Modelling sport results with extreme value methods

Second semester project

Author: *Csáfordi József*
Matematika MSc

Supervisor: *Dr. Zempléni András*
Department of Probability Theory and Statistics

Eötvös Loránd University
Faculty of Science

Budapest, 2023

Contents

1 Introduction **3**

2 Data **3**

3 Pareto-distribution **3**

3.1 Theoretical Background 3

3.2 Data fitting 5

4 Analysis and estimation **6**

5 Summary **6**

6 Goals **8**

1 Introduction

In my second phase of my project we left the Berlin marathon due to the few and unreliable data. From this time, we are working with the data of the Boston Marathon, where we have much more information about every competitor from 1898 to 2019. (data: [5] & [3]) Our goal has been to try to predict the best expected results of the following years based on our knowledge. For this purpose, we've used the Pareto distribution to see if there would ever be a sub-two hours Boston Marathon time in the near future. Throughout my work, I continued to use the R programming language.

2 Data

During this semester, cleaning and processing the data also proved to be a serious task, since at the beginning of the 20th century the results were not recorded as precisely as they are today. Furthermore, on the competition's official page ([5]) the data of the half-marathoners or wheelchair marathoners have not yet been separated, which would significantly distort the final picture. After solving these problems and cleaning them, approx. 614000 rows were left. (in the ratio of 1:2 women to men). Since our primal goal was to fit the Pareto distribution for each year, we have to consider the amount of data. Given that there were enough recorded finishers, the process of trimming the data below the threshold proved to be effective in obtaining acceptable estimates based on the retained information. In the case of male competitors, from 1975, while for women from 1981, that we do not get misleading parameters when fitting the distribution. As in the first semester, the running results are understood in seconds and received a multiplier of -1, since the models are looking for maximum, while we are looking for the best (i.e. minimal) result.

3 Pareto-distribution

3.1 Theoretical Background

The Pickands–Balkema–De Haan [2] theorem is a fundamental result in extreme value theory, establishing a connection between the tail behavior of a distribution and the Generalized Pareto Distribution (GPD). This theorem states that, for a wide class of distributions and a sufficiently high threshold u, the distribution of exceedances $X - u$ above this threshold, properly normalized, converges in distribution to the GPD. The

GPD is characterized by its cumulative distribution function:

$$F(x) = 1 - \left(1 + \alpha \frac{x - \beta}{\gamma}\right)^{-1/\alpha,} \tag{1}$$

where γ is the scale parameter, β is the location parameter, and α is the shape parameter. And $\frac{x-\beta}{\gamma}$ for $\alpha \geq 0$, $\frac{-1}{\alpha} \geq \frac{x-\beta}{\gamma} \geq 0$ for $\alpha < 0$ and $F(x) = 1 - e^{-z}$ if $\alpha = 0$ This result is particularly valuable for modeling extreme events, as it allows practitioners to focus on the tail behavior of the distribution and estimate parameters crucial for risk assessment and extreme value prediction. The Pareto distribution, denoted by $X\ Pareto(\alpha, x_m)$ is a continuous probability distribution defined for $x \geq x_n$ with parameters $\alpha > 0$ (shape parameter) and $x_m > 0$ (scale parameter). The probability density function of the Pareto-distribution is given by:

$$f(x; \alpha, x_m) = \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}}, x \geq x_m \tag{2}$$

Here, α governs the shape of the distribution and x_m is the minimum possible value. The cumulative distribution function is:

$$f(x; \alpha, x_m) = 1 - \left(\frac{x_n}{x}\right)^{\alpha}, x \geq x_m \tag{3}$$

The expected value (μ) and variance (σ^2) of the Pareto-distribution are the following:

$$\mu = \frac{\alpha x_m}{\alpha - 1}, \text{ for } \alpha > 1 \sigma^2 = \frac{\alpha x_m^2}{(\alpha - 1)^2(\alpha - 2)}, \text{ for } \alpha > 2 \tag{4}$$

