Supercentenarians

Abstract

A supercentenarian, defined as an individual who has lived for at least 110 years, must register their age with the government. However, memory decline associated with advanced age or some other reasons can lead to inaccurate birth date reporting, raising concerns about potential fraud. To investigate this issue, we analyzed data from a study tracking over 1,700 supercentenarians. Our analysis included exploratory data analysis of the birthdate distribution (including a Chi-squared test for independence), as well as a non-parametric Wilcoxon rank-sum test. Based on these analyses, we conclude that evidence suggests the presence of fraudulent supercentenarian registrations, particularly prior to the implementation of mandatory government birth registration.

Introduction

Thanks to the wonders of modern technology, our lifespans have increased dramatically. However, despite reaching an average lifespan of around 80 years, only a select few reach the extraordinary age of 100. Moreover, the ratio of supercentenarians who live for more than 110 years is very rare : only about 0.0000025 percent of the population can reach that age. In 2015, Gerontology Research Group(GRG) collected the data on registered supercentenarians(SC) globally. Our study aimed to investigate the potential presence of fraudulent claims in the data.

Data

Our data was sourced from the Supercentenarian website, which maintains a comprehensive database of global supercentenarian cases. This database provided us with information including birth date, and birthplace encompassing a total of 1,739 cases. The population of each state of the US from 1790 to 1950, and the year of introduction of birth registration in each state in the US was used for implementing this study. Details are summarized below.

- **Birth date :** The specific data of birth for an individual, including day, month and year.
- **Birth place** : Country and state (in case of US-born individuals) where the individual was born.
- **The population of each state in the US** : the population of every decade from

1790 - 1950.

- **Year of birth registration implemented by state** (1841 to 1919).

Data Analysis

To investigate the potential presence of fraudulent claims, we implemented two approaches.

The first approach investigates the distribution of supercentenarians' birthdates to determine whether it follows a uniform distribution. Theoretically, we expect each birthdate to have an equal probability of occurring. A uniform distribution implies that all birthdates are equally likely. By analyzing the actual distribution of supercentenarian birthdates, we can identify any statistically significant deviations from this expected pattern, potentially indicating the presence of fraudulent claims. Here, Chi-square test is implemented to analyze the birthdate distribution for potential deviations from uniformity.

The second approach is to compare the ratios of supercentenarians before and after the introduction of birth registration in the USA. This analysis focuses exclusively on supercentenarians residing in the United States due to limited data availability regarding birth registration and population in other countries. We segregated the supercentenarians into two groups based on whether they were born before or after the introduction of birth registration in their respective state. We expected a higher supercentenarian ratio in the pre-registered group, because birth registration should hinder fraudulent claims of supercentenarian status. This expectation implies that unusually high supercentenarian ratio in the pre-registration group is indicative of fabric data. Since the distribution of the birthdates is heavily right-skewed, it violates the normality assumption of the t-test. Therefore, a non-parametric Wilcoxon rank sum test was implemented to analyze the differences between the two groups.

1) Analyzing the birth date distribution of supercentenarians for uniformity

Analyzing the birth date distribution of supercentenarians for uniformity, at first we visualized the distribution of the birth date of supercentenarians by month.

Figure1.

