UCLouvain

FACULTE DES SCIENCES

Faculté des sciences

Analysis of Combined Events Scoring systems

Models for comparing Athletic performances

Author: **Alex WLODAWER**
Supervisor: **Christian HAFNER**
Reader: **Philippe LAMBERT**
Academic year 2024–2025
Master [120] en science des données, orientation statistique

Preface

This master thesis has been prepared in partial fulfillment of the requirements for the master degree in Data Science (statistical orientation) delivered by the Université Catholique de Louvain.

I would like to express my gratitude to my supervisor, Pr. Christian M. Hafner. I sincerely thank him for the freedom he gave me to accomplish this thesis.

Contents

Chapter 1

Introduction

The question of who is the best athlete in the world arises frequently and is often the subject of debates among sports enthusiasts. For instance, is Usain Bolt, world record holder in the 100m and 200m, a better athlete than Javier Sotomayor world record holder of the High Jump? While both athletes are true legends in their respective sports, such a comparison is difficult to make. Determining who is "better" becomes an exercise of subjectivity, as their performances exist within entirely different domains, especially since they do not participate in the same sport.
One could try to answer this based on the number of gold medals, the number of weeks spent at the top of the world rankings, or even physiological and mechanical factors. However, each of these criteria has its limitations and cannot provide a definitive answer.

The decathlon provides a solution to this comparison challenge. Indeed, as a multi-event competition testing athletes across diverse disciplines, the decathlon relies on scores rather than raw performances. Instead of summing all the performances together, each performance is converted to a score using an event-dependent formula. Then, the points from the 10 events are summed to rank the athletes. Obviously, being better in one event than another would lead to a better score in that first event. Using this performance-to-points system, one could compare performances in different disciplines to evaluate which athlete is the best.

The existing scoring system used by World Athletics has been in place for decades but shows limitations when it comes to fairly comparing performances across the different disciplines of the decathlon. Thus, while the current system is functional, there is room for improvement in creating a system that reflects the value of each performance.

Basile Grammaticos (Grammaticos, 2007) proposed linking point attribution directly to the complementary cumulative distribution function (CCDF) of performances. His CCDF-based approach inspires the methodology developed in this thesis.

The goal of this research is to evaluate the accuracy of existing decathlon scoring model, particularly the World Athletics model. By examining alternative statistical models that could better capture the complexities of performance across different events, this thesis seeks to develop a more robust scoring system. This involves analyzing the limitations

of the current model, exploring new distribution models for athletic performances, and testing these models with real-world data. The outcome will be a new scoring system that better reflects the value of performance.

After evaluating the current model's fit using statistical and visual test. I will propose six candidates, the Weibull, Log-Normal, Gamma, Generalized Gamma, Burr XII and Inverse Burr, to model our track-and-field data. Following this presentation, I will present three different parameter-estimation methods, two frequentist approaches (maximum likelihood and quantile matching estimation) and a Bayesian approach. After fitting each model's density, I will conduct statistical tests on different regions of the distribution to assess the goodness-of-fit of the theoretical model. Using these statistical tests' results I will construct rankings in order to select the best candidate in this framework. Finally I will implement my model and demonstrate the new scoring system in action, analyzing Olympic-level and decathlon performances to illustrate its advantages.

Chapter 2

Context

2.1 Comparing Sport Performance

Competition has always been deeply rooted in human nature. In Ancient Greece, as sport began to gain importance, the pursuit of excellence helped shape the competitive spirit. Above all, it was victory that mattered most (Leonard, 2016). Second place was, indeed, considered the first loser. As sport evolved, so did measurement systems, making it possible to establish meaningful rankings and to recognize the achievements of competitors beyond the winner.

At that time, ranking systems worked well because athletes competed in the same events, enabling the audience to compare performances within a common context. But one fundamental question remains: how can we compare a thrower to a sprinter or a swimmer? Indeed, the indicators differ across sports. In a stadium race, performance is measured in units of time, while in a throwing event, performances are measured in units of distance. Beyond these unit differences, the complexity, duration, and overall nature of the sport may vary significantly. Nevertheless, the debate over the greatest athlete of all time has long persisted.

2.2 The Sport of Decathlon

Decathlon is a sport in which comparisons take a major space. The decathlon is a track and field discipline belonging to the family of combined events. This unique competition extends over two consecutive days, during which athletes must participate in ten predefined events in a specific order. Competing in a decathlon means that athletes must take part in four running events, three jumping events, and three throwing events.

