

7 Avril 2020

Remarque préliminaire

La présente newsletter de la Société Luxembourgeoise de Statistique exprime l'opinion personnelle de son président. La situation actuelle fait qu'une prise de position concertée de la Société aurait peut-être pris du temps supplémentaire, alors qu'une prise de position rapide nous a paru nécessaire.

COVID-19 – La statistique prise au dépourvue ?

La crise que nous vivons actuellement est d'abord une crise sanitaire, mais pas seulement. Les conséquences se font et se feront sentir à travers tout l'édifice socio-économique. Un nouveau virus a fait son apparition fin 2019 et il continue à se propager sous forme de pandémie. Les compteurs tournent et tous les jours, ou même plusieurs fois par jour, nous sommes inondés de leurs enregistrements : nombre de personnes testées positives, nombre de personnes hospitalisées, en soins intensifs, en réanimation et hélas aussi, nombre de personnes décédées.

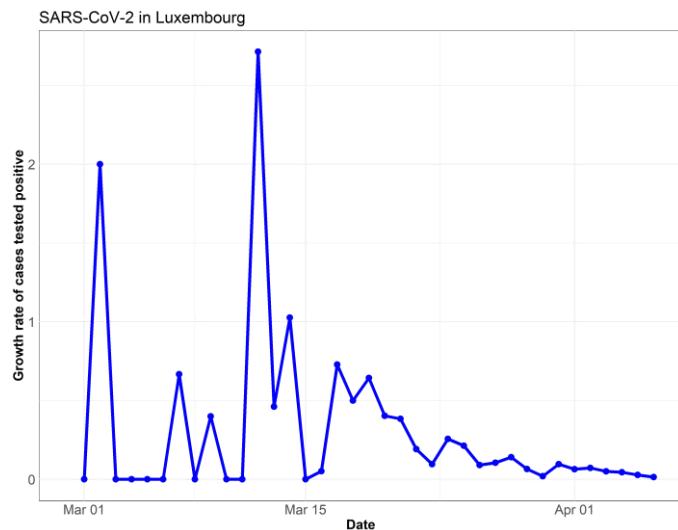
Et la statistique dans tout cela, en général et ici au Luxembourg ? Avouons-le, elle a été prise de court. Les statisticiens ont souvent l'habitude de réfléchir longtemps pour savoir ce qu'il faut observer, comment il faut observer, comment mesurer les faits observés, comment et par quelles techniques établir des statistiques fiables et de qualité. La statistique veut cerner la réalité le mieux et le plus précisément possible et cela comporte en général des délais.

Rien de tout cela avec le COVID-19. Les nombres sont tombés, ils ont été de suite accaparés par les médias et on semble n'avoir rien demandé aux statisticiens de service. Certes, le monde de la recherche luxembourgeois n'a pas dormi. Une [Task Force](#) avec plus de douze groupes de travail a été créée et la statistique y est fort présente. Le STATEC a réagi aussi rapidement, tout particulièrement pour évaluer les conséquences économiques et sur son site un [dossier](#) COVID-19 a été créé. Avec la propagation de la pandémie, les gouvernements ont été appelés à agir en vitesse en s'appuyant certes sur l'expertise et les modèles d'épidémiologistes, comme ceux du professeur Neil Ferguson de l'Imperial College de Londres.

Le monde des nombres absolus

La statistique, ou du moins la statistique publique, revendique pour elle de vouloir fournir des bases scientifiques sur lesquelles construire des décisions (« evidence based decision making » est le terme consacré en anglais). Il faut bien avouer que les décideurs ont cette fois-ci dû prendre des décisions sur base de statistiques finalement assez précaires, voire rudimentaires. Les compteurs tournent toujours, comme ceux du [worldometer](#) ou de [l'université John Hopkins](#), et même à [Luxembourg](#), mais de préférence en termes absolus.

Certes, l'évolution de la pandémie en nombres absolus est importante, tout particulièrement pour le monde hospitalier qui doit pouvoir essayer d'ajuster ses capacités à différents niveaux pour ne pas être débordé. Reconnaissions qu'à Luxembourg d'importants efforts y ont été consacrés avec succès et même si le monde médical dans son ensemble est extrêmement sollicité, on a pu éviter des situations comme dans d'autres pays ou régions plus près de chez nous. La communication du Premier ministre et du Ministre de la Santé s'est aussi efforcée d'être plus différenciée au fil des jours, distinguant par exemple entre résidents et non-résidents. Mais pour pouvoir apprécier l'évolution de la propagation on aurait pu faire mieux. Il aurait été approprié de donner par exemple un tableau résumant l'évolution des statistiques journalières, voire de présenter un graphique de l'évolution du taux de croissance journalier comme on a essayé de le faire sur le site de la [Société Luxembourgeoise de Statistique](#).