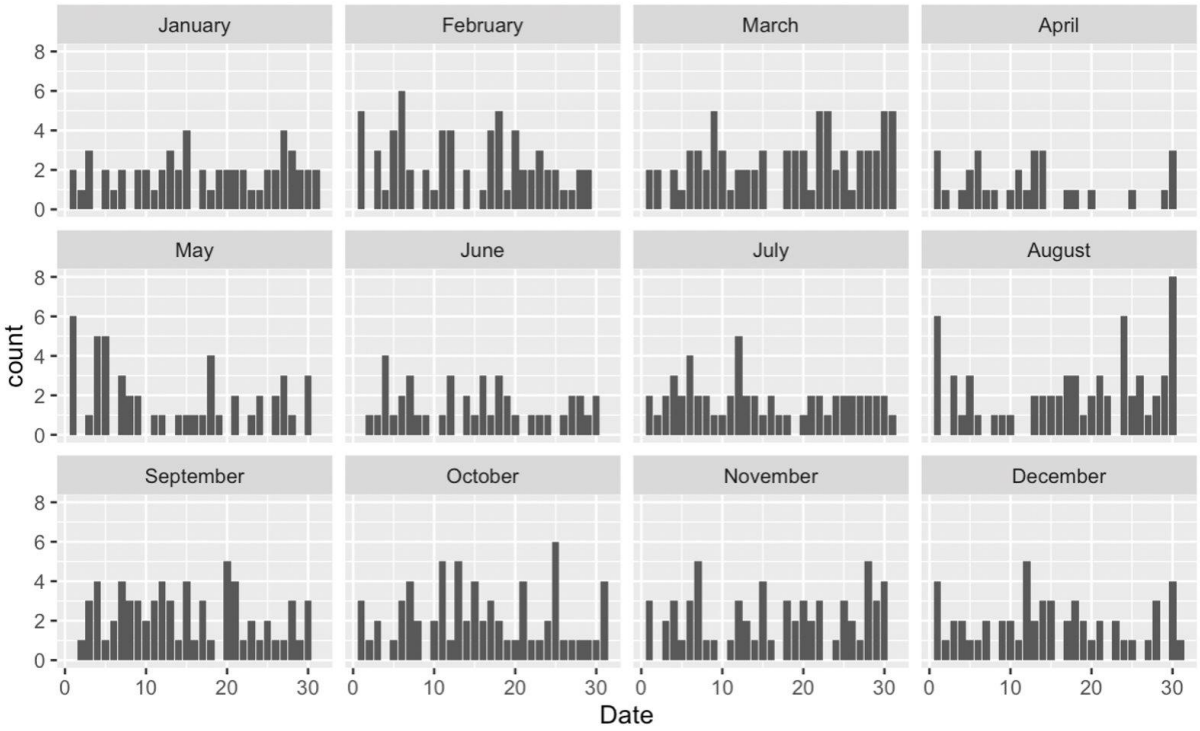

Figure1 depicts the distribution of birthdays for each month. As the figure demonstrates, August consistently exhibits the highest number of births compared to all other months. Conversely, April shows the lowest number of births. This discrepancy indicates a lack of uniformity within the distribution. To statistically support this observation, we implemented a Chi-square test, and the results are presented below.

Table 1.

```
        Chi-squared test for given probabilities

data:  Bdate$freq
X-squared = 437.59, df = 365, p-value = 0.005371
```

A very small p-value of 0.005 provides statistically significant evidence of fraud in the birth date records of Supercentenarians. This suggests that the birthdate distribution is not uniform, making it unlikely to be the result of a natural and unbiased population.

2) Analyzing the differences of the supercentenarian's ratio in the two groups.

Before performing the numerical comparison, we visualized the population distribution of two groups of supercentenarians; those born before and those born after the introduction of the birth registration.

Figure2.

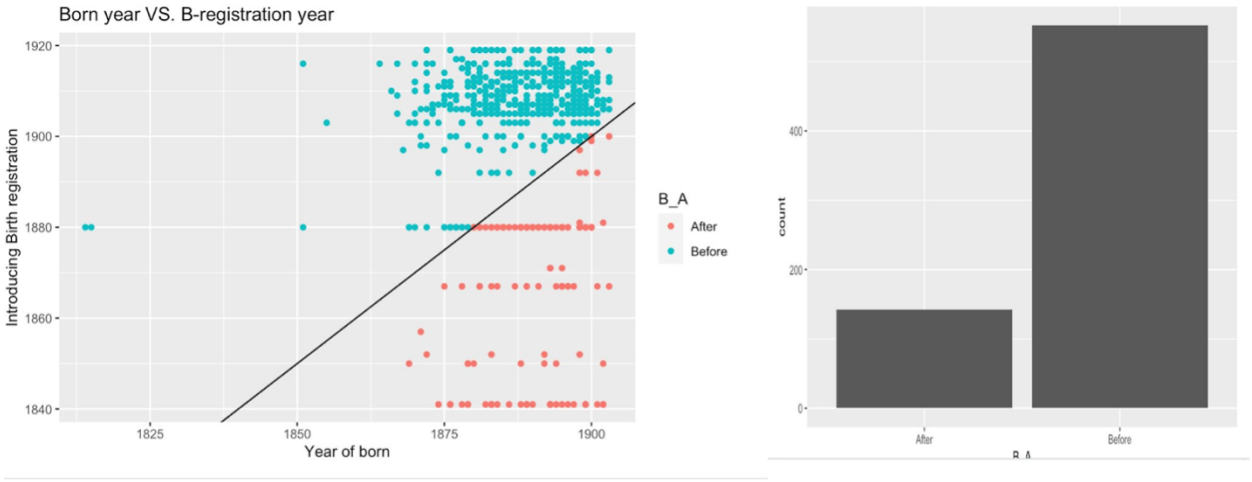

Figure2 shows that there are more supercentenarians before the introduction of birth registration. In the left side panel, the x-axis represents the birth year of supercentenarians and the y-axis represents the year of birth registration introduction. The diagonal line identifies cases where birth and registration years are the same. The blue dots over the diagonal line indicate the individuals who were born before the birth registration was introduced and the red dots below the line indicate those who were born after it. Just by looking we see that blue dots exceed red ones. The bar graph on the right side tells the same story.

Figure3.