More specifically, on the first day, athletes compete in the 100m race, long jump (LJ), shot put (SP), high jump (HJ), and 400m race. On the second day, they go on to compete in the 110m hurdles (110mH), discus throw (DT), pole vault (PV), javelin throw (JT), and the 1500m run.

Each performance in each individual event is converted into points using scoring tables established by World Athletics (formerly known as the IAAF/WA, International Association of Athletics Federations, the governing body for the sport of athletics). After

each event, the athlete's points are accumulated to determine their overall score and final ranking in the competition.

Other similar types of combined events exist in athletics, depending on the athlete's gender, location, and the number of events. However, I will mainly focus on the decathlon.

The decathlon is primarily a test of versatility, endurance, and consistency, where athletes compete as much against themselves as against their opponents. Unlike traditional races, where the objective is simply to finish first, the decathlon aims to reward overall performance across ten disciplines. Athletes aim to maximize their individual scores in each event, accumulating as many points as possible rather than focusing solely on outperforming their competitors in one single discipline. This structure emphasizes the importance of being a well-rounded athlete rather than excelling in just one or two events. Ultimately, the winner is the competitor who demonstrates the highest overall ability across all ten events, making the decathlon a true test of all-around athleticism. But should it be the case? I will discuss and analyze that further in this thesis.

Decathlon is particularly useful for developing a system that enables comparisons between results from different events. In the decathlon, every performance is awarded a certain number of points, typically ranging from 0 to 1300 points.

For a given performance, the score p is calculated using one of the following equations (Spiriev, 2022):

$$\begin{aligned}
&\text{Track}: p = a(b - T)^c, \quad \text{where } T \text{ is time in seconds; e.g., 10.43 s for 100 meters} \\
&\text{Jumps}: p = a(D - b)^c, \quad \text{where } D \text{ is distance in centimeters; e.g., 808 cm for LJ} \\
&\text{Throws}: p = a(D - b)^c, \quad \text{where } D \text{ is distance in meters; e.g., 16.69 m for Shot Put}
\end{aligned}$$

with:

- a : the scaling parameter.
- b : the baseline parameter.
- c : the progressiveness parameter that rewards elite performances more importantly than average or low performances.

Those formulas have been subject to many changes across time during the 19th and 20th centuries to comply with multiple requirements in the field of physiology, physics, and statistics, but also according to the requirements of the experts, athletes, and coaches (Grammaticos, 2007).

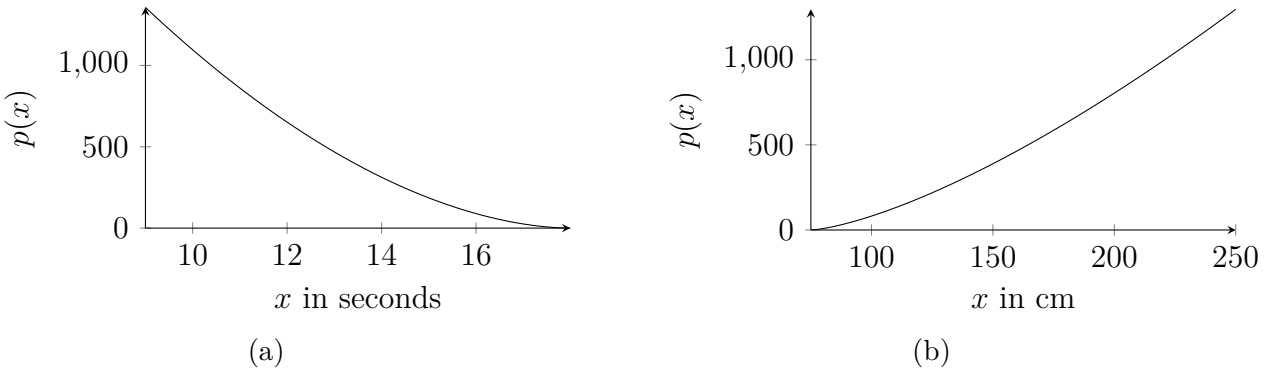


Figure 2.1: Current points attribution for (a) 100m and (b) High Jump

2.3 Key Requirements for Performance Comparison in Decathlon

In practice, comparing performances in the decathlon involves a twofold process (Grammaticos et al., 2022).

