembre 2016, *La Statistique dans la cité* avait conclu que, contrairement aux idées reçues, ce n'étaient pas les sondeurs qui avaient annoncé trop hâtivement la probable victoire d'Hillary Clinton, mais les commentateurs, les médias et les spécialistes de sciences politiques.

Cette année encore, il a été reproché aux sondeurs d'avoir surestimé les scores en faveur de Joe Biden et surtout en faveur du parti démocrate pour l'élection des représentants qui avait lieu le même jour. Ce reproche n'est pas sans fondement, même si le résultat final a bien été celui que les intentions de vote laissaient attendre ; mais les scores ont été beaucoup plus serrés. Pour ne prendre que l'exemple du total des votes au niveau fédéral, celui de l'écart entre les deux candidats, les intentions de vote annonçaient un écart souvent de l'ordre de 8 % en faveur de Joe Biden et, finalement, cet écart n'aura été que de 4,4 %. Il semble bien que la mobilisation d'électeurs du parti républicain ait été sous-estimée, soit qu'ils aient été insuffisamment représentés dans les échantillons des sondages, soit que les électeurs sondés aient refusé de déclarer leur véritable choix, comme cela a parfois été supposé en France il y a quelques années pour expliquer les sous-estimations du vote FN. Il y a eu manifestement, en particulier dans les "swing states", une sous-estimation du vote Trump même si cela n'a pas conduit cette année à inverser les résultats par rapport aux prévisions ; mais par contre pour ce qui est de l'élection des représentants, le vote républicain qui avait été sous-estimé a conduit, contrairement aux prévisions, à une forte diminution de la majorité démocrate à la Chambre des représentants par rapport à l'élection de 2018.

La discussion sur les méthodes et les résultats des instituts de sondage doit donc se poursuivre : elle est légitime et saine dans une démocratie !

Prévisions électorales : trois métiers

Les résultats des récentes élections aux États-Unis ont beaucoup surpris de part et d'autre de l'Atlantique. Pourtant, une armée d'organismes américains avait tenté d'éclairer a priori cet événement. On peut les classer en trois groupes : les instituts de sondage, les agrégateurs et les prévisionnistes.

Les instituts de sondage ont réalisé de très nombreuses enquêtes d'intention de vote auprès des futurs électeurs, au niveau fédéral et encore plus au niveau de chacun des cinquante États, ce qui est nécessaire compte tenu des modalités très particulières du scrutin aux USA ; les sondages ont été particulièrement nombreux dans les États les plus disputés (les fameux « swing states »). Face à cette profusion, interviennent les agrégateurs de sondages, comme « RealClearPolitics » ou « 270 to Win ⁽⁵⁾ ». En première analyse, les agrégateurs font la moyenne des sondages nationaux et la publient.

Mais ceux qui reçoivent désormais le plus d'attention, ce sont les prévisionnistes, qui avancent un pronostic sur le résultat de l'élection, au niveau national et surtout État par État. Cette catégorie se subdivise elle-même en trois : les experts politiques, qui font de la prévision « au doigt mouillé » - par exemple l'organe de presse « Politico » ; les marchés de paris sur les candidats qui affichent une « cote » de chacun à partir des paris qu'ils reçoivent (« PredictIt ») ; et enfin les modélisateurs

statisticiens. Ces derniers tentent de synthétiser de nombreuses informations pour en déduire des prédictions : informations sur les élections passées, sondages évidemment, mais aussi variables contextuelles socio-démographiques (sur le corps électoral) ou économiques (dont on sait l'influence sur l'élection). Un statisticien américain, Nate Silver, s'est acquis une forte réputation par la qualité de ses prédictions lors d'élections précédentes ; son site (et son entreprise), « FiveThirtyEight ⁽⁶⁾ », est très suivi. Une autre équipe de statisticiens a été constituée pour cette élection par *The Economist* autour d'Andrew Gelman, professeur à l'université Columbia reconnu dans la communauté des statisticiens américains. Gelman est un bayésien, spécialiste des « modèles multiniveaux » ; il plaide sur son blog en faveur de l'emploi des méthodes statistiques dans les sciences sociales et humaines, avec un accent particulier pour promouvoir une transparence complète sur les données et sur les méthodes.

Bien sûr, les frontières entre ces diverses catégories sont poreuses. Les instituts de sondage utilisent aussi des modèles statistiques, notamment pour « redresser » leurs résultats bruts à partir de données issues des élections précédentes.

