Contents

1 Introduction

In my third phase of my project we have followed to analysing sport data. we are working with the data of the Boston Marathon, where we have much more information about every competitor from 1898 to 2019. We continue where we left off in the previous semester. First, we make up for the missing things and algorithms that did not fit into the previous semester's documentation, then we move on to the construction of the two-dimensional model, and finally we draw a conclusion about the 3 semesters of work, in more detail about the last 2 semesters.

In my previous semester our goal have been to try to predict the best expected results of the following years based on our knowledge. For this purpose, we had used the Pareto distribution to see if there would ever be a sub-two hours Boston Marathon time could be expected in the near future. In this semester, our goal is to fit a two-dimensional model to the data, from which we form a copula and thus try to estimate a best result. We continue to use the R programming language

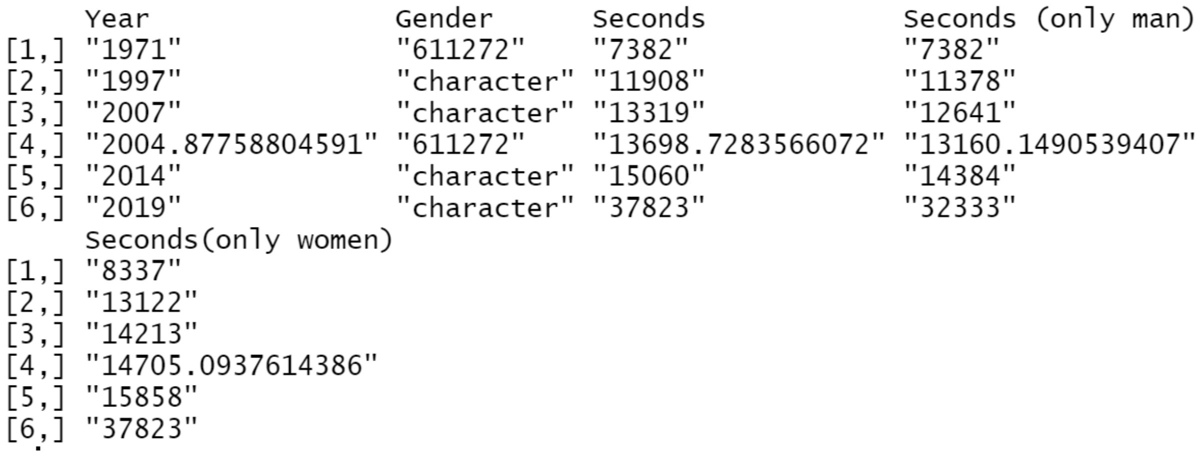
2 Summary of the previous semester

During the second phase of our project, we transitioned from analyzing the Berlin Marathon to focusing on the Boston Marathon due to the availability of more reliable data spanning from 1898 to 2019. Our aim was to predict the best expected results for future years based on historical data. To accomplish this, we utilized the Pareto distribution to explore the possibility of a sub-two-hour Boston Marathon time. Our analysis was conducted using the R programming language. Cleaning and processing the data proved to be a significant task, especially considering the lack of precision in early 20th-century records. Additionally, the official Boston Marathon page ([2] and [4]) did not separate data for half-marathoners or wheelchair marathoners, which could distort the final analysis. After addressing these issues and cleaning the data, approximately 614,000 rows remained, with a ratio of 1:2 women to men. Our primary goal was to fit the Pareto distribution for each year, necessitating careful consideration of the amount of data available. Trimming the data below a certain threshold proved effective in obtaining acceptable estimates based on the retained information. For male competitors, this threshold was set from 1975 onwards, while for females, it was from 1981. Running times were expressed in seconds and multiplied by -1 since the models aimed to maximize performance, whereas we sought to minimize running times. The Pickands–Balkema–De Haan theorem [1] is fundamental in extreme value theory, linking the tail behavior of a distribution to the Generalized Pareto Distribution (GPD). This theorem establishes that, for a wide range of distributions and a sufficiently high threshold, the distribution of exceedances properly normalized converges to the GPD. The GPD, characterized by its cumulative distribution function, is particularly valuable for modeling extreme events and assessing risk. We applied the Pareto distribution to our dataset annually, fitting it with quantiles ranging from 50 to 99 for each year. Selection of the appropriate threshold was crucial to ensure accurate parameter estimates. After selecting quantiles and assessing the goodness of fit, we obtained parameter estimates for each year. Diagnostic plots were used to validate the fits, ensuring they were acceptable. Diagnostic plots for male competitors in 1976 and 2019 showed minimal differences, indicating consistency in data distribution. After selecting appropriate quantiles, we generated parameter estimates and used them to make predictions for each year. Trimming the top and bottom 5% of estimates resulted in a more accurate representation of the data. The application of the Pareto distribution provided valuable insights into future race winners at the Boston Marathon. Despite challenges in data cleaning and processing, our analysis demonstrated the effectiveness of extreme value theory in predicting extreme events. However, a notable trend in forecasted winning times was not observed, suggesting that achieving a sub-two-hour marathon time may take considerable time. Overall, our project highlights the importance of utilizing advanced statistical methods, such as the Pareto distribution, in predicting extreme events. By leveraging historical data and mathematical models, we can gain valuable insights into future outcomes and trends in marathon running. At the end it turned out that there was hardly any trend in the forecasted best possible times, so it looks that it might take quite a long time, till the winning time here will go below the magic 2 hours.