2.3.1 Vertical Comparison

The first comparison occurs within a single event and is quite intuitive. For instance, two performances in the 100m can easily be compared. The best athlete is simply the one with the fastest time.

At first, a linear relationship between time and points might seem sufficient. However, this is not the case. One must consider that top performances are significantly harder to improve than average ones. For example, reducing a 100m time from 20 seconds to 19 is much easier than reducing it from 10 to 9 seconds. This reasoning applies across all events in track and field.

This is referred as the vertical comparison. Indeed, exceptional performances should be rewarded with an exceptional number of points. In the previously presented formulas, this progressiveness is reflected by the exponent c being greater than one.

2.3.2 Horizontal Comparison

As previously mentioned, the decathlon is a test of versatility, designed to reward well-rounded athletes, those who perform consistently across all disciplines. An athlete who excels in only one event (e.g., the 100m) might earn a large number of points there, but if he performs poorly in other events like the javelin throw, he would receive very few points. Due to the point assignment system, such an athlete would not rank well compared to someone who performs reasonably well across all events.

As an illustrative example, I consider Derek Drouin, Olympic and World Champion in the high jump. In April 2017, he participated in a decathlon in the United States and

delivered an extraordinary jump of 2.28m, setting the all-time best high jump performance within a decathlon. However, due to his over-specialization in high jump, he was below average in the other events and finished with a total score of 7150 points.

To demonstrate that decathlon rewards all-around ability, I compare his result with Jente Hauttekeete's best performance, the Belgian U18 decathlon record holder.[1]

Event	**JH performance**	**JH point**	**DD performance**	**DD point**
100	11.16	825	11.42	769
LJ	6.81	769	6.75	755
SP	15.38	813	12.10	612
HJ	2.00	803	2.28	1071
400	50.51	791	51.81	733
110h	14.36	929	14.47	916
DT	42.67	719	34.85	561
PV	4.30	702	3.80	565
JT	48.58	568	49.39	580
1500	4.49.60	621	4.54.51	592
Total	-	7540	-	7150

Table 2.1: Points and performances comparison between Jente Hauttekeete (JH) and Derek Drouin (DD)

From this small example, I can clearly see that in the decathlon, it is more valuable to be complete and perform well across all events than to be exceptionally strong in just one. Jente Hauttekeete did not outperform in any single event but received consistent rewards across all disciplines.

Now the question arises: How should the different events be compared to one another? This is the essence of horizontal comparison. In the current formulas, this comparison is handled through the parameter a, which serves to bring all point scores onto a common scale. Indeed, by adjusting the magnitude of a, one could favor a specific event over another. Conversely, horizontal comparison in the decathlon would be considered fair if, and only if, a performance at an elite level say, in the top 5 percentile of the 100m earned the same number of points as a performance in the top 5 percentile of the long jump or the discus throw.

I need to mention that a world record or some extreme elite performances should in fact be awarded with a larger number of points than an elite performance. Indeed a 100m ran in 10.00 sec is extraordinary, but that same 100m ran in 9.58 is on another planet and deserves more points. Maybe Derek Drouin should have received more points for its high jump as he jumps nearly 30cm more that its opponent. Note that elite-level decathlete could for some of them participate to individual events and perform as well as specialists. This causes that some of them are indeed very well awarded for some performances. Nevertheless, the best decathletes in the world still perform incredibly well in all ten events.

[1]Results retrieved from the World Athletics website

2.4 Current formulas by World Athletics (IAAF/ WA)

The current tables were defined in the 1980s by V. Trkal and K. Ulbrich, under the request of the Technical Committee of World Athletics (IAAF/WA), in response to feedback from athletes, coaches, and criticism of earlier scoring tables. V. Trkal and his team proposed nine principles for designing these tables (Trkal, 2003). I believe that some of these principles are valuable more generally when comparing performances across different disciplines.

- (2) The results in different disciplines that are evaluated with approximately the same point value should be comparable in terms of quality and the difficulty of achieving them.
- (3c) The tables in all disciplines should be very slightly progressive.
- (4) The tables must be applicable to combined events for beginners and juniors as well as for elite athletes.