En France, nous connaissons les instituts de sondage depuis plus de quatre-vingts ans ; les agrégateurs ont fait leur apparition au moment des dernières élections présidentielles. Mais notre pays n'a pas (encore) donné la vedette à des prévisionnistes professionnels ; les paris en ligne sur les candidats ne sont pas autorisés en France et, si des statisticiens de valeur se sont déjà attachés à prédire les résultats des élections en liaison avec des instituts de sondage, leurs travaux sont jusqu'à présent restés relativement confidentiels, sans doute parce que la demande ne s'est pas encore exprimée. Mais cela ne devrait certainement pas tarder !

(5) 270 est le nombre minimum de grands électeurs qu'il faut gagner pour être élu

(6) 538 est le nombre des grands électeurs

Outils

R_0 ? R_0 ? vous avez bien dit R_0 ? Comme c'est bizar R_0 ?

Rapportant au quotidien les dispositions prises par les pouvoirs publics, les médias ont, au fil de la Covid-19, popularisé un indicateur « R » jusque là mal connu des non-spécialistes. Il s'agit du « taux de reproduction » : nombre moyen de personnes qu'une personne contaminée contamine à son tour. S'il est supérieur à 1, le nombre de malades croît de façon exponentielle et l'épidémie explose. Si au contraire R est inférieur à 1, on a une décroissance exponentielle et l'épidémie s'éteint d'elle-même. Dès lors, pouvoir jouer sur cette valeur est essentiel pour contenir l'épidémie. La valeur initiale R_0 au démarrage de l'épidémie dépend des caractéristiques de l'agent infectieux mais aussi de diverses circonstances physiques et du comportement des gens. Une politique sanitaire va tenter de jouer sur celui-ci : quarantaine, distanciation, « gestes barrière », etc. Pour les épidémiologistes, le problème n'était pas nouveau.

Une équipe anglaise ⁽⁷⁾ avait, en 2013, proposé une méthode qui repose sur l'hypothèse que l'« histoire moyenne » de la maladie reste à peu près stable dans le temps (même si l'histoire individuelle diffère d'un sujet à l'autre). Dès lors, un événement (par exemple l'apparition de certains symptômes de degré repérable) présente le même profil temporel que l'infestation, avec un décalage de plusieurs jours mais constant. Bien sûr, il faut que le symptôme servant de signal soit observé de façon suffisamment homogène et stable également. De la même façon, le résultat peut différer suivant ce que l'on a retenu comme signal pour dénombrer les cas : ainsi, nous avons actuellement, en France, deux estimations de R : l'une fondée sur les diagnostics confirmés, l'autre sur les hospitalisations. Mais, dans les faits, elles demeurent voisines et cela suffit pour éclairer convenablement la politique menée.

La méthode a été testée sur les archives de cinq grandes épidémies historiques, ce qui semble en confirmer sa robustesse. Les auteurs ont développé un outil prêt à l'emploi (sous Microsoft Excel) disponible en ligne : il peut être mis en œuvre par des personnels sans compétence statistique ou épidémiologique particulière, soit un avantage notable, de coût, de délai et de ressource humaine.

(7) Voir Estimate Time-Varying Reproduction Numbers During Epidemics, Anne Cori et al., Imperial College London, *American Journal of Epidemiology*, Vol. 178, No. 9, September 15, 2013

L'enquête de recensement de 2021 sera reportée

L'évolution de la situation sanitaire conduit l'Insee à reporter à 2022 la prochaine enquête annuelle de recensement ; un communiqué de presse a été publié sur le site de l'Insee le 27 novembre 2020. En effet, la collecte sur le terrain, devant débiter le 21 janvier 2021, aurait entraîné de nombreux

déplacements et contacts avec les habitants, difficilement compatibles avec le contexte sanitaire. Néanmoins, comme chaque année, l'Insee, en adaptant les méthodes de calcul et en mobilisant davantage les sources administratives, publiera les populations légales de la France et de toutes ses communes fin 2021.