3 Data

As we can see in the 1. figure, we are looking at data from 1971 onwards to 2019 because we did not have sufficient quantity and quality of data from previous years. This is because the world's best runners did not participate in this race from the beginning, but started taking part over time, which means that results from earlier years could significantly deteriorate our findings. Furthermore, we see that the best result ever recorded is 7382 seconds for men, while it is 8337 seconds for women. On average, they completed the marathon in 13699 seconds, which is approximately 4 hours ($\approx 3:48$).

In the 2. figure, we see the number of competitors starting each year. It is evident that from 1971 to 2010, the number of competitors increased exponentially. However, due to COVID-19, there was a decline in the number of competitors starting in 2020. As a result, we did not consider data from these years because realistically, competitions will return to such levels only after 2023, and some competitions were even canceled during



```
     Year              Gender      Seconds           Seconds (only man)
[1,] "1971"            "611272"    "7382"            "7382"
[2,] "1997"            "character" "11908"           "11378"
[3,] "2007"            "character" "13319"           "12641"
[4,] "2004.87758804591" "611272"   "13698.7283566072" "13160.1490539407"
[5,] "2014"            "character" "15060"           "14384"
[6,] "2019"            "character" "37823"           "32333"
     Seconds(only women)
[1,] "8337"
[2,] "13122"
[3,] "14213"
[4,] "14705.0937614386"
[5,] "15858"
[6,] "37823"
```

Figure .a: Summary

this period. We also see that the highest number of competitors was in 2014.