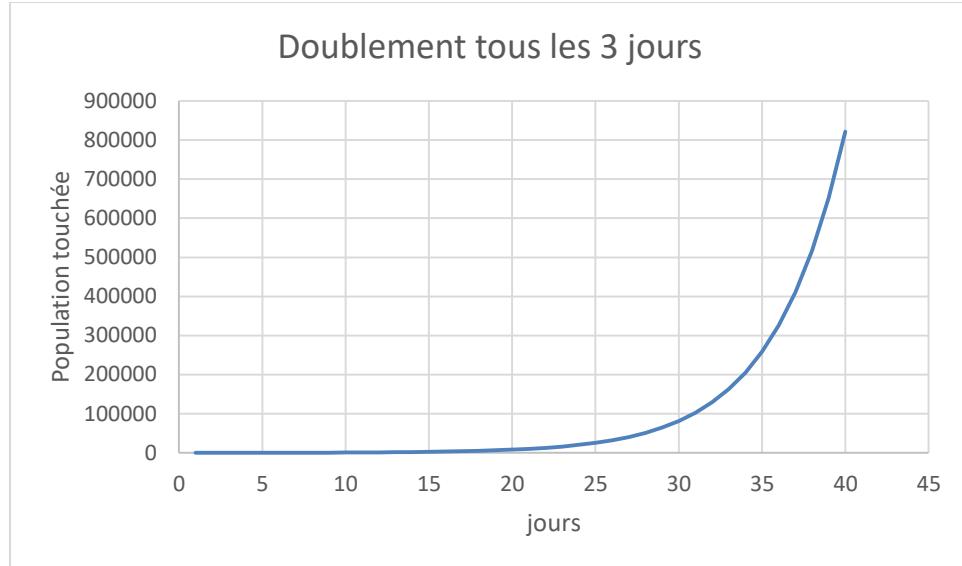


Des statistiques descriptives élémentaires sont sans doute disponibles, mais il faut regretter qu'elles ne soient pas publiées, car elles pourraient fournir une meilleure information (même graphique) sur la distribution par exemple par âge. Les statisticiens auraient pu expliquer davantage et ne point se borner à laisser cette tâche aux politiciens. Tout le monde a probablement une idée de ce que représente une moyenne. Il n'en va pas de même pour la médiane. Si l'âge médian des personnes décédées est de 86 ans, cela veut dire que la moitié des personnes décédées avaient plus de 86 ans et l'autre moitié moins de 86 ans. La médiane est moins sensible aux extrêmes, tout particulièrement quand on a peu d'observations. Si parmi ceux de moins de 86 ans il y avait malheureusement eu un enfant, cela aurait tiré la moyenne vers le bas. En cas de distribution symétrique, la moyenne et la médiane sont très proches l'une de l'autre. Pour le nombre de personnes testées positives, nous apprenons que la moyenne d'âge est de 46, mais il serait intéressant d'avoir plus d'informations sur la distribution, même sous forme graphique (par exemple sous forme de box-plot ou boîte à moustache). Il est regrettable que cette information statistique ne soit pas fournie.

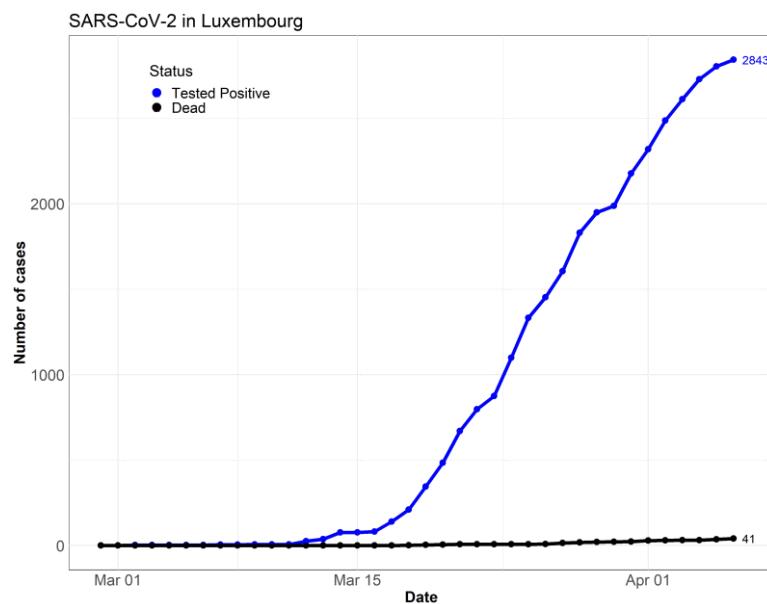
Cours de statistique et d'épidémiologie ad hoc ?