On peut par ailleurs signaler dans l'édition datée du 18 novembre du quotidien *Le Monde* un article plutôt élogieux (*Avec la Covid-19, l'Insee en terre inconnue*) sur l'adaptation contrainte des statisticiens et la capacité d'innovation de l'Insee dans ces circonstances difficile, ainsi que deux séances organisées récemment dans le cadre des Jéco⁽⁸⁾. Rappelons que le n° 20 (juin 2020) de *La Statistique dans la cité* proposait déjà un article sur les modalités d'adaptation du service statistique pour fournir des informations démographiques et économiques fiables face au choc subi par la société lors du premier confinement.

(8) Voir ci-dessous dans les notes de lecture

Vie des institutions

Suite du feuilleton sur la plateforme des données de santé (PDS)

Le précédent numéro de notre lettre évoquait les mésaventures juridiques de la PDS dite aussi *Health Data Hub* (HDH) concernant le choix de confier à *Microsoft Azure* la gestion des opérations d'hébergement, d'appariement et de mise à disposition des données de santé. L'invalidation le 16 juillet dernier par la Cour de justice de l'Union européenne du *Privacy Shield* (« bouclier de confidentialité »), a rendu illégaux en effet les transferts de données sensibles de l'Europe vers les États-Unis.

À ce sujet, *Mediapart* a révélé dans un article du 22 novembre :

- un courrier du 10 novembre de la présidente de la Cnil demandant au gouvernement que les données de santé à caractère personnel ne soient plus confiées à des entreprises soumises à la loi des États-Unis, tout en admettant une période de transition aussi courte que possible ;
- la réponse du ministre Olivier Véran où il écrit « qu'il n'existe pas à court terme de solution optimale du point de vue technique » mais que le gouvernement s'engage à faire migrer les données chez un opérateur de droit européen dans un délai de deux ans.

Mais au-delà du risque présenté par le *nuage* états-unien, a-t-on vraiment besoin du dernier cri des prestations en nuage et pour quoi faire ? C'est en substance le sujet d'une tribune des épidémiologistes Marcel Goldberg et Marie Zins dans *Le Monde* daté du 30 octobre.

Rappelons que le HDH n'est plus seulement comme le prévoyait la loi de janvier 2016 un guichet unique chargé d'instruire les demandes de traitement de données de santé à caractère personnel pour le compte des demandeurs et de la Cnil. Allant au-delà des préconisations du rapport Villani sur l'Intelligence artificielle (IA) de mars 2018 (pp. 200 et suivantes⁽⁹⁾), la loi du 24 juillet 2019 a fait de l'ex *Institut national des données de santé* une plateforme centralisant et mettant à disposition elle-même via un opérateur sous contrat l'ensemble des données de santé financées par la collectivité :

- non seulement le Système national des données de santé « historique » (SNDS) : données de l'assurance maladie + données d'activité médicale servant de base au financement des hôpitaux + causes de décès, le tout pseudonymisé et chaîné de sorte qu'on puisse suivre les données d'une personne sans que son identité figure dans la base ;
- mais aussi les dossiers médicaux (dossiers hospitaliers en particulier), les données d'enquêtes (et de cohortes) lorsqu'elles sont appariées avec les données du SNDS historique, les données de prévention, les données de la PMI, des Maisons départementales des personnes handicapées (MDPH), etc.

L'intention est louable : il manquait dans le SNDS historique des informations utiles pour comprendre les maladies ou évaluer les thérapies (e.g. données génomiques, antécédents familiaux et autres facteurs de risque comme masse corporelle ou consommation de tabac) ; or la Cnil s'opposait à ce que des données extérieures au SNDS lui fussent appariées de manière pérenne en vue de faciliter des études futures.

Mais la loi de 2019 et plus encore son *projet de décret d'application* n'ont-elles pas tordu le bâton dans l'autre sens en confiant à la PDS la mission de réunir toutes ces bases en une seule ? Goldberg et Zins font observer que la plupart des données ainsi rassemblées seraient hétérogènes en qualité et sur le plan sémantique, donc de peu d'utilité pour développer des algorithmes d'Intelligence artificielle. Un effort d'harmonisation (interopérabilité) afin que certaines informations soient structurées et

définies de la même manière permettra des croisements ultérieurs mais cela n'exige pas que ces données soient toutes hébergées sur une plateforme nationale unique.