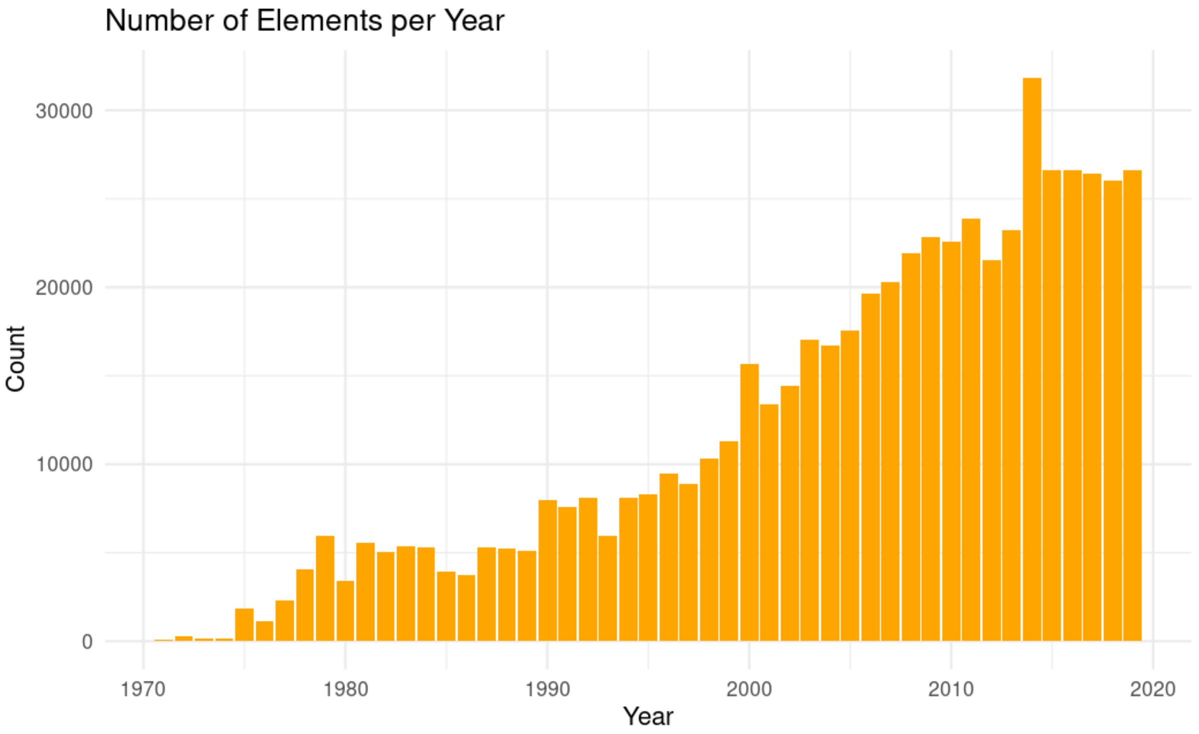


Figure .b: Histogram

4 Copula

4.1 Definition

Copulas are tools used to model dependency between variables. When studying dependencies among multiple variables, it is often insufficient to consider only the distributions

