

La doyenne du Québec, Marie-Louise Meilleur, vécut jusqu'à l'âge de 117 ans et 230 jours. Son record sera-t-il battu ?

Peut-on faire mieux que Meilleur ?

Léo Belzile
HEC Montréal
Christian Genest
Université McGill

On appelle *supercentenaire* une personne ayant dépassé l'âge de 110 ans. Au Québec, on en a recensé à ce jour une vingtaine, toutes des femmes. Au sommet de cette liste, présentée au tableau 1, trône Marie-Louise Meilleur, née à Kamouraska le 29 août 1880 et décédée à East Ferris (Ontario) le 16 avril 1998. Durant les derniers mois de sa vie, elle fut même la doyenne de l'humanité suite au décès de la Française Jeanne Calment (1875-1997). Une autre Québécoise, Julie Winnefred Bertrand, originaire de Coaticook, a également brièvement détenu le titre mondial. À ce jour, cependant, Jeanne Calment demeure l'être humain ayant vécu le plus longtemps, soit 122 ans, 5 mois et 14 jours, parmi toutes les personnes dont l'âge a fait l'objet d'une validation rigoureuse.

Au-delà de la fascination qu'exercent les records de toutes sortes, la question de la longévité humaine est un sujet d'intérêt pour tous. Aussi est-il légitime de se demander quelles sont les chances qu'au Québec, quelqu'un vive plus longtemps que Mme Meilleur. Pour répondre à cette question, il faut d'abord avoir une idée précise du nombre d'aspirants

potentiels au titre. Puis, on doit évaluer les chances qu'un de ces « concurrents » batte le record actuel, ce qui nécessite de recourir à un modèle qui permette d'extrapoler au-delà des données recueillies à ce jour.

C'est à cet exercice délicat que nous allons nous livrer ici en nous appuyant sur deux sources de données fiables : la base de données sur la longévité canadienne (BDLC) et l'*International Database on Longevity* (IDL). Ce sera l'occasion de mettre en lumière certains des enjeux conceptuels et techniques que recèle la modélisation de données démographiques.

Évolution du nombre de centenaires depuis 100 ans

Depuis le milieu du 20^e siècle, le nombre de centenaires dans la population mondiale ne cesse de croître. C'est aussi le cas chez nous, comme l'illustre la figure 1 qui montre l'évolution du nombre (gauche) et de la proportion (droite) de centenaires nés au Québec entre 1921 et 2022. Ces deux courbes, extraites de la banque de données BDLC, s'avèrent quasi linéaires (sur

Tableau 1 : Liste des 20 supercentenaires les plus âgés nés au Québec (un astérisque indique que cette observation fait partie de l'IDL).

Marie-Louise Meilleur (1880-1998)	117 ans, 230 jours
Julie Winnefred Bertrand* (1891-2007)	115 ans, 124 jours
Cecile Klein (1907-2022)	114 ans, 212 jours
Hazel Campbell (1903-2016)	112 ans, 336 jours
Alexina Saint-Pierre Loyer (1907-2020)	112 ans, 183 jours
Yvonne Delorme-Martel (1909-2022)	112 ans, 167 jours
Marie-Laure Nadon* (1889-2001)	112 ans, 128 jours
Julia Houde* (1894-2006)	112 ans, 61 jours
Simonne Leduc (1911-2023)	111 ans, 360 jours
Marie-Blanche Portelance (1909-2021)	111 ans, 349 jours
Colombe Benoît-Leclerc (1903-2015)	111 ans, 343 jours
Marie-Louise Meunier* (1887-1998)	111 ans, 223 jours
Eva Carrier-Delisle (1908-2019)	111 ans, 207 jours
Émilienne Marcil-Pigeon (1911-2022)	111 ans, 204 jours
Marie Bibeault* (1882-1993)	111 ans, 179 jours
Gisèle Chalifour (1907-2018)	111 ans, 85 jours
Linda Yale (1912-2024)	111 ans, 74 jours
Louise-Anna Turcotte-Brouillette (1905-2016)	111 ans, 68 jours
Clarina Hudon (1905-2016)	111 ans, 34 jours
Judith Pinard* (1898-2009)	110 ans, 345 jours