La logique d'une plateforme unique (« regroupons et conservons : ça peut toujours servir ») s'autodétruit : pourquoi alors se limiter aux données de santé, ou aux données publiques, ou aux données françaises ? Bien sûr, il faut rendre possibles des études médico-économiques ou des projets de recherche sur données internationales ; bien sûr il peut être avantageux de disposer d'un opérateur de référence dans un pays (c'est le rôle que joue aujourd'hui le *Centre d'accès sécurisé aux données* pour l'Insee, le CNRS, la DGFiP et d'autres producteurs de données, y compris de santé) mais cela suppose surtout des dictionnaires de données partagés et des capacités de rapprochement et de traitement sécurisés, non une plateforme totalisante donnant aux algorithmes d'intelligence artificielle un terrain de jeu illimité sans considération pour la confidentialité ni l'autonomie des producteurs de données ni la liberté des personnes concernées. La loi de juillet 2019 a posé en principe que les données de santé financées sur fonds publics doivent pouvoir être réutilisées librement à des fins d'intérêt public mais en pratique, il est nécessaire que les producteurs de données aient un intérêt non seulement à accéder aux données des autres mais aussi à mettre en forme les leurs pour les partager avec d'autres... L'incitation en cette matière fonctionne mieux que l'autorité : la PDS ne devrait-elle pas, comme le suggèrent Goldberg et Zins « se concentrer sur des activités réellement utiles comme par exemple une cartographie analytique des bases de données disponibles, leur mise en réseau, la promotion de règles harmonisées de partage des données, etc. » ?

(9) [aiforhumanity.fr/pdfs/9782111457089 Rapport Villani accessible.pdf](http://aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf)

La saga du recensement de 2020 aux USA ⁽¹⁰⁾ (suite et, sans doute, pas fin)

Le n° 10 de *La Statistique dans la Cité* (juin 2018) avait fait état de la polémique créée par l'initiative du Secrétaire américain au commerce, Wilbur Ross, qui avait décidé, au printemps 2018, d'introduire, pour la première fois depuis 1950, une question sur la citoyenneté : toutes les personnes recensées auraient dû déclarer si elles avaient ou non la citoyenneté américaine. Beaucoup de voix s'élevaient pour contester l'opportunité de ce changement, en estimant que l'inclusion de cette question provoquerait une augmentation du taux de non-réponse au recensement parmi ceux qui ne sont pas citoyens américains. Le n° 16 (octobre 2018) avait relaté les suites judiciaires données à cette demande, qui s'étaient notamment traduites par un refus de la Cour suprême de donner suite à cette demande de l'administration Trump.

Enfin le n° 21 (octobre 2020) de *La Statistique dans la cité* faisait état du mémorandum présidentiel du 21 juillet 2020 donnant l'ordre au *Bureau of the Census* de retrancher les effectifs des sans-papiers, dont le nombre est estimé à environ 10 millions aux États-Unis, de la population résultant du recensement. D'autres demandes qui auraient été de nature à biaiser les résultats accompagnaient ce mémorandum.

En fait, l'objectif visé par l'administration Trump est de ne plus caler la répartition des sièges à la Chambre des représentants sur la totalité de la population résidente, en expliquant ne plus vouloir « donner une représentation parlementaire à des étrangers qui entrent ou restent illégalement dans ce pays ». Il n'est évidemment pas question dans cet article de discuter du bien-fondé de cet objectif. Mais, comme le quatorzième amendement prévoit que « les représentants seront répartis entre les divers États proportionnellement à leur population respective, calculée en comptant tous les habitants de chaque État » et qu'il est très difficile de modifier la Constitution (tout amendement doit être ratifié par les trois-quarts des États), l'administration Trump s'est attaquée à la définition de la population dans le recensement. Plusieurs États, dont celui de New York, qui abritent de nombreux migrants, ont rapidement saisi la justice pour maintenir le statu quo et obtenu des victoires en première instance. L'administration Trump a alors demandé à la Cour suprême d'intervenir en urgence, puisque le Président est censé transmettre début janvier au Congrès les résultats du recensement de 2020 et le nombre de sièges attribués à chaque État. Celle-ci s'est réunie le 30 novembre et rendra sa décision avant le 1er janvier ; mais il semble que les juges pourraient botter en touche, en se bornant à des questions de procédure.