The Pareto distribution is particularly useful in modeling situations where a small fraction of the population accounts for a large proportion of the total, such as income distribution, city population sizes, and wealth distribution. In the introduction of the Generalized Pareto Distribution (GPD), we acknowledge that extreme values in a system are infrequent and concentrated, deviating from the behavior of typical distributions. The GPD is theoretically designed to model values that surpass the average behavior of the system, and it plays a key role in the concept of limit distributions, notably the Maximum Domain of Attraction (MDA). The MDA condition is precisely where the GPD limit is satisfied.(1) This condition implies that extreme values associated with certain types of distributions, occurring at large magnitudes within the system, approximately follow the same distribution. Therefore, when the MDA condition is met, the GPD becomes a suitable model for extreme values. This relationship is fundamental in understanding and characterizing extreme events, as the GPD efficiently enables the modeling of extreme values. It does so by employing parameters that govern the deviation and variability of values from the mean in the extreme range, providing a valuable tool for the analysis and estimation of

	Year	Quantil(%)	Data size	Threshold	Estimated Scale par	Estimated Shape par	AD-test Scale par	AD-test Shape par	AD-test p value	Estimation	AD-test Standard error
1	1975	50	927	-10918.16	1418.4108	-0.4456972	1411.8555	-0.4440295	7.546909e-03	-7738.516	49.51586
2	1975	51	908	-10884.00	1396.9488	-0.4434493	1617.5068	-0.5214033	8.744642e-10	-7781.782	49.44621
3	1975	52	890	-10805.84	1302.4495	-0.4198453	1299.2676	-0.4185289	8.232981e-02	-7701.472	48.62920
4	1975	53	872	-10787.68	1304.1006	-0.4233710	1303.6204	-0.4234854	1.547786e-01	-7709.368	48.92214
5	1975	54	854	-10769.52	1307.9152	-0.4281928	1468.9626	-0.4890195	1.562653e-05	-7765.626	49.12656
6	1975	55	836	-10752.36	1314.0362	-0.4337263	1505.4042	-0.5050790	1.230791e-06	-7771.828	49.38706
7	1975	56	818	-10732.20	1315.4204	-0.4377978	1533.2460	-0.5182084	4.725113e-08	-7773.456	49.63352
8	1975	57	800	-10715.08	1322.0075	-0.4433014	1322.3081	-0.4438671	1.814391e-02	-7736.017	50.04544
9	1975	58	781	-10657.76	1265.8460	-0.4306276	1447.2625	-0.5009250	4.374272e-06	-7768.580	49.52528
10	1975	59	763	-10618.88	1237.7245	-0.4259474	1389.1585	-0.4869013	8.556213e-05	-7765.820	49.36464

Figure 1: Parameter-table

extreme events in diverse systems. Therefore, the GPD enables effective modeling and estimation of extreme values, aiding in the understanding and handling of phenomena occurring in the extreme range of a system. Not to forget to mention that, in the case of a negative shape parameter the distribution is bounded from above. Since we modeled the exceedances this semester, we did not need the location parameter (β) in the formulas, which is set to 0.

3.2 Data fitting

As a result of this, we applied the Pareto distribution to our dataset annually, fitting it with quantiles ranging from 50 to 99 for each year to obtain the optimal parameters separately for each year usint the gpd.fit function (see:[1]). For this purpose, we first fitted the distribution with quantiles from 50 to 99, saving the results along with necessary information in a table (1). The selection of the threshold was also a crucial step in fitting the distribution for us. Based on the theoretical background of the Generalized Pareto Distribution (GPD), we know that with an inappropriate threshold, our parameter estimates will be inaccurate. Therefore, we examined so many quantiles and the goodness of fit to find a threshold where the fit becomes acceptable. After selecting the quantiles, it became apparent to us that as we progress through the years, the threshold decreases during the fitting process, reaching a point where the fit becomes acceptable. Subsequently, we selected the best values using the Anderson-Darling test [4], considering the p-value (it must be grater than 0.05), standard error, ensuring that the parameters obtained during the Anderson-Darling test did not significantly deviate from our fitting estimation (A few estimates could exhibit differences ranging from several hundred to even several thousand, thus, despite their p-values, we had to reject them) and the estimated value did not differ significantly from the original fastest time. We re-examined the fits based on the selected quantiles and assessed the diagnostic plots for these.

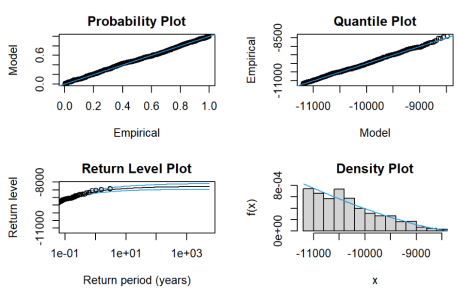

Figure 2: 1976 diag.