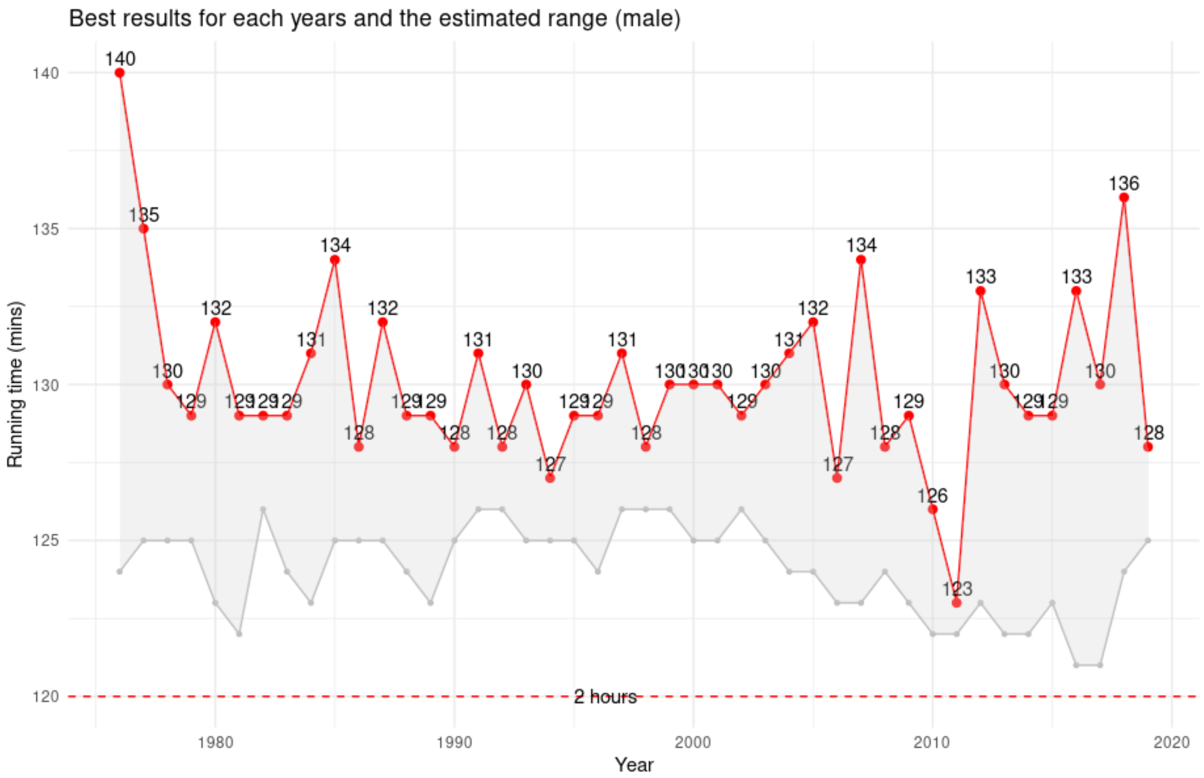


Figure 4: Estimated Male plot

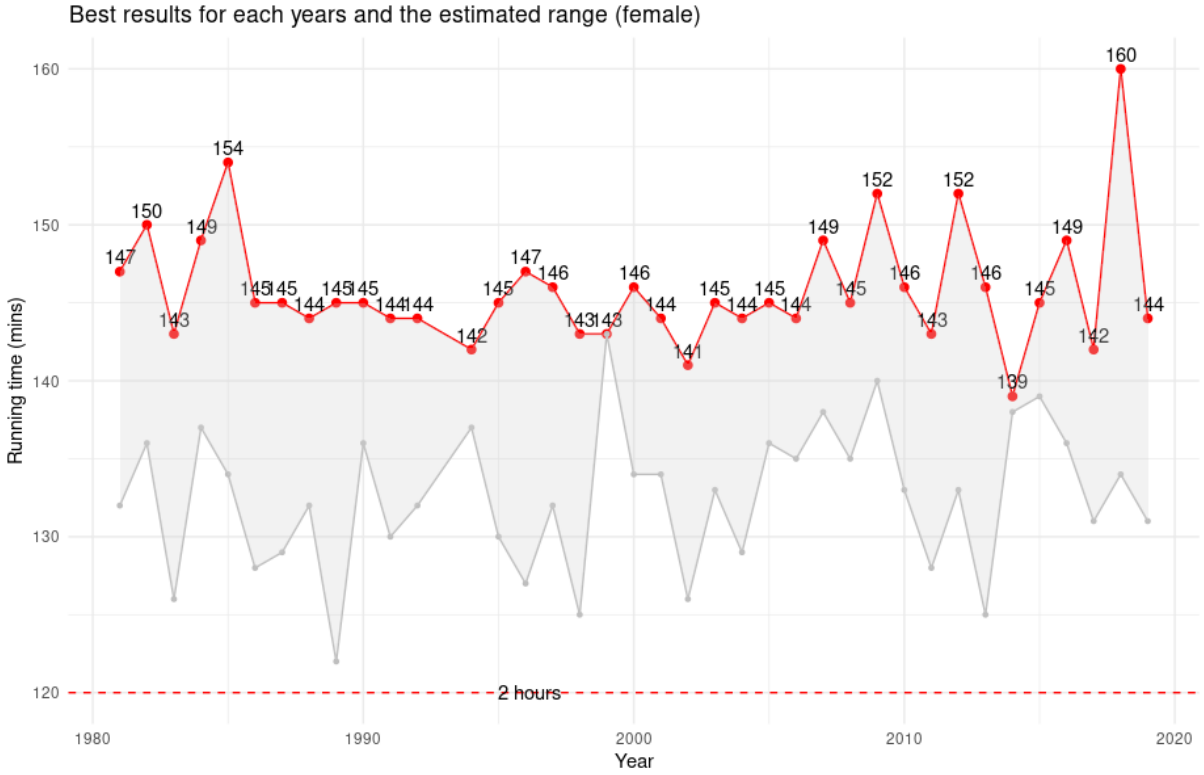


Figure 5: Estimated Female plot

of the variables and their correlation coefficients. Copulas are particularly useful in cases where there are non-linear relationships between variables or their distributions are complex. A copula is a multivariate distribution function that specifies the joint distribution of variables given the marginal distributions. Not to forget to mention that, they have uniform margins. The copula separates the dependency between variables from the marginal distributions, allowing the use of independent marginal distributions in dependency modeling. Let F be a multivariate distribution function, and $F_1, F_2, \ldots, F_d$ be the marginal distribution functions of the individual variables. The copula, denoted by C, is a function such that: $F(x_1, x_2, \ldots, x_d) = C(F_1(x_1), F_2(x_2), \ldots, F_d(x_d))$. [3]