I describe here an analysis of the current formulas used by World Athletics in a formal way:

- Formulas used for throwing and jumping events (Long Jump, Shot Put, High Jump, Discus Throw, Pole Vault, Javelin Throw):

$$p(x) = a(x - b)^c, \tag{2.1}$$

  where x is a measure of distance in meters for throwing events and in centimeters for jumping events.

- Formula used for track events (100m, 400m, 110mH, 1500m):

$$p(x) = a(b - x)^c, \tag{2.2}$$

  where x is a measure of time (in seconds).

Here, to simplify the interpretation of the formulas, I find it useful to replace the raw performance x in track events with the athlete's average speed. This transformation is insightful because, unlike in jumping or throwing events where we seek to maximize a distance or height, in running we aim to minimize time. By expressing performances in terms of speed, all disciplines can be unified under the same formula. Under this transformation, Equation (2.2) would become equivalent in form to Equation (2.1).

Event	a	b	c
100m	25.4347	18.00	1.81
Long Jump	0.14354	220.0	1.40
Shot Put	51.39	1.50	1.05
High Jump	0.8465	75.0	1.42
400m	1.53775	82.0	1.81
110m Hurdles	5.74352	28.5	1.92
Discus Throw	12.91	4.00	1.10
Pole Vault	0.2797	100.0	1.35
Javelin Throw	10.14	7.00	1.08
1500m	0.03768	480.0	1.85

Table 2.2: IAAF/ WA Decathlon Scoring Parameters

For running events, as previously discussed, I need to change the form of the formula so that it becomes monotonically increasing in speed v, rather than decreasing in time t. Specifically, the transformation rewrites the original formula $p(t) = a(b-t)^c$ as:

$$p(v) = a'(v - b')^{c'}.$$

Here, the new parameters a', b', c' differ from the original a, b, c, and can be estimated by minimizing the absolute difference between the original and transformed scoring functions (L1 norm). This can be done using the following optimization objective:

$$\min_{a',c'} \sum_i \left| a'(v_i - b')^{c'} - a(b - t_i)^c \right|, \tag{2.3}$$

with:

- $b' = \frac{d}{b}$, where d is the race distance,
- t_i and v_i are corresponding times and speeds (i.e., $v_i = \frac{d}{t_i}$),
- The sum runs over a densely sampled range of times, from the current world record up to some baseline time b.

I also computed the R^2 between the old and new scoring function to see whether they align. $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ with y_i, the true scores, $\hat{y}_i$, the estimated scores.

In order to maximize my fit, I also find the R^2 values when minimizing (2.3) with a L2 norm. I find equal R^2 values up to the third digits.

Event	a'	b'	c'	R^2
100m	178.29708	5.54	1.23946	0.9969
400m	198.29314	4.88	1.23090	0.9968
110mh	190.95337	3.86	1.19563	0.9950
1500m	263.63712	3.13	1.13175	0.9943

Table 2.3: IAAF/ WA Running Scoring Parameter based on the speed and R^2's assessment of the transformation

Given the R^2 obtained by this analysis, the transformed parameters seem to be reflecting the original one. Using this new parametrization and the common Formula (2.1), I can easily interpret the meaning of each parameter:

- a, the Scaling factor :
  - Determines the overall magnitude of the scores in a range of 0 to 1300 points.
  - It serves to normalize the event so that different disciplines contribute proportionally to the total decathlon score.
- b, the Baseline performance :
  - Represents a reference performance corresponding to a null score, shifting the scoring function, under which the score is 0.
  - It ensures that realistic performances yield positive scores.
- c, the Exponent for performance growth :
  - Determines how the score increases as performance improves.
  - Higher values of c mean that small improvements in performance result in larger increases in score, especially at elite levels.
  - It defines the progressiveness of the scoring function.

2.4.1 Mathematical derivation of scoring parameters of World Athletics

Very few insights are available regarding the data used to compute the parameters of the scoring function. However, here is a realistic attempt to reconstruct how the parameters may have initially been calculated.

Selection of the baseline performance b

The parameter $b > 0$ must have been chosen based on historical performance data, serving as a reference threshold corresponding to a null score:

$$b = \text{performance}_{\alpha}(\text{ distribution}),$$

where α represents a chosen percentile. This approach ensures that:

- In running events, b corresponds to a reference time.
- In jumping and throwing events, b corresponds to a minimum threshold distance or height.