Nous venons déjà de voir quelques notions élémentaires de statistiques utiles dans l'analyse des données relatives au Covid-19. La Société Luxembourgeoise de Statistique pourrait aussi profiter de l'occasion pour essayer de familiariser le public avec ces notions élémentaires. Comme elle comprend aussi des membres du Luxembourg Institute of Health (LIH), très actifs pour l'instant – reconnaissons-le, elle pourrait établir un glossaire à l'instar de [celui](#) publié par la Royal Statistical Society, ou ne serait-ce qu'un lien vers cette page.

Les collègues du Royaume Uni mentionnent aussi la **croissance exponentielle** dont on entend parler presque tous les jours. Effectivement les épidémies ont tendance à leur début à se propager pendant une certaine période avec une croissance exponentielle. Qu'est-ce que cela veut dire ? Par exemple pour le nombre de cas testés positifs ? Cela veut dire que de jour en jour le nombre de cas croît à un certain pourcentage. Prenons comme exemple une croissance de près de 26% ($\sqrt[3]{2} - 1$ pour être exact). Si l'on part 100 de personnes testées positives on constate qu'avec une telle croissance ce nombre va doubler au bout de 3 jours (200) et 3 jours plus tard il doublera encore : 400 etc.

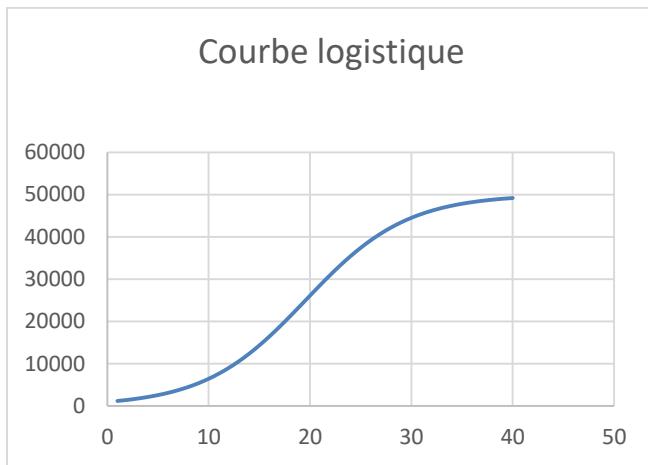


Mais bon, avec une telle croissance on observerait que l'ensemble de la population du Luxembourg serait contaminée au bout de 39 jours. Le graphique ci-dessous représente l'évolution du nombre de testés positifs et de décès à la date du 5 avril. Le site du STATEC fournit aussi un graphique similaire, mais assez curieusement on y met la date en ordonnée.



Si l'on observe bien une croissance exponentielle (ou géométrique) pendant une certaine période de la propagation, il y a en général un moment où la courbe s'infléchit. On parle alors d'un point d'inflexion et l'allure générale de la courbe prend l'allure d'une **courbe logistique**. La courbe logistique ou fonction logistique est attribuée au mathématicien belge Pierre-François Verhulst pour