Where $x_1, x_2, \ldots, x_d$ are the values of the variables. Thus, the copula arises from the joint function of the marginal distribution functions of the individual variables. Copulas are commonly used in various fields such as financial modeling, risk management, actuarial science, environmental science, and many others. Their benefits include more accurate modeling of complex dependency structures between variables, which is crucial for effective decision-making and risk management.

The Gaussian copula for a given correlation matrix $R \in [-1, 1]^{d\times d}$, the Gaussian copula with parameter matrix R can be written as

$$C_R^{\text{Gauss}}(u) = \Phi_R\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right),$$

where Φ^{-1} is the inverse cumulative distribution function of a standard normal and Φ_R is the joint cumulative distribution function of a multivariate normal distribution with mean vector zero and covariance matrix equal to the correlation matrix R.

The Frank copula is a popular member of the family of copulas, named after Józef Frank. It is particularly suitable for modeling symmetric dependency structures and can handle both positive and negative correlations.

The Frank copula is an Archimedean copula, which can be parameterized by a single parameter, θ, which indicates the degree of dependence. The closed-form expression for the Frank copula is:

$$C_\theta(u, v) = -\frac{1}{\theta}\log\left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}\right),$$

where u and v are the marginal distributions on the $[0, 1]$ interval.

Parameters:

- θ: The dependence parameter, where $\theta \in (-\infty, \infty)$ and $\theta \neq 0$.
  - If $\theta > 0$, the variables exhibit positive dependence.
  - If $\theta < 0$, the variables exhibit negative dependence.
  - If $\theta = 0$, the variables are independent.

5 Two-dimensional model

Continuing the previous work, our first goal was to create a two-dimensional model. We were able to do this by pairing the male and female competitors each year in such a way that the best male competitor was matched with the best female competitor, the second best female competitor with the second best male competitor, and so on.

For each gender, we needed a threshold based on which we could perform this matching and, since we needed to fit a Pareto distribution to this data, it was obvious to use the Anderson-Darling test used in the previous semester to determine with which values below the threshold we can work well, we know this with the p-value of the test. The Anderson-Darling Test is a goodness-of-fit test that determines how well your data fits a given distribution. Based on experience, we can make very few acceptable matches, in our case, even in the case of the chosen threshold of 9200 sec for the females and 8000 sec for the males, with a p-value of 0.04 for women. The p-value of the test for the males is smaller than we would have liked, but if a better fit for males is preferred, we will not get an acceptable female match.

6 Copula model

The 281 points after determining the limit are shown in 3. figure. Then we normalized the points into the $[0, 1]$ interval with the pgpd function, so we can get what copula model fits our data. For this, the BiCopSelect function, which gives us not only the copula model, but also the corresponding parameter and the Kendall-tau value. This is why we got the Gaussian copula and the second best copula was the Frank copula, which we talked about in more detail above. Then we generated a data set containing 1000 points for the Gaussian copula and for the Frank copula, a comparison of this a can be seen in the figure 4, where the generated Gaussian points are represented by red dots and the Frank points by black dots.

Following this, we conducted a brief test using gofCopula which is a goodness-of-fit tests for copulas based on the empirical process comparing the empirical copula with a parametric estimate of the copula derived under the null hypothesis. to assess whether the fitted copula is appropriate for our data. The returned values are depicted in the 5. figure, from which, based on the p-value, we can accept the null hypothesis that the distribution originates from a Gaussian copula.