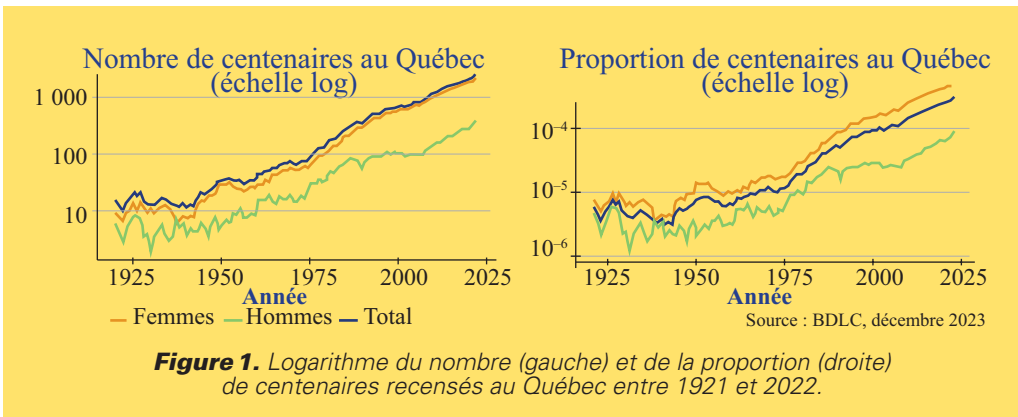


Figure 1. Logarithme du nombre (gauche) et de la proportion (droite) de centenaires recensés au Québec entre 1921 et 2022.

l'échelle logarithmique) depuis 1955, ce qui suggère que le nombre d'aspirants au titre de centenaire est en croissance exponentielle. Au 31 décembre 2022, on estime qu'ils étaient au total 2677, dont 2270 femmes.

Description du modèle

On peut facilement calculer le taux de survie avant 105 ans de façon empirique, mais une fois passé ce cap, le nombre d'individus se réduit comme peau de chagrin. Il faut donc recourir à un modèle statistique pour extrapoler au-delà de la plage des durées de vie observées. Vu l'augmentation du nombre de centenaires, il semble plausible que le record de longévité québécoise de Marie-Louise Meilleur finisse un jour par être battu. Notre modèle servira à déterminer les chances que cela se produise.

Appelons X la durée de vie restante d'une personne ayant

déjà atteint un grand âge. À des fins d'illustration, nous fixerons ici ce seuil à $s = 106$ ans, mais d'autres valeurs voisines conduiraient substantiellement aux mêmes conclusions. Puisque la valeur de X varie d'un individu à l'autre, on la conçoit comme une variable aléatoire dont la loi F donne, pour tout nombre réel x , la probabilité que $X \leq x$, c'est-à-dire $F(x) = \Pr(X \leq x)$.

C'est un fait bien documenté que le modèle exponentiel fournit une excellente représentation de la variation entre les durées de vie résiduelles des individus au-dessus d'un seuil s suffisamment grand. Dans ce modèle,

$$F(x) = F_{\theta}(x) = 1 - e^{-x/\theta}$$

pour tout $x \geq 0$, où $\theta > 0$ est un paramètre représentant la durée de vie résiduelle moyenne pour la population de référence. Une fois la valeur de θ choisie, on peut aisément procéder à divers calculs, notamment

- la probabilité qu'une personne ayant atteint le seuil s décède entre t et $t + 1$ ans plus tard, soit

$$\begin{aligned} \Pr(t \leq X < t + 1) &= F_{\theta}(t + 1) - F_{\theta}(t) \\ &= e^{-t/\theta} - e^{-(t+1)/\theta}; \end{aligned}$$

- le nombre attendu de décès dans chacune des tranches d'âge $(s + t, s + t + 1)$ étant donné n individus ayant atteint le seuil s , soit $n \times \Pr(t \leq X < t + 1)$ pour tout $t \in \{0, 1, 2, \dots\}$.



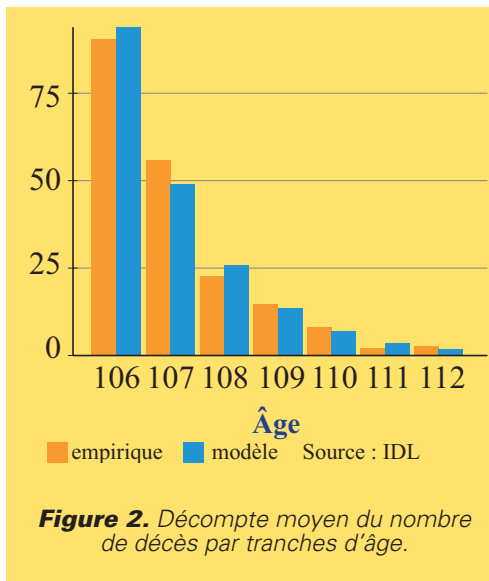


Figure 2. Décompte moyen du nombre de décès par tranches d'âge.