For instance, Harder arbitrarily suggested that the null score should be assigned to a performance achievable by 95% of the population (Harder, 2001). This requires not only knowledge of the sport and the statistical distribution of performances (Grammaticos

et al., 2022), but also a clear definition of the population under consideration.

Indeed, should the distribution only consider decathletes, as they are the first users of the scoring table? Or should it include all track and field athletes, assuming that they could all potentially compete in a decathlon? Even more broadly, should it consider the entire population, given that virtually anyone can attempt to run 100m, even if it takes them 2–3 minutes or more? If so, should such performances be rewarded with any points at all?

J. G. Purdy defined the null score b supported by physically grounded arguments:

- For running events, Purdy argued that "running is not walking" (Purdy, 1974, 1975, 1977), and proposed the expression:

$$v_0 = 2 - \frac{d}{99950} \quad \text{(m/s)},$$

  where d is the distance of the running event.

- For the long jump, he suggested that the null score could correspond to the length of a large walking step. For the high jump, he proposed using the height of the center of gravity, typically around $h_0 = 80\,\text{cm}$.

- For throwing events, the derivation was more complex, and Purdy arbitrarily fixed the null-score distance at 5% of the distance corresponding to 1000 points.

It turns out that using this set of assumptions results in null scores that are very close to those adopted by World Athletics (WA).

Calibration of the exponent c

The exponent $c > 1$ was chosen to model the nonlinear relationship between performance and the difficulty of improvement. Setting up two reference pairs of performance-points (x_i, P_i). I can find c such that it is the difference in slopes between the logarithms of the performances and the transformed performance values, given by the formula:

$$c = \frac{\log P_1 - \log P_2}{\log(x_1 - b) - \log(x_2 - b)}, \quad x_1, x_2 \geq b,$$

where:

- (x_1, P_1) represents a world-class performance (e.g., a world record, typically yielding $P_1 \approx 1200$–1300 points),

- (x_2, P_2) represents a benchmark elite performance (e.g., a score in the range of 900–1000 points).

This calibration ensures that incremental improvements at elite levels lead to exponentially larger gains in points, reflecting the increasing difficulty of marginal improvements at high performance levels required by the verticality comparison paradigm.

Determination of the scaling factor a

Once b and c were determined, the scaling factor $a > 0$ was computed to normalize the points distribution such that a reference elite performance corresponds to a predefined score:

$$a = \frac{p}{(x-b)^c}, \quad x > b,$$

where p is the target score (typically 1000 points), and x is the associated performance level.

To ensure fairness across all events, I assume that a must have been fine-tuned using least-squares minimization over a representative set of historical decathlon performances:

$$\min_a \sum_i \left(p_i - a(x_i - b)^c\right)^2,$$

where (x_i, p_i) are performance–score pairs drawn from competition data.

Final parameter validation

The derived parameters a, b, c were validated against historical competition results to ensure that:

- The scoring system fairly rewards well-rounded athletes across all disciplines.
- The point distribution remains consistent and comparable across different events.
- The system is robust and scales appropriately as performance levels improve over time.

2.5 General design of the performance/point relation

In a general context, it appears intuitive that comparing two performances from different sports requires the use of scoring tables that assign points to those performances. As previously mentioned, these tables must be both vertically and horizontally informative.

Horizontally, Harder suggested in 2001 (Harder, 2001) that "two performances are equivalent when the number of athletes who obtain them is comparable", where "number of athletes" refers to the practitioner of the latter event. It seems only fair, for comparison purposes, that the best achievable performance in the long jump should translate to the same number of points as the best performance in the 1500m.

Vertically, we need a function linking performance to points that makes sense according to the principles discussed earlier.

By combining these two requirements, Harder proposed a quantitative approach that assigns:

- 100 points to a performance that 50% of practitioners can achieve,

- 200 points to a performance that 5% can achieve,
- 300 points to a performance that only 0.5% can achieve, and so on.

To implement such a scheme, assumptions about the distribution of performances are required. Hence, following Harder's intuition, I can link the attribution of points to the underlying distribution.

2.5.1 Mathematics

First, I need to define the set of possible performances, which in the case of track and field can generally be considered as ranging from 0 to $+\infty$. Of course, this range should be adjusted for each specific event. Within this positive domain, some values are clearly unrealistic, running 100m in less than 5 seconds, for example, is far beyond what is physically achievable.