(10) Voir aussi ci-dessous une note de lecture sur un témoignage

Notes de lecture

Lu dans le blog ⁽¹¹⁾ de discussions quotidiennes entre les membres de l'American Statistical Association (ASA Connect) : qu'est-ce que l'efficacité d'un vaccin ? En phase 3 des tests, on compare le nombre de diagnostiqués positifs dans un groupe ayant été vacciné et dans un groupe de

même taille ayant reçu un placebo. Cette comparaison suppose qu'au total des deux groupes, un nombre minimum de testés positifs ait été atteint. Par exemple, avec le vaccin Moderna, 196 patients ont été testés positifs parmi lesquels on a compté 11 positifs dans le groupe vacciné et 185 dans le groupe placebo. L'efficacité du vaccin est réputée être de $(185 - 11)/185$, soit 94,1 %

(11) Résumé écrit par la rédaction

Lu dans *Variances*, le Web-Magazine d'Ensaë Alumni, daté du 19 novembre 2020 : un article sur les contaminations et les décès dus à la Covid-19 au miroir de l'ethnicité : une plus forte contamination et une plus forte mortalité dues au Covid-19 ont été constatées parmi les minorités ethniques, comme les Noirs, ou issues de migrations du Sud dans plusieurs pays occidentaux. En France, partant de ce constat, plusieurs analyses ont mis en évidence une prépondérance de facteurs sociaux plutôt qu'ethniques dans ces surcroûts de contaminations.

Vu dans le compte-rendu des Journées de l'Economie 2020 : La Fondation pour l'Université de Lyon organise à Lyon chaque année depuis 2008 les *Journées de l'Economie (Jéco)*. Cet événement réunit des économistes, des chefs d'entreprise, des experts de l'administration, des journalistes, des acteurs sociaux et des responsables politiques. Les Jéco 2020 ont été organisées cette année du 17 au 19 novembre uniquement sous forme de visioconférences en raison des mesures de confinement. Deux de ces conférences traitent des instruments pour mesurer la croissance et des outils utilisés pour pallier les lacunes dans la collecte des statistiques habituelles dues à la pandémie, « *Les statistiques face à la crise* » et « *Google en sait-il plus que l'Insee ?* ». Ces deux conférences sont, comme toutes les autres, accessibles sur le site des Jéco :

<http://www.journeeseconomie.org/index.php>.

Lu dans *Le Monde* daté du 18 novembre 2020 : Comment départager deux candidats à une élection « serrée » ? Dans son numéro daté du 18 novembre 2020, le journal *Le Monde* ouvre ses colonnes à une tribune d'Étienne Ghys, secrétaire perpétuel de l'Académie des sciences, intitulée « *Départager deux candidats ? Si simple à la majorité simple* ». Il y montre que, dans une élection serrée entre deux candidats, une erreur aléatoire très faible dans le dépouillement (de l'ordre de un pour dix-mille) a une probabilité non négligeable (plusieurs pour cents) d'inverser le résultat souhaité par les électeurs. Si l'élection est à deux niveaux, comme dans le cas de l'élection présidentielle aux États-Unis où l'on vote d'abord par État, avant de procéder au vote final, cette probabilité augmente. Elle est d'autant plus élevée que le nombre de niveaux est grand. Il conclut que, finalement, le type d'élection le moins sensible aux « bruits » représentés par ces inévitables erreurs de dépouillement est le vote direct à la majorité simple.

Lu dans *Le Monde* daté du 1er décembre 2020 : une tribune d'un chercheur au CNRS et à l'École d'économie de Paris ; l'auteur constate que, si l'on dispose de données statistiques en quantité suffisante pour une approche épidémiologique de la Covid-19, nous manquons de données sur l'environnement et la situation socio-économique des personnes positives au SARS-CoV2 ; ces données seraient selon lui essentielles pour comprendre comment se propage le virus.

Lu dans le magazine mensuel *The Atlantic* ⁽¹²⁾ : un témoignage plutôt préoccupant écrit par un agent du *Bureau of the Census* ayant participé aux opérations de collecte du recensement 2020 de la population aux USA. En introduction de son article, il écrit (traduction assurée par la rédaction) :

Il est vrai que chaque décompte décennal rencontre des défis spécifiques.

Le Président George Washington, dans une lettre adressée au Gouverneur Morris, un autre des « pères fondateurs », s'était plaint que le premier recensement, mené en 1790 par des officiers de police à cheval (« US marshalls ») s'était traduit par un sous-dénombrement très important, ce qui constituait une menace potentielle pour la sécurité nationale. L'exploitation des données du recensement de 1880 a pris près de huit ans. Le Congrès a été amené à rejeter dans leur totalité les résultats du recensement de 1920.