La figure 2 donne une idée de la qualité de l'ajustement du modèle exponentiel pour les données québécoises, pour lesquelles on a posé $\theta = 1,547$. Nous reviendrons sur ce choix un peu plus loin. Pour chaque tranche d'âge $[s + t, s + t + 1)$, où $s = 106$ ans et $t \in \{0, 1, 2, \dots\}$ ans, deux colonnes permettent de comparer le décompte empirique des décès (en orange) à celui prédit par le modèle (en bleu), étant donné $n = 2677$. D'emblée, l'ajustement semble bon. On peut le confirmer au moyen d'un test du khi-deux, comme on le verra à la toute fin de l'article.

Exprimé autrement, on peut s'attendre à ce qu'en moyenne, une seule personne sur 2048 âgées de 106 ans vive au-delà de 117 ans. Sachant qu'à l'heure actuelle, environ 2677 personnes nées au Québec ont plus de 106 ans, il y a donc de bonnes chances pour que le record de Mme Meilleur soit éventuellement battu. Ce n'est qu'une question de temps.

Conséquences

En prenant pour acquis le modèle exponentiel de paramètre $\theta = 1,547$, on peut déduire plusieurs faits intéressants concernant la longévité d'un individu ayant déjà atteint le seuil de 106 ans. D'une part, l'espérance de vie d'un tel individu est de $\theta = 1,547$ ans, ce qui est relativement court. De fait, la probabilité que cette personne vive encore au moins un an est

$$\Pr(X > 1) = e^{-1/\theta} \approx 0,523.$$

D'autre part, cette probabilité de vivre au moins un an de plus reste la même une fois que l'individu a atteint l'âge t de 107 ans, 108 ans, 109 ans, et ainsi de suite. En effet, sachant que $X > t$, la probabilité conditionnelle de vivre un an de plus est donnée par la formule

$$\begin{aligned} \Pr(X > t + 1 | X > t) &= \Pr(X > t + 1) / \Pr(X > t) \\ &= e^{-(t+1)/\theta} / e^{-t/\theta} \\ &= e^{-1/\theta} \approx 0,523. \end{aligned}$$

Cette propriété, dite d'absence de mémoire, caractérise les lois exponentielles : ce sont les seules lois de probabilité sur $[0, \infty)$ qui la possèdent.

Ainsi, tout se passe comme si, une fois atteint l'âge de 106 ans, les chances qu'un individu souffle une bougie de plus sont de l'ordre de 50%, soit l'équivalent du jet d'une pièce de monnaie, avec comme résultat survivre un an de plus si on obtient (disons) pile, ou décéder dans le cours de l'année à venir (face). Vu sous cet angle, le record de Marie-Louise Meilleur est remarquable. Une fois âgée de 106 ans, ce qui est un exploit en soi, elle a réussi à obtenir « pile » 11 fois de suite avant que la pièce du destin tombe sur face au douzième coup, entraînant son décès entre les âges de 117 et 118 ans ! Les chances d'observer une telle séquence sont de

$$1/2^{11} = 1/2048 \approx 4,88 \times 10^{-4}.$$

On peut aussi aborder le phénomène du vieillissement humain à des âges très avancés à l'aide du concept de taux de défaillance. Cette notion s'exprime comme la limite, quand Δt tend vers zéro, du risque instantané de décès dans l'intervalle $[t, t + \Delta t]$, sachant que la personne est vivante à l'instant t . En termes mathématiques, cette limite, si elle existe, est donnée par la formule

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \rightarrow 0} \Pr(t < X < t + \Delta t | X > t) \\ &= \lim_{\Delta t \rightarrow 0} \Pr(t < X < t + \Delta t) / \Pr(X > t). \end{aligned}$$