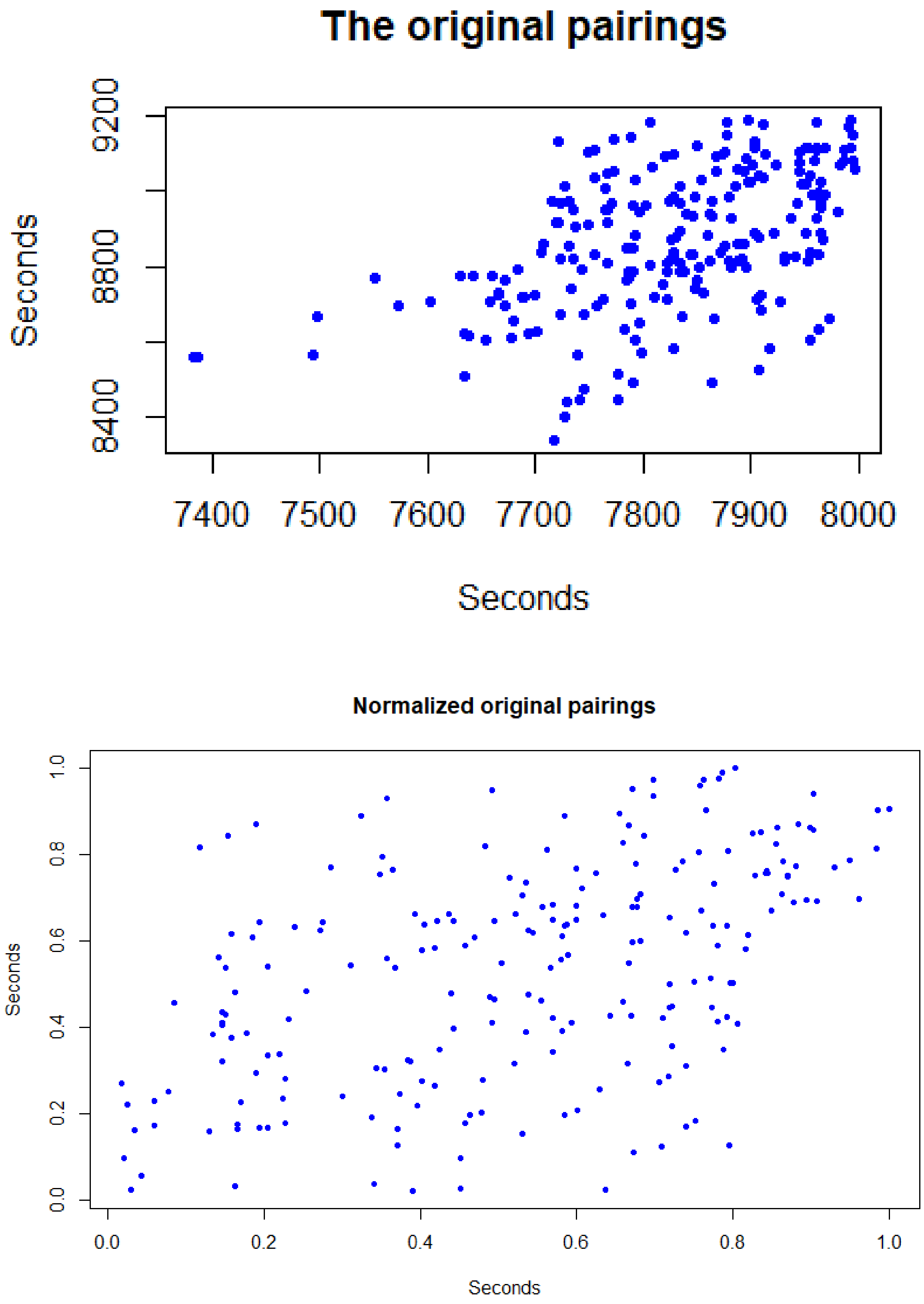


Figure 5: Normalized pairs