Mais les problèmes auxquels fait face le recensement de 2020 sont d'une ampleur historique. Même avant que les opérations de collecte n'aient commencé, l'administration Trump avait exigé que le recensement comprenne une question sur la citoyenneté, ce qui avait déclenché une bataille juridique d'une année à la suite de laquelle la Cour suprême a finalement donné raison au Bureau du recensement qui avait averti que les taux de réponse diminueraient probablement si une telle question était posée ⁽¹³⁾. D'autres problèmes ⁽¹⁴⁾, notamment une politisation sans précédent dans les nominations à des

postes de responsabilité, des catastrophes naturelles importantes, des troubles civils, une insuffisance de financement, un taux de rotation élevé des employés, des changements constants dans les délais annoncés et des litiges presque incessants ont finalement miné le projet de recensement. Il faut bien sûr ajouter au tableau la pandémie.

L'article complet peut être lu à travers le lien

<https://www.theatlantic.com/politics/archive/2020/11/failure-2020-census-trump-pandemic/617188/>.

(12) *The Atlantic* est aussi un media multi-plateforme

(13) Voir l'article paru dans le n° 10 (juin 2018) de *La Statistique dans la Cité*

(14) Voir les articles parus dans le n° 16 (octobre 2019) et dans le n° 21 (octobre 2020) de *La Statistique dans la Cité*

Annonces

11e Colloque international francophone sur les sondages

Le 11ème Colloque International Francophone sur les Sondages, qui devait initialement avoir lieu en octobre 2020, se tiendra du mercredi 6 au vendredi 8 octobre 2021 à l'Université Libre de Bruxelles. Le mardi 5 octobre 2021 sera consacré à des ateliers de formation. Pour plus d'information (inscriptions en ligne, soumission d'une communication, ...), consulter le site du Colloque :

<https://sondages2020.sciencesconf.org/>

Actualités du groupe Statistique et Enjeux publics

En 2020, seules trois séances des *Cafés de la statistique* ont pu se tenir les 14 janvier, 10 février et 11 mars avant la première période de confinement. Les contraintes sanitaires ont ensuite conduit le groupe SEP à suspendre les séances, tout en maintenant la publication régulière de *La Statistique dans la Cité*. Devant l'impossibilité durable d'organiser les Cafés en présentiel au café du Pont-Neuf, le bureau du groupe a décidé de tenir quelques séances virtuelles en visioconférence. La première séance de ce type s'est déroulée le 8 décembre. Elle portait sur les inégalités salariales entre les femmes et les hommes et était introduite par une présentation de Nila Ceci-Renaud ⁽¹⁵⁾, département des salaires et conventions salariales à la Dares. Malgré quelques problèmes de connexion, cette séance a été un succès, puisqu'elle a regroupé 78 participantes et participants et que le flot des remarques, questions et interactions entre les participants a été tout à fait satisfaisant. Comme à l'accoutumée, la séance a donné lieu à un enregistrement vidéo qui sera prochainement mis en ligne ⁽¹⁶⁾. Une prochaine séance en visioconférence aura lieu le mardi 19 janvier sur le thème de l'utilisation des indicateurs pour la gestion de la pandémie ; elle sera introduite par Dominique Costagliola ⁽¹⁷⁾, épidémiologiste et biostatisticienne à l'Inserm.

(15) Nila Ceci-Renaud aurait du normalement intervenir le 19 décembre 2019, puis le 22 avril 2020. Ces deux séances avaient dû être annulées, d'abord en raison des grèves dans les transports, puis à cause du confinement

(16) Le lien permettant d'accéder aux comptes-rendus et aux vidéos des Cafés est :

https://www.sfds.asso.fr/fr/statistique_et_enjeux_publics/499-les_cafes_de_la_statistique/

(17) Le Monde daté du 9 décembre 2020 a consacré à Dominique Costagliola un Portrait à l'occasion du Grand Prix de l'Inserm qui lui a été décerné pour l'année 2020

Responsable de l'infolettre : Chantal Cases, présidente du groupe SEP

Rédacteur en chef : Jean-Louis Bodin

Secrétaire de rédaction : Jean-Pierre Le Gléau

Webmestre : Érik Zolotoukhine