From this set of possible values, I can define a probability density function, $f(x)$ that provides a likelihood of a performance.

The cumulative distribution function (CDF), denoted by $F(x)$, gives the probability that a randomly chosen athlete achieves a performance less than or equal to x. For the purposes of this approach, Grammaticos, in (Grammaticos, 2007), worked instead with the complement of the CDF defined as:

$$G(x) = 1 - F(x), \quad x \in \{\text{possible performances}\}, \tag{2.4}$$

which is referred to as the CCDF (Complementary Cumulative Distribution Function), or the survival function.

Following Harder's work, Grammaticos suggested a general formula linking the number of points to the complement of the cumulative distribution function:

$$p(x) = -\ln(1 - F(x))/\lambda = -\ln(G(x))/\lambda, \tag{2.5}$$

where λ is a positive constant used to scale the result and ensure that the score remains on a meaningful and positive scale. The motivation behind scaling the relation with λ was to ensure that the sum of points scored by world-class athletes should remain approximately the same (Trkal, 2003) comparing to the previous tables.

Using these relations, I can construct different scoring systems by fitting the collected performance data to various candidate distribution models. This function will allocate to larger smaller values of the CCDF, $G(x)$, larger amount of points in a progressive way.

2.5.2 Multiple distribution models

In this study, I aim to analyze different performance distribution models in order to assess the fairness and accuracy of scoring systems. By examining real-world performance data collected from French athletics competitions during the 2010s and 2020s, I will evaluate

how the current IAAF (WA) model fits empirical data and suggest a new optimized scoring system.

To construct a fair and robust scoring system, I consider multiple probability distributions that could realistically model athlete performances. The idea behind examining new distribution model was designed by Grammaticos himself and J.G. Purdy. They suggested a modeling function that explicitly accounts for the fact that both practitioners and non-practitioners can attempt an event. A revised and extended version of this model was later studied by the same group of statisticians (Grammaticos et al., 2022).

Chapter 3

Modeling

As mentioned in the previous section, using Harder's quote: "Two performances are equivalent when the fraction of the practitioners who obtain them is comparable", I can create a fair scoring system by linking the appropriate point attribution function, $p(x)$, to the Complementary Cumulative Distribution (CCDF), $G(x)$.

3.1 Data

To illustrate the comparison of athletic performances across all potential participants, I compiled track and field data from individual events of all ten decathlon disciplines from 2015 to 2025, sourced from the French Federation of Track and Field. Instead of working on decathlon performance only, I decided to work on a larger scope in order to award a score to each performance, and in a second time to review this new system on decathlon data. This analysis has the potential to contribute to a global ranking of athletes based on their best performances. For example, World Athletics holds an annual ceremony to recognize top athletes across various criteria, including performances. The dataset includes each athlete's best performance over the ten-year period (2015–2025) among members of French track-and-field clubs who competed in at least one official meet.

For the ease of the following, I decided to transform all time marks into speed marks in order to have positive relation between scores and performances.

Event	Count	Mean	Best FR	Best World	Decathlon World Best
100m (m/s)	33,506	7.99	10.21	10.44	9.88
LJ (cm)	27,953	517	826	895	845
SP (m)	10,149	8.69	21.36	23.56	19.17
HJ (cm)	22,254	152	233	245	228
400m (m/s)	20,950	7.10	8.88	9.30	8.89
110H (m/s)	3,381	6.26	8.43	8.59	8.23
DT (m)	8,798	24.86	68.96	74.35	57.70
PV (cm)	8,966	295	605	627	576
JT (m)	11,750	32.15	87.58	98.48	79.80
1500m (m/s)	22,170	5.41	7.17	7.28	6.28

Table 3.1: Dataset descriptive statistics and World Record and Decathlon World Best.

