Connor Moss and Jackson Kennedy
Olympics Project | Executive Summary
4-25

Deep Dive: Forecasting Swimming Records Through Time Series and EVT Analysis

Our objective was to analyze how Olympic swimming times have evolved. We used advanced statistical measures such as exponential smoothing and Extreme Value Theorem (EVT) to explore both the *average* times and *fastest* times across different years for various strokes, while in the process attempting to forecast future record-breaking performances.

We began by constructing a variety of exponential smoothing models in an attempt to predict future times in the Men's 100m Freestyle. Our initial plan was to use more traditional time series techniques, such as autoregressive integrated moving average (ARIMA) modeling; however, after struggling to stationarize the Men's 100m Freestyle data, we transitioned to exponential smoothing, which does not require/assume stationarity. (Although not entirely sure why it was so difficult to stationarize the data, we think that either the linear trend was too strong or the data set was simply too small; it could also be a combination of both of these things. We could only stationarize the data by taking the fourth (!) difference, which resulted in a notably smaller set of priors (16 → 12) and also, presumably, introduced correlations that weren't present in the original data.) Before constructing our exponential smoothing models, we took into consideration 1) which of the three exponential smoothing methods (simple, Holt's, Holt-Winters) was most appropriate, and 2) how to tune our method-of-choice's parameters in order to produce the best results. We landed on Holt's method as most appropriate due to our data's clear downward trend and lack of seasonality; we also identified the smoothing (α = 0-1) and dampening (T/F) parameters as potentially having significant effects on the models' predictions. Then, after constructing three different models, each with slightly different parameter values, we crowned the *dampened* model with α = 0.3 as the best. This model predicts male swimmers to continue to improve in the 100m Freestyle, but at an increasingly slower rate. In an effort to learn more about the likelihood of this model over- or under-estimating the progression of Olympic swim times, we turned to our good friend, Extreme Value Theory.

EVT focuses on the tails of probability distributions, which, in this case, are the fastest (and slowest) times. We compiled a list of the right-tail (i.e. fastest) observations, creating a distribution of extreme values. We used the Augmented Dickey-Fuller test to check for stationarity in the time series data. We did not have statistically sufficient evidence to reject our H_0, which was that the data was non-stationary (i.e. trend is present). This could lead to biased data and incorrect conclusions, so we addressed this differently than the aforementioned fourth-order difference method. We addressed this by detrending the data with a simple linear regression model and using the residuals to make our predictions. This separated the variation in swim times that aren't explained by the year variable, effectively removing the "noise" from our dataset so that we could appropriately fit a *general* extreme value model with the *extRemes* package.

Our results from the model came from looking at the return levels, which tell us the times that swimmers are expected to beat only once in a certain number of years. We used the years (4, 8, 12, 16) to simulate the next four Olympic Games. In doing so, for the 100m freestyle, we found that the model predicted a time of 45.95 to happen in this upcoming Summer Games and 45.93 to happen once in the next four. With the current record standing at 47.02, there is still more than a second of improvement for the athletes. This wasn't the case with the 100m backstroke as Ryan Murphy (USA) has already broken the record that our model predicted with a blistering 51.85, where our model predicted 51.90 to happen once in the next four Games. The results for breaststroke and butterfly suggest that swimmers' times have started to plateau, as our models showed room for improvements of only 0.2 and 0.4 seconds, respectively.

To conclude, through our time series and extreme value analyses, we gained valuable insight into both the historical and anticipated progression of Olympic swimmers. Most notably, our models suggest that record-breaking times in the 2024 Olympics are more likely in some events (such as the Men's 100m Freestyle) than others, and also that the general rate of improvement seems to be decreasing, or plateauing, across all events; these findings fall closely in line with both of the articles on which this project is (loosely) based. If given more time and/or a slightly more extensive data set, we would like to more carefully examine the differences between men and women, with the focus remaining on rate of progression. Additionally, we think it would be interesting to compare the evolution of swimmers to that of track and field athletes; despite the sports' obvious differences — both historically at the Olympics, and in terms of what they ask of the human body — this could potentially help to better inform our predictions of record-setting swim times.