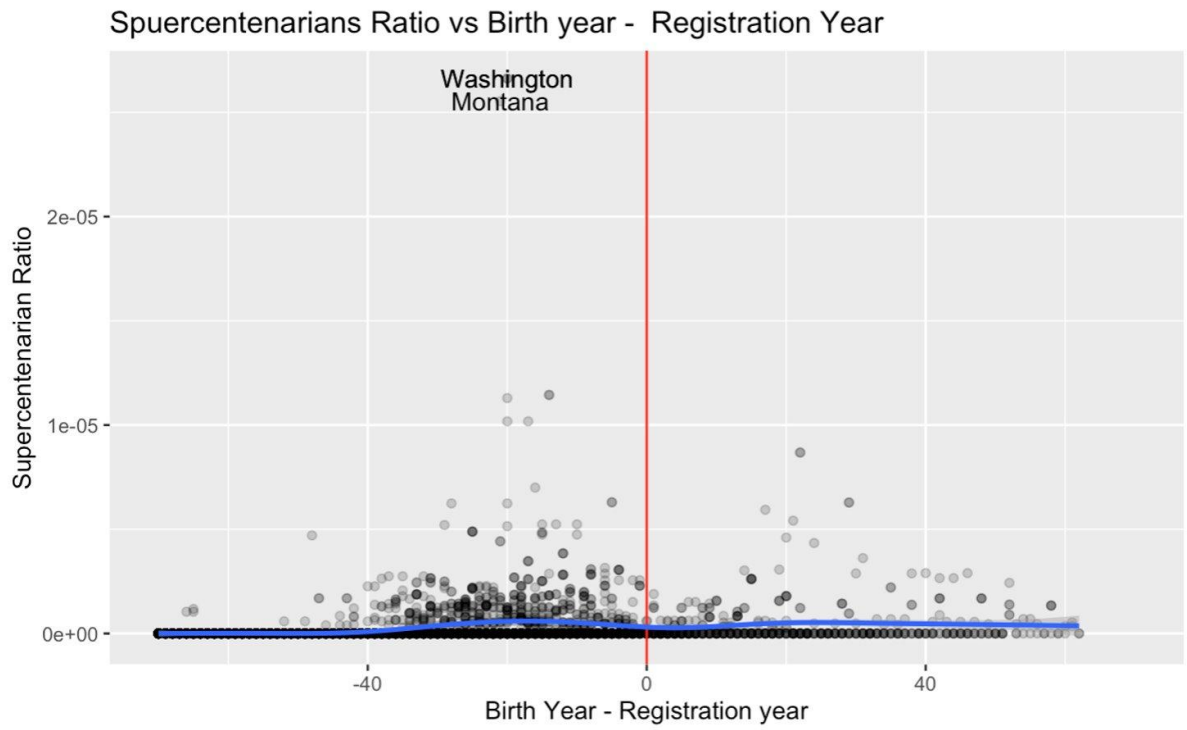

Figure3 illustrates how the ratio of supercentenarians varies with the completion of birth registration throughout the US. The x-axis indicates the differences between supercentenarians' birth year and birth registration year in each state. The y-axis is the ratio of supercentenarians. We can observe that the blue smooth line continuously decreases from a point approximately -20 years and stabilizes at the point zero, where all the states in the US completed the birth registration. The plot reveals the higher ratio of supercentenarians before the completion of the birth registration. To statistically confirm this observation, we employed a nonparametric Wilcoxon rank-sum test.

Table2.

```
        Wilcoxon rank sum test with continuity correction

data:  before$Ratio_log and after$Ratio_log
W = 939374, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

The Wilcoxon test yielded a highly significant p-value close to zero, confirming our observation of a difference in the supercentenarian ratio before and after birth

registration implementation. This result aligns with our previous finding from the Chi-square test for uniformity of the birthdate distribution.

Conclusions

From our study we were able to extrapolate several interesting findings. We first learned from our exploratory data analysis that the distribution of birth dates is not uniform, so it can be suspected as fraud. Next we learned from our Non-parametric test(Wilcoxon rank sum test) that the differences between the groups before and after the introduction of birth registration is statistically significant enough to warrant some suspicion. Additionally, we did see a decrease in Supercentenarian registration once all states had implemented a birth registration mandate.

Limitations

We need detailed records, including photos, references, and other key factors, to substantiate the claim of supercentenarian status. Specifically, more comprehensive evidence, such as reason of the death, place of death, and references from family members, would allow us to assign numerical scores, introduce a reliability variable, and statistically examine potential inaccuracies in conjunction with birth year and other records. This presents an excellent opportunity to explore the statistical aspects of false reporting.

Additionally, a variable that requires more detailed observational data is race. When examining the racial distribution in each state, there was a noticeable difference from the distribution of supercentenarians we observed in the graph. About 10 out of over 50 states showed such discrepancies, suggesting statistically questionable data distributions, raising suspicion that many states may have falsely reported records rather than showing a natural result based on racial distribution. Therefore, if the ongoing distribution of supercentenarians, including data from recent times, is included, it will provide an opportunity to examine the distribution based on race in more detail. This could be helpful in predicting statistical inaccuracies, anticipating false reporting, and predicting the actual distribution of supercentenarians.