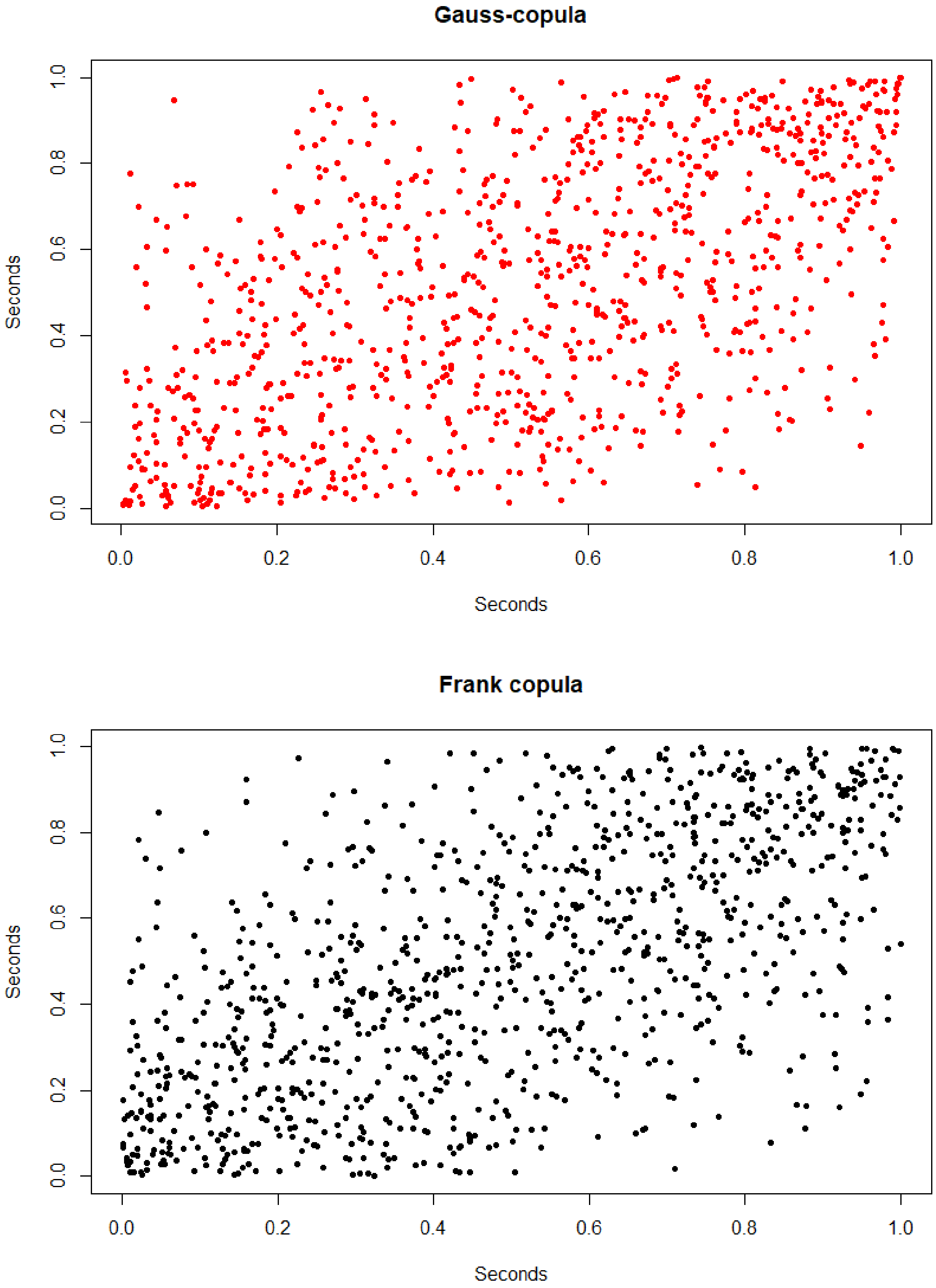
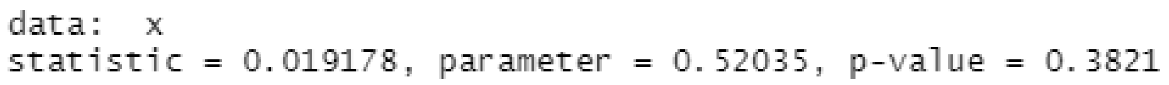