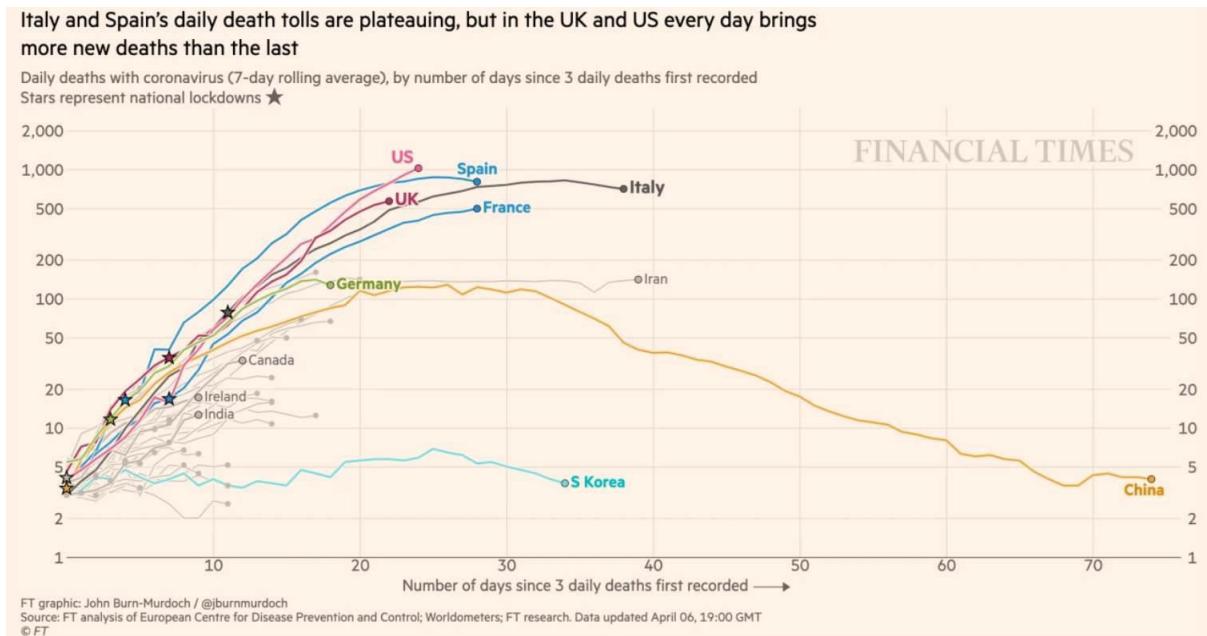
modéliser l'évolution d'une population. Cette courbe est en forme de S. On parle aussi dans ce cas d'une sigmoïde.



Le lecteur intéressé trouvera évidemment plusieurs sites sur Internet consacrés à la croissance exponentielle. Le lien [suivant](#) discute la croissance exponentielle dans quelques pays au début du mois de mars 2020. En voici un [lien](#) qui montre l'évolution dans sa perspective logistique.

En termes de graphiques on retrouve donc des allures similaires à celles que nous venons de décrire. Un exemple est celui du Coronavirus Resource Center de l'Université John Hopkins qui utilise des [moyennes mobiles](#) sur 5 jours.

Une autre approche intéressante est de faire des graphiques en échelle semi-logarithmique. De tels graphiques permettent de comparer aisément les taux de croissance donnés par la pente des courbes. Ci-après un graphique repris du Financial Times.



N'oublions pas non plus, comme le montre la courbe de la Chine, qu'après une croissance exponentielle, logistique, il est permis d'espérer que la courbe va décroître en fonction de l'efficacité des mesures prises.

Tests, sondages et décisions politiques

Une des grandes questions est évidemment de connaître à un certain moment le nombre de personnes contaminées par le virus. Le problème provient du fait que ce nombre n'est pas connu. On ne connaît que le nombre de personnes testées positives, tout comme on connaît le nombre de ceux qui sont hospitalisés, qui sont en soins intensifs, décédés ou ayant été testé positifs mais guéris entretemps. Entretemps il est clair aussi que le nombre de testés positifs varie avec le nombre de tests, et dans certains cas d'une ampleur similaire. Plus on fait de tests, plus on détecte de personnes porteuses du virus. Au Luxembourg, on constate par exemple une diminution du nombre de personnes testées positives le weekend parallèlement à une diminution du nombre de tests.

Une des priorités de la Task Force Research Luxembourg est justement d'estimer la prévalence Covid-19. La prévalence Covid-19 se veut d'établir le nombre de personnes atteintes par le virus, soit à un moment donné (prévalence instantanée) ou durant une période (prévalence sur un temps donné). Il ne faut pas confondre cette notion avec celle d'incidence, qui désigne le nombre de nouveaux cas au cours d'une période donnée.

