

# HOW STATISTICS CAN PREDICT LIFE LONGEVITY

## Introduction

Various statistical studies attempt to answer the question ‘Is there a bound to human life?’ However, these studies differ in their conclusions due to problems related to the reliability of the data, the uncertainty of being able to study this issue with current data, and the inapplicability of mathematical models, such as the extreme value theorem, to a large amount of data. This is why we wanted to try our hand at studying human life expectancy through a statistical model used to make future projections, the ARIMA, trying to understand its limitations and benefits.

## Reliability of Data

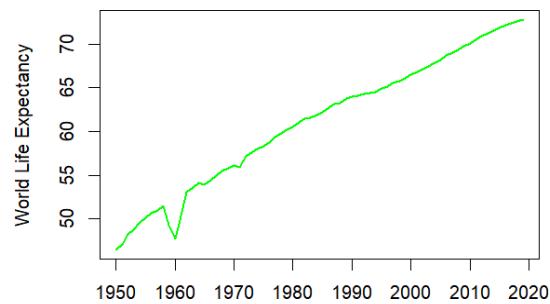
The data on which such research is based comes from a dataset of the Population Division of the Department of Economic and Social Affairs published by the United Nations ([2022 Revision of World Population Prospects](#)), in which estimates are reported on the Life Expectancy from 1950 to 2022. We chose to start directly with past estimates as we considered them fundamental for the correct functioning of the model.

## The bound around year 100

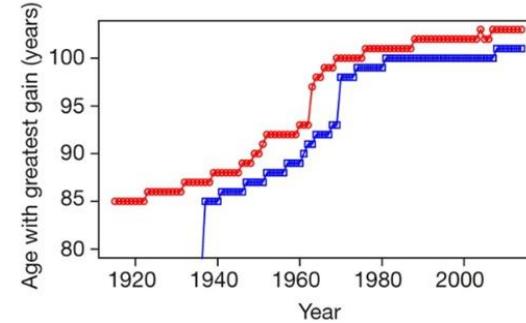
As shown in Figure 1, life expectancy at birth has grown almost constantly through the last seventy years, apart from a downward spike in the 1960s. However, as explained in past research ([Evidence for a limit to human lifespan](#)), even though life expectancy is increasing, the maximum lifespan has not increased since the 1990s. It has been proven, with the available data, that the gains in survival peak around 100 years of age and then rapidly decline. The age with the most rapid gains has increased over the century, but its rise has been slowing and it appears to have reached a plateau (Figure 2).

## Hypothesis and Truncation

The 2020 and 2021 data have been removed from the time series as it is reasonably believed that these exceptions are the result of the 2020 pandemic and an ARIMA model isn’t capable of taking into account these values while producing a proper prediction. It is plausible that similar events may happen in the future, even then, the model’s estimate is fairly accurate compared to reliable data like the UN projections of world life expectancy.



[Figure 1] Life Expectancy of the World

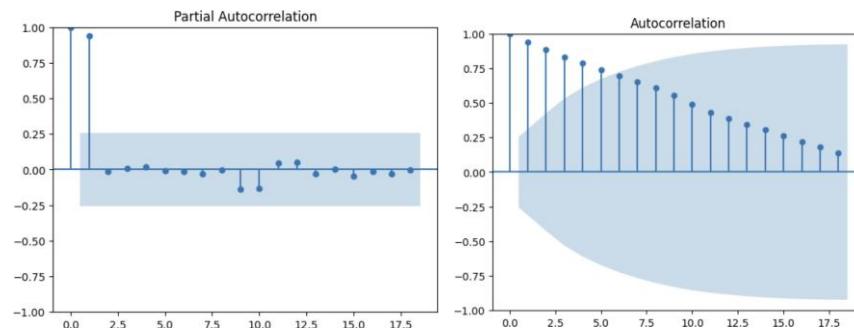


[Figure 2] Relationship between years and the age experiencing most rapid gains in the past 100 years.

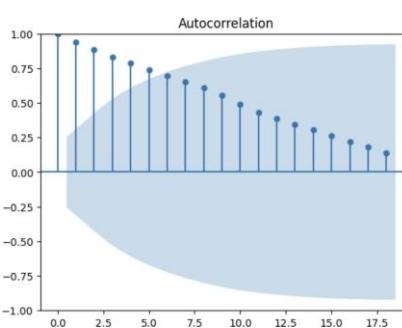
## The ARIMA Model

An ARIMA, which stands for Auto Regressive Integrated Moving Average, is a statistical model that analysts use to make time-series forecasts and similar predictions. For example, it was used to predict life expectancy in Saudi Arabia up to the year 2030 ([Prediction of life expectancy in Saudi Arabia by 2030 using ARIMA models](#)). The model works by using a stationary time series to predict future values based on the previous trend. This is done by analyzing the past lag correlations of the data frame. To implement the model, we first checked

if our time series was stationary, which can be done using various tests. We chose to use ACF (Figure 3, Autocorrelation function) and PACF (Figure 4, Partial Autocorrelation function) plots of our data to visualize the correlations between the lags and throughout the entire



[Figure 3] Autocorrelation function



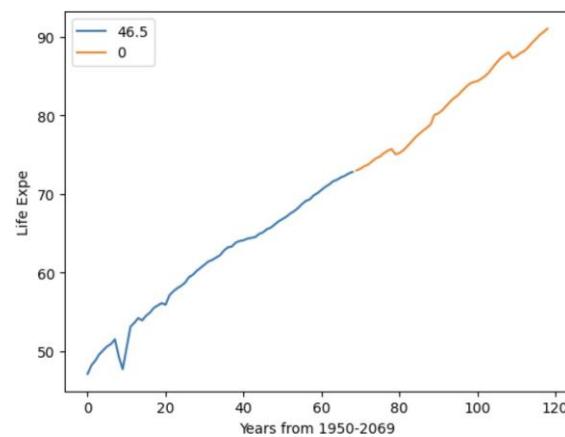
[Figure 4] Partial Autocorrelation Function

Series. These plots helped us determine suitable values to input

into the model for a more accurate forecast. Another test we conducted was the Adam Dickey-Fuller test, which checks for the rejection of the null hypothesis and identifies the root of the time series. Unfortunately, our p-value was too high, indicating that we couldn't reject the null hypothesis and declare our time series as stationary. Therefore, we had to differentiate our data. After differentiating the data, we obtained a p-value of less than 0.05, which was optimal, allowing us to proceed with inserting reasonable values into our model. ARIMA modeling can be complex and time-consuming, but we can identify plausible values by using the differentiated ACF and PACF plots and looking for significant spikes. After this process, we settled on an 8,1,4 ARIMA model, which, when compared to the actual values in our dataset, provided fairly accurate results correlating 0.99. Finally, we decided to forecast 50 years into the future. To improve accuracy, we performed the forecast in steps of 10 years, repeating the process 5 times.

## Results and conclusion

As shown in Figure 5, the result of the experiment is clear: life expectancy will continue to rise quite linearly. This is due to the very nature of the model, which was "trained" on the data of the years 1950-2019. Therefore, this estimate as specified above is not intended to be a scientifically precise prediction, but rather a reference to how much life expectancy could continue to grow. It can be assumed that as long as technological and economic development continues to grow, then on average we will live longer. However, as Figure 2 suggests, the fact that life expectancy will rise does not imply that the bound will disappear. Indeed, we live more and more on average, but it is not enough to say that we will get to the point of living forever.



[Figure 5] Prediction of the life expectancy from 2019 to 2069, obtained using the ARIMA model

**Folder for Python Script, R Script, and Dataset used:**  
[Resources statistics project](#)