Lorsque la loi F admet une dérivée f en tout $t > 0$, on a alors

$$\lambda(t) = f(t) / \{1 - F(t)\}.$$

Nous savons que le taux de défaillance du corps humain est relativement bas jusqu'à l'âge de 30 ans et qu'il entreprend ensuite une lente ascension, laquelle s'accélère après l'âge de 70 ans. Toutefois, si le modèle exponentiel est une bonne représentation de la durée de vie résiduelle des personnes ayant déjà soufflé 106 bougies, il se trouve que pour cette cohorte d'exception, le taux de

défaillance est élevé, certes, mais essentiellement constant ! En effet, selon le modèle exponentiel, on a

$$f_{\theta}(t) = \theta^{-1} e^{-t/\theta}$$

et donc, quel que soit $t \in [0, \infty)$,

$$\begin{aligned} \lambda(t) &= \theta^{-1} e^{-t/\theta} / e^{-t/\theta} \\ &= 1/\theta \approx 0,646. \end{aligned}$$

Il s'agit donc d'un modèle bien particulier, dont l'estimation et la validation méritent notre attention.

Estimation du paramètre θ

Supposons pour le moment que le modèle exponentiel soit une représentation appropriée de la durée de vie résiduelle des personnes nées au Québec et ayant atteint l'âge de 106 ans. Comment procéder à l'estimation du paramètre θ de ce modèle et mesurer l'incertitude qui l'entoure ?

Lorsque les données X_1, \dots, X_n dont on dispose constituent un échantillon aléatoire simple, la théorie statistique classique nous dicte que la valeur la plus plausible de θ est celle qui maximise la fonction de vraisemblance

$$L(\theta) = f_{\theta}(x_1) \times \dots \times f_{\theta}(x_n)$$

ou, de façon équivalente, qui maximise son logarithme, ce qui a l'avantage de transformer le produit en somme et facilite les calculs :

$$\begin{aligned} l(\theta) &= \log\{L(\theta)\} \\ &= -n \log(\theta) - (x_1 + \dots + x_n)/\theta. \end{aligned}$$

En dérivant la fonction l par rapport à θ , on trouve que le maximum se produit lorsque $\theta = (x_1 + \dots + x_n)/n = 1,397$, c'est-à-dire lorsque l'on choisit la moyenne expérimentale comme valeur de θ .

Cette estimation étant fondée sur un échantillon, elle est sujette à une incertitude que l'on peut qualifier par une marge d'erreur, exactement comme on le fait pour les sondages d'opinion publiés dans les médias.

Ainsi peut-on affirmer que la véritable valeur de θ n'est pas forcément égale à 1,397, mais qu'elle se situe entre 1,268 et 1,545 « 19 fois sur 20 ». La valeur $\theta = 1,397$ s'avère toutefois la plus plausible au vu des données.

Le calcul de l'intervalle de confiance, que nous ne détaillerons pas ici, s'appuie sur le théorème central limite, lequel stipule que dans la mesure où la taille n de l'échantillon est assez élevée, la variation intrinsèque de la moyenne échantillonnale $(x_1 + \dots + x_n)/n$ obéit à une loi normale, aussi appelée courbe de Gauss.

Source des données

En pratique, on ne peut pas procéder à l'estimation du paramètre θ par la méthode décrite ci-dessus, et donc la valeur estimée $\theta = 1,397$ n'est pas la bonne, car les seules données dont on peut disposer pour une étude sérieuse de la longévité sont de nature démographique et ne constituent pas du tout un échantillon aléatoire simple.

Au Québec, des données fiables en matière de durée de vie humaine sont recueillies par le Directeur de l'État civil et l'Institut de la statistique du Québec. Cependant, mêmes les registres de l'état ne sont pas exempts d'erreurs administratives. Ces données ont donc été contre-vérifiées manuellement par les démographes Robert Bourbeau et Bertrand Desjardins, professeurs à l'Université de Montréal, en recoupant les données de recensement, les certificats de baptêmes et de mariage, etc., pour éviter la confusion liée notamment aux homonymes. Certaines personnes étant nées au 19^e siècle, ce travail de moine nécessite en outre de consulter les archives religieuses, avec toutes les difficultés que cela comporte (dispersion des registres, etc.).

Le fruit du travail de Bourbeau et Desjardins est disponible dans la base de données de l'IDL et contient tous les individus décédés après 1983 (pour les supercentenaires) et 1985 (pour les personnes mortes entre les âges de 105 et 110 ans). Toutefois, leur collecte s'est arrêtée en 2009, ce qui explique le fait que plusieurs records récents ne s'y trouvent pas, tel que précisé dans le tableau 1. Par ailleurs, seules les personnes nées et décédées au Québec figurent dans la base de données de l'IDL, ce qui exclut certaines personnes décédées ailleurs, dont Marie-Louise Meilleur. Néanmoins, ces données sont quasi-exhaustives, si on exclut les migrations.