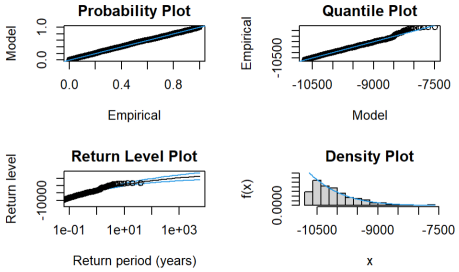

Figure 3: 2019 diag.

4 Analysis and estimation

In Table 2 and 3, we observe diagnostic plots for men in 1976 and 2019. The figure clearly shows that the probability plots and quantile plots are quite similar with minimal differences. The density plot also reflects the earlier point that in 2019, data is more densely packed over a shorter time interval than in 1976. If we examine the other plots, the density plots consistently indicate an improving trend year by year. After selecting the appropriate quantile, we obtained the parameters and the covariance matrix from the corresponding fit. From this, we generated a sample of 200 elements for each year according to a normal distribution with the given parameters. We used these parameters to make 200 estimates for each year. As a filtering step during the estimates, we trimmed the top and bottom 5 percentage. This process is illustrated in the Table 4 (with male data) and the 5 (with female data). The red line indicates the actual best running result, while the gray area represents where our estimates fell. Our initial assumption was that the gray area would narrow over time, but based on the obtained estimates, this assumption is not fulfilled (or only to a small extent).

5 Summary

This semester, our primary focus was on applying the Pareto distribution. We examined data from the Boston Marathon to understand how well the distribution could be fitted and explored optimal parameter selection methods to enhance our estimations. The project aimed to provide predictions for future race winners, and the application of extreme value theory, including the Pareto distribution, proved to be an effective method. There is hardly any trend in the forecasted best possible time times, so it looks that (assuming the current level of participants' and the course setting to hold) it might take quite a long time, till the winning time here will go below the magic 2 hours