Figure 5: Frank copula

7 Evaluation

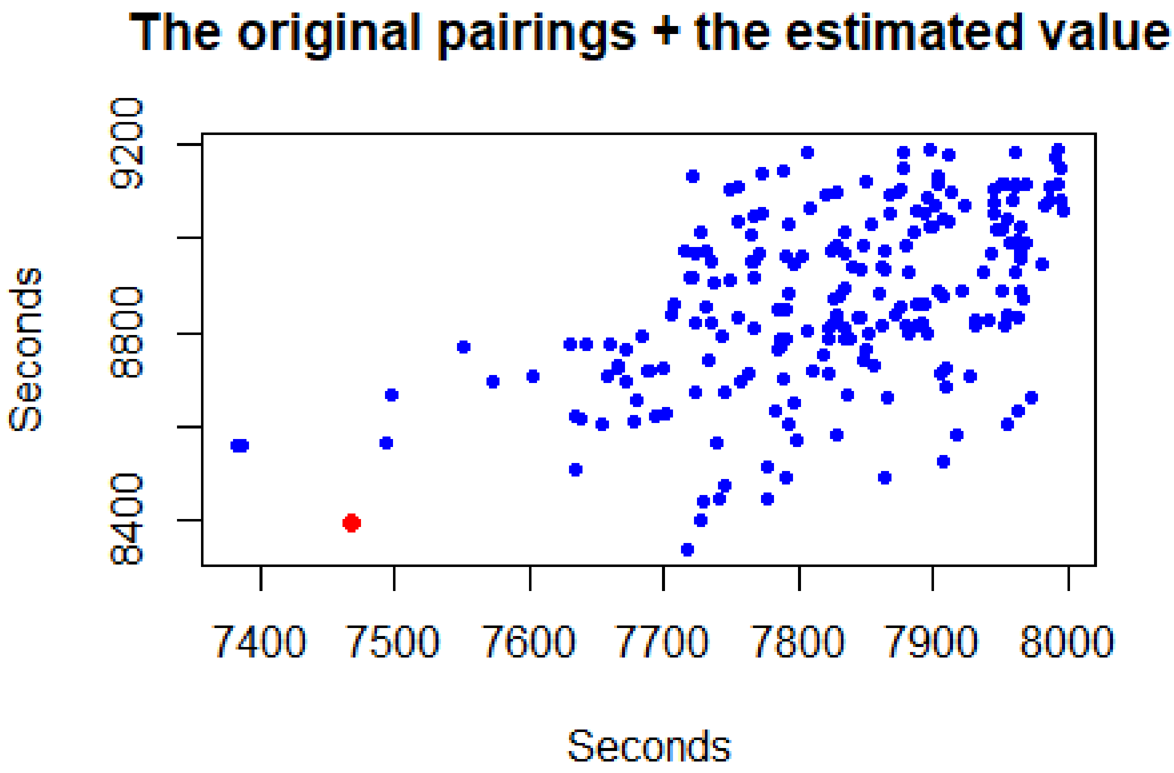
After the success of the test, we aimed to estimate a value that we could claim both genders will certainly achieve in the future. We based this on a 100-year level, as we



```
data:  x
statistic = 0.019178, parameter = 0.52035, p-value = 0.3821
```

Figure 4: Test

observed no expected improvement in results in the near future, as seen in the past semester. Subsequently, using the qgpd function (where the parameters are those used in the [0,1] transformation), we transformed the estimated value back into the original range. This is depicted in the 6. figure, where the blue dots represent our original pairings, and the red dot represents the estimated value, which we are confident both genders will achieve in the next 100 years. For males, this value is 7468 seconds, while for females, it is 8390 seconds. Thus, we can expect a race in the next 100 years where both the male and female winners achieve this result.



8 Conclusion

Our project endeavors to predict future outcomes of the Boston Marathon, leveraging historical data and statistical modeling techniques. Despite inherent data challenges, we successfully applied the Pareto distribution to gain insights into potential winning times. However, the elusive sub-two-hour marathon time remains a distant milestone. Continuing our work, we constructed a two-dimensional model by pairing male and female competitors. Copula analysis provided insights into gender performance dependency, with the Gaussian copula emerging as the preferred model. Validation tests further bolstered confidence in our modeling approach. With validated models at hand, we projected future outcomes, identifying threshold values for both male and female winners with confidence, but there are still tasks left (e.g. quantifying the uncertainty of the estimate, examining the dependence on the thresholds).

References

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

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

[3] https://mathworld.wolfram.com/SklarsTheorem.html.

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