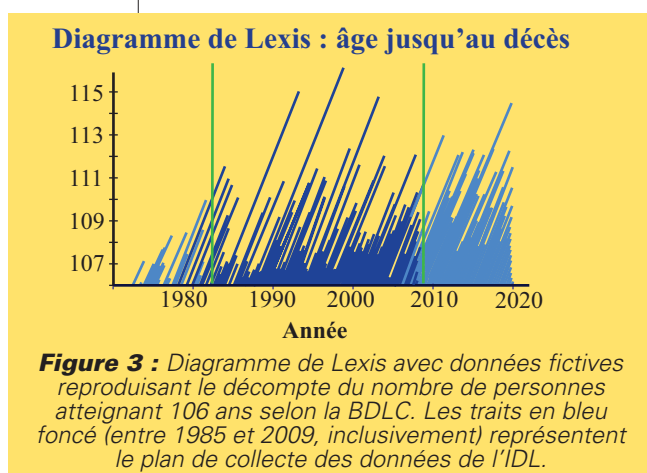


Figure 3 : Diagramme de Lexis avec données fictives reproduisant le décompte du nombre de personnes atteignant 106 ans selon la BDLC. Les traits en bleu foncé (entre 1985 et 2009, inclusivement) représentent le plan de collecte des données de l'IDL.

Fantômes et troncature

Puisque l'*International Database on Longevity* (IDL) ne recense que les personnes décédées entre 1985 et 2009, l'échantillon sur lequel s'appuie l'ajustement du modèle exponentiel est sujet à un mécanisme appelé « troncature », qui engendre des biais si on en fait fi.

Pour comprendre la nature et l'impact de la troncature, considérons le diagramme de Lexis présenté à la figure 3. Sur ce graphique, l'axe horizontal représente le fil du temps, alors que l'axe vertical est dévolu à l'âge des individus. Les lignes diagonales représentées sur le graphique montrent l'âge de chacun d'entre eux en fonction du temps jusqu'à leur décès ; le passage du temps étant le même pour tous, les pentes des droites sont toutes égales (à 1 si les deux axes sont à la même échelle). Comme seules les personnes ayant vécu plus de 106 ans sont incluses, l'axe des ordonnées commence à 106 et la ligne de vie de la personne s'arrête à son décès.

Sur la figure 3, réalisée à partir de données fictives mais réalistes, les traits bleu foncé représentent les individus inclus dans la collecte. En revanche, les traits bleu pâle constituent des fantômes, c'est-à-dire des personnes qui n'ont pas pu être répertoriées dans la base de données IDL car elles sont soit décédées avant 1985, soit après 2009. On note qu'en raison de l'augmentation de l'espérance de vie de la population québécoise, on s'attend à ce qu'il y ait densification des traits diagonaux à mesure que le temps passe, tel que représenté sur la figure.

La nature même de la base de données IDL fait que tous les fantômes en sont exclus. Pour mieux prendre conscience du problème que cela pose pour l'analyse des données, considérons la figure 4, qui montre l'âge au décès en fonction de l'année de naissance des entrées québécoises de l'IDL (croix bleu pâle), avec l'âge maximum observé par année en bleu foncé. Comme on peut le constater, la troncature affecte la fenêtre d'âges potentiels pour chacun des individus : par exemple, une personne née en 1900 ne sera incluse dans l'IDL que si elle décède au plus tard en 2009, donc au plus à l'âge de 110 ans. Toutes les données répertoriées doivent donc forcément se situer entre les deux droites qui déterminent les âges minimum et maximum d'inclusion dans la base de données. Ces contraintes varient d'une personne à l'autre et leur effet est plus marqué pour les individus nés tardivement; en outre, ces derniers sont plus nombreux.

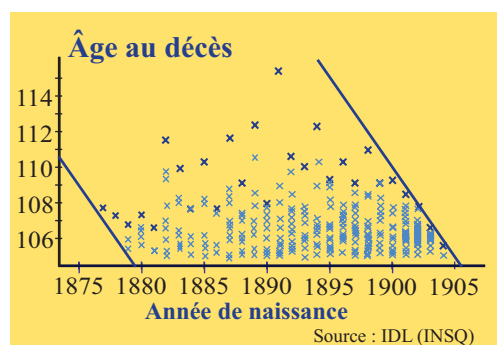


Figure 4 : Âge au décès en fonction de l'année de naissance; chaque croix représente un individu. L'âge maximum au décès observé est rapporté en bleu foncé pour chaque année de naissance.