Avec la propagation rapide du virus, les gouvernements ont néanmoins dû prendre des mesures ad hoc, sans toujours pouvoir s'appuyer sur des études statistiques et scientifiques. On peut parfois lire aujourd'hui des commentaires *a posteriori* qui donnent l'impression que l'estimation de la prévalence serait une tâche somme toute assez facile. Citons à cet égard le professeur Wolfram Meyerhöfer dans une [contribution](#) publiée le 3 avril 2020 dans le journal FAZ . Il dit, à juste titre, qu'on ne sait pas dire combien de personnes contaminées ont besoin d'une assistance médicale, car on ne fait que des tests parmi les personnes montrant des symptômes manifestes: « Dabei ist erkennbar, dass die Gefährlichkeit eines neuen Virus verlässlich abgeschätzt werden kann, indem man eine repräsentative Stichprobe der Bevölkerung testet und dann den Krankheitsverlauf der dabei positiv Getesteten verfolgt.¹ » Bref, il suffirait de tirer un échantillon représentatif et de faire les tests appropriés. L'auteur qui enseigne « Mathematikdidaktik » à l'université de Paderborn plaide pour des décisions basées sur des faits établis et appelle les politiciens à reconsiderer un certain nombre de décisions prises à la lumière de statistiques confirmées.

En réalité, ce n'est peut-être pas aussi simple. Ainsi, Gerd Bosbach, connu pour son livre « Lügen mit Zahlen » et qui avait été l'invité d'une conférence du STATEC en 2012, affirmait dans une [interview](#) du 26 mars 2020 au sujet de l'estimation de la prévalence : « Die Zahl der Infizierten ist auf jeden Fall deutlich höher, aber niemand kann sagen, um welchen Faktor. Um dies zu beantworten, bräuchten wir eine repräsentative Stichprobe aus der Bevölkerung. Das ist zur Zeit mangels Testkapazitäten in Deutschland nicht machbar. » Le manque de moyens et les priorités de traitement médical rendent donc difficilement concevable un test basé sur un échantillon représentatif. Le même Gerd Bosbach va néanmoins affirmer quelques jours plus tard dans une autre interview : « Weitreichende Entscheidungen bedürfen gesicherter Grundlagen. Genau das ist bisher vernachlässigt worden. Die wiederholte Gleichsetzung der Zahl positiv Getesteter mit der Zahl der Infizierten vernebelt den Blick, die Zählweise bei Corona-Toten ebenfalls. Statt wichtige Zahlen anhand einer repräsentativen

¹ Il est manifeste que le caractère dangereux d'un nouveau virus peut être évalué de manière fiable en testant un échantillon représentatif de la population, puis en suivant l'évolution de la maladie chez ceux qui se sont révélés positifs.

Stichprobe zu erheben, werden Daten aus einer völlig verzerrten Teststatistik übernommen². » Même si les discours de Gerd Bosbach sont instructifs et intéressants à maints égards (je vous invite à lire ses propos), on ne peut néanmoins s'étonner que ce qui paraissait impossible quelques jours auparavant, soit devenu possible en un rien de temps. A moins que la quantité des tests disponibles chez nos voisins allemands ait augmenté pour se permettre ce « luxe statistique ».

Bref, il faut peut-être se garder de vouloir paraître plus malin a posteriori. Nous sommes aujourd'hui dans une situation où la connaissance médicale du Covid-19 progresse encore de jour en jour. Tout n'est pas connu, ni des cas symptomatiques, ni des cas asymptomatiques. Une étude basée sur un échantillon tiré selon les règles de l'art n'est probablement pas aussi facile qu'on pourrait le rêver. De plus il est peut-être aussi prétentieux d'affirmer qu'en tant que statisticien, on serait aujourd'hui à la hauteur de fournir les bases solides sur lesquelles les décisions politiques pourraient s'appuyer, alors qu'on a été largement pris au dépourvu. Il y aura des décisions politiques portant sur le court et le moyen terme, d'autres concerteront le long terme. La complexité de la situation d'aujourd'hui et de celle du futur ne permettra pas de réponses simples. En ce sens il faut saluer la collaboration et les synergies entre tous les acteurs scientifiques engagés dans la Task Force et espérer que ces scientifiques, ensemble avec leurs réseaux internationaux, contribuent à fournir aux décideurs des bases solides en temps utile et que le monde politique puisse s'en inspirer pour trouver les bonnes réponses à hauteur de la complexité de la situation actuelle et future.

Nico Weydert
Président
Société Luxembourgeoise de Statistique

² Des décisions de grande envergure nécessitent des bases solides. C'est exactement ce qui a été négligé jusqu'à présent. L'équation répétée du nombre de personnes testées positives avec le nombre de personnes infectées a obscurci la vue, tout comme le comptage des décès dus au coronavirus. Au lieu de collecter des données sur la base d'un échantillon représentatif, les données sont extraites d'une statistique de test complètement biaisée.