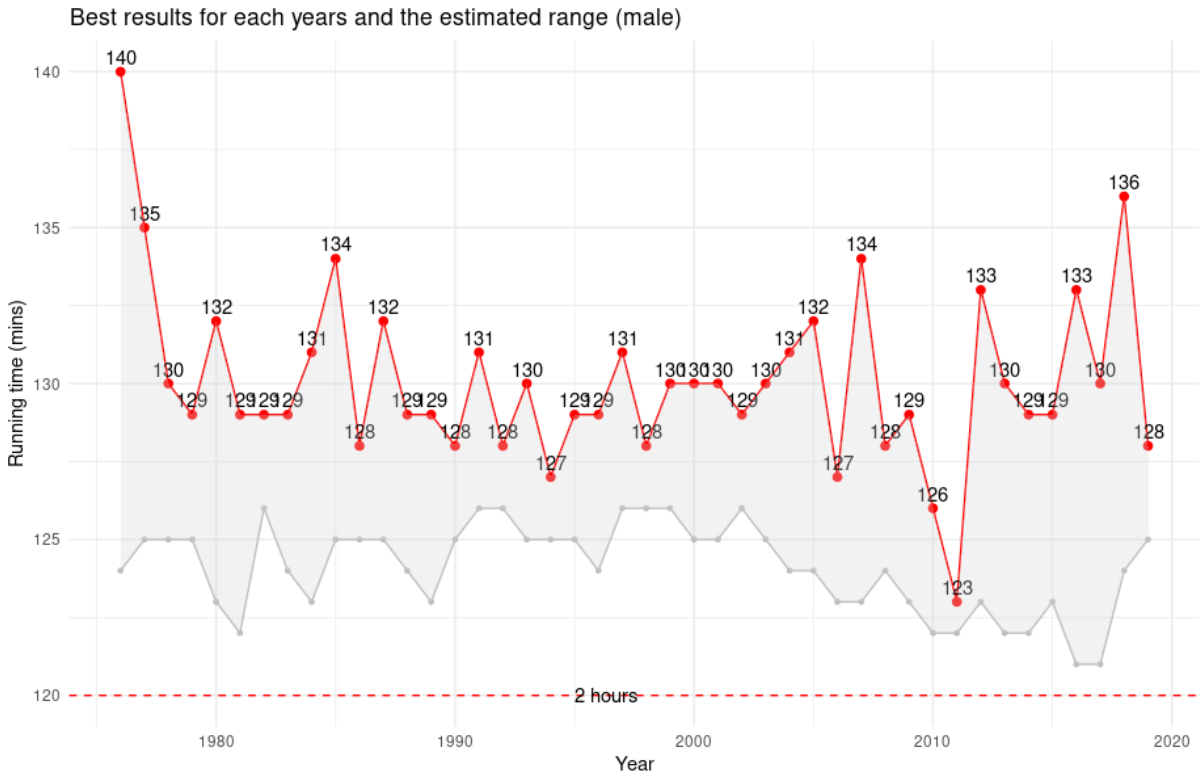

Figure 4: Estimated Male plot

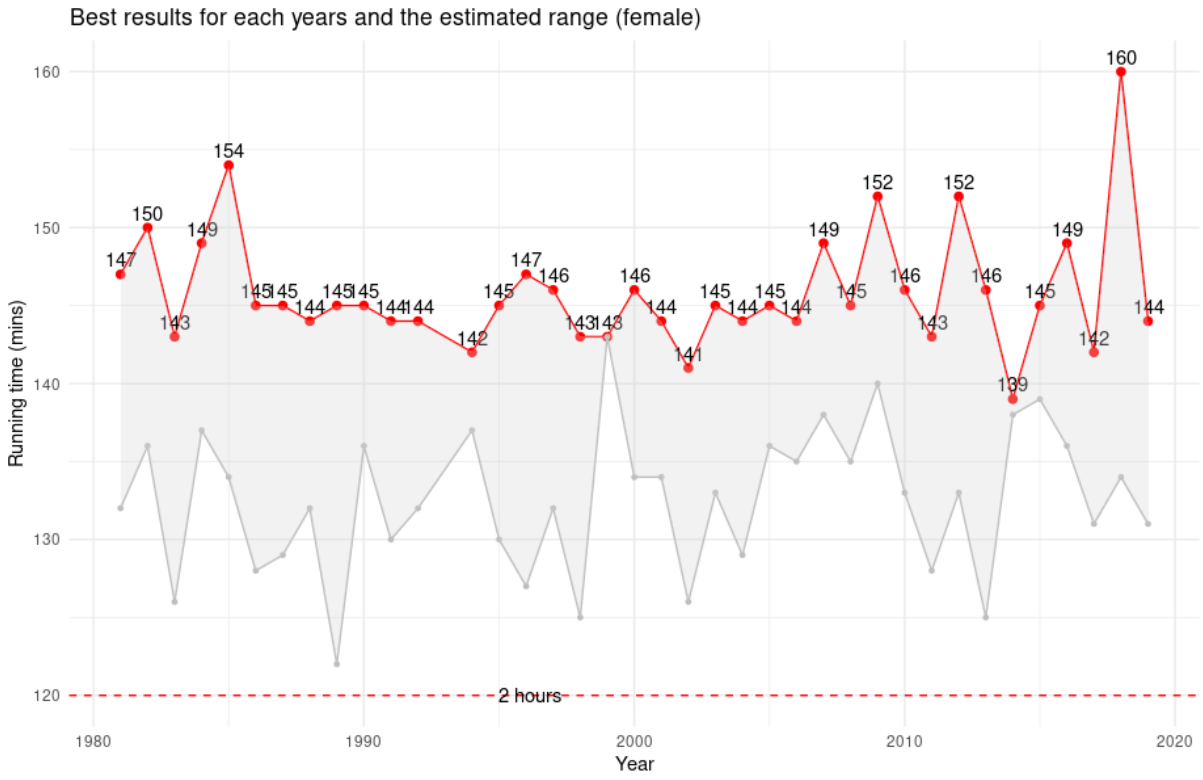

Figure 5: Estimated Female plot

6 Goals

It would be worthwhile to examine the relationship between the Berlin Marathon, analyzed in the previous semester, and the Boston Marathon used in the current analysis in the upcoming semester. Additionally, applying multidimensional models could facilitate the comparison of estimations between female and male participants. If there are opportunities throughout the semester to explore data from other races and compare them, considering various additional variables could also be insightful and a more accurate mathematical representation.

References

[1] *https://CRAN.R-project.org/package=ismev*.

[2] https://en.wikipedia.org/wiki/Pickands-Balkema-De_Haan_theorem.

[3] https://github.com/adrian3/Boston-Marathon-Data-Project/tree/master.

[4] https://search.r-project.org/CRAN/refmans/eva/html/gpdAd.html.

[5] https://www.baa.org/races/boston-marathon/results/search-results.