Vraisemblance corrigée

Pour tenir compte de cette limitation majeure dans les données, il faut modifier la vraisemblance de façon à refléter la troncature à laquelle est soumis chacun des 197 âges de décès de plus de 106 ans observés dans l'IDL. On incorpore ainsi l'information à l'effet que l'âge observé devait forcément se situer dans un intervalle donné, dont les bornes sont déterminées par les droites de la figure 4 selon l'année de naissance, en soustrayant le seuil de 106 ans. Pour l'individu i , la borne inférieure a_i est donnée par le maximum entre 0 et l'âge de l'individu au 1^{er} janvier 1985 moins 106 ans tandis que la borne supérieure b_i est l'âge que cette même personne aurait atteint si elle avait survécu jusqu'à la fin de décembre 2009, moins 106 ans. Si l'âge au décès est $x_i + 106$ ans, alors la vraisemblance $L(\theta)$ s'écrit

$$\frac{f_{\theta}(x_1)}{F_{\theta}(b_1) - F_{\theta}(a_1)} \times \dots \times \frac{f_{\theta}(x_n)}{F_{\theta}(b_n) - F_{\theta}(a_n)}.$$

C'est donc cette fonction de θ qu'il faut maximiser, et non la précédente. Malheureusement, ce problème d'optimisation n'admet pas de solution explicite. En revanche, on peut facilement tracer la courbe correspondante (figure 5) et se tourner vers l'optimisation numérique pour obtenir la valeur du paramètre la plus plausible au vu de l'échantillon, à savoir $\theta = 1,547$.

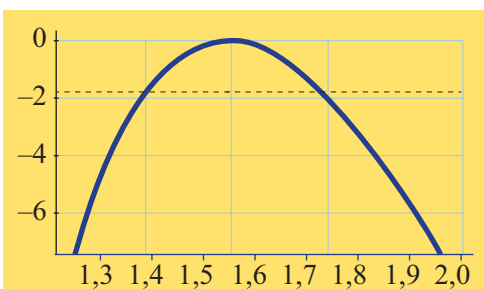


Figure 5:
Fonction de log-vraisemblance $l(\theta)$,
décalée pour que la hauteur maximale
de la courbe soit à zéro.

Il est instructif de comparer cette valeur à la moyenne empirique $(x_1 + \dots + x_n)/n$, soit 1,397. Cette dernière a tendance à sous-estimer la durée moyenne de vie car elle ne tient pas compte de la troncature. Ceci vient du fait qu'en raison du nombre croissant de centenaires, beaucoup plus de gens sont affectés par la borne supérieure (à droite de la figure 4) que par la borne inférieure.

Sur la figure 5, qui montre l'allure de la courbe $l(\theta)$, on peut apercevoir une ligne horizontale en trait pointillé à hauteur de -1,92. L'abscisse des deux points de croisement de cette ligne avec la courbe $l(\theta)$ permet de repérer les bornes de l'intervalle de confiance à 95% associé à l'estimation de θ . Il s'agit en l'occurrence des valeurs 1,384 et 1,741, qui donnent les bornes de la marge d'erreur « 19 fois sur 20 » (c'est-à-dire 95%).

Ajustement du modèle

Il existe divers moyens de confirmer la qualité de l'ajustement observé à la figure 2. La façon la plus facile et standard de procéder consiste à calculer la somme des carrés des écarts entre, d'une part, les valeurs prédites par le modèle exponentiel de paramètre $\theta = 1,547$ et, d'autre part, les fréquences empiriques pour les groupes $[106, 107)$, $[107, 108)$, ..., $[111, \infty)$. La valeur observée de cette statistique χ^2 est alors comparée à la loi du khi-deux à 5 degrés de liberté, soit le nombre de catégories moins un. En l'occurrence, on trouve $\chi^2 = 1,98$, ce qui correspond au quantile d'ordre 15% de la loi de référence. Le seuil observé du test est donc de 85%, ce qui ne soulève absolument aucun doute quant à la pertinence du modèle exponentiel. Les conclusions que nous avons tirées de l'analyse des données de l'IDL¹ sont donc valables.

1. International Database on Longevity (2024). Institut national d'études démographiques (Ined) (hôteur). Extrait de : <https://www.supercentenarians.org/fr/> (11/11/2024)