3.2 Current model assessment

In this section, I will first assess the current World Athletics scoring system using provided parameters and the underlying distribution derived under Harder's assumptions (Grammaticos et al., 2022)

3.2.1 Underlying model

Let's examine back the initial scoring formula (2.1)

$$p(x) = \begin{cases} a(x-b)^c, & x \geq b \\ 0, & x < b \end{cases} .$$

Reversing Equation (2.5), I easily find that :

$$G(x) = e^{-\lambda p(x)}, \quad x \in \{\text{possible perf.}\}. \tag{3.1}$$

Using the World Athletics (WA/IAAF) Equation (2.1), the complementary cumulative distribution is defined by :

$$G(x) = e^{-a\lambda(x-b)^c}, \quad \forall x \geq b. \tag{3.2}$$

From that expression I can recover the formulation of the PDF, the distribution of the performance in a specific event.

$$f(x) = \lambda e^{-\lambda p(x)} p'(x). \tag{3.3}$$

Deriving $p(x)$, as $p'(x) = ac(x-b)^{c-1}$, I obtain a distribution of the form :

$$f(x) = \lambda ac(x-b)^{c-1} e^{-\lambda a(x-b)^c}, \quad \forall x \geq b. \tag{3.4}$$

In the Equation (3.4), I can already see that the parameters λ and a are only jointly identifiable because their effects are twisted together in both the exponential and power terms. However, since both parameters influence the overall behavior of the distribution together, they cannot be separated from each other without additional information. Consequently, λ and a must be estimated together, as changes in one parameter can offset changes in the other.

3.2.2 World Athletics Modeling

The Weibull distribution

The previous probability density function (3.4), is a specific case of the Weibull distribution.

This distribution is a continuous probability distribution named after Waloddi Weibull. The theoretical ground of the following has been found in "Generalized Weibull Distributions" (Lai, 2014).

In its general settings, the probability density function (PDF) of the Weibull distribution is given by:

$$f(x;\lambda,c) = \begin{cases} \frac{c}{s}\left(\frac{x}{s}\right)^{c-1} e^{-(x/s)^c}, & x \geq 0 \\ 0, & x < 0 \end{cases} ,$$

where:

- $s > 0$ is the scale parameter controlling the spread of the distribution.
- $c > 0$ is the shape parameter that determines the distribution's form.

The cumulative distribution function (CDF) of the Weibull distribution, which gives the probability that a random variable X is less than or equal to a given value x, is:

$$F(x) = 1 - e^{-(x/s)^c}, \quad x \geq 0.$$

The Weibull distribution is commonly used as a distribution to describe survival function, when we are interested in data that has not happened yet, in other words, $1-F(x)$ (Mudholkar et al., 1996). In this framework, it is particularly helpful to determine the proportion of people that are able to attain a level of performance. That fraction of the population decreases as the performance improves.

$$G(x) = 1 - F(x) = e^{-(x/s)^c}, \quad x \geq 0.$$

This is of a similar type as equation (2.5).

3.2.3 Performance distribution using Weibull

The Weibull distribution I've described here is relevant for human athletic performances modeling for multiple reasons:

- As I mentioned in Section 2.3.1 regarding athletic performances, achievements follow a non-linear improvement curve, small gains become exponentially harder at elite levels.
- Performance distributions are often skewed, as events are, in most cases, achievable by many people, regardless the age category, or the level of training. Weibull distribution can capture this asymmetry.

In this framework, I observe that even though the Weibull distribution corresponds to the initial density estimation, I need to refine the classical formula of the Weibull distribution, which normally includes two parameters, to a three-parameter distribution model. This allows to shift the scoring function to the right in order to allocate the points starting from a given performance under which any performance will be equal to 0. Also, I proceed to a reparameterization of (3.7) according to (Grammaticos et al., 2022) with $a' = \lambda \times a = \frac{1}{s^c}$

First, recalling equation (3.4), $\lambda \times a$ is actually a scaling parameter. To avoid identification problems and ensure clarity in this model, I combine a and λ into a single

scaling parameter a', resulting in the revised form of the Weibull distribution. This avoids confusion between the two parameters and ensures proper model fitting. Thus, I rewrite the WA-Weibull model as:

$$f(x) = a'c(x-b)^{c-1}e^{-a'(x-b)^c}, \quad x \geq b, \tag{3.5}$$

with :

- a' : The new combined scaling parameter.
- b : The location parameter. In other words, I am facing a Shifted Weibull distribution where $x \geq b$. In this context, b represents the minimum performance needed to be eligible for scoring. In World Athletics, this is referred as the null score performance, the performance at which an athlete earns zero points. It shifts the entire distribution to the right.
- c : The shape parameter that controls the skewness and the steepness of the distribution. The larger the c, the narrower the peak and the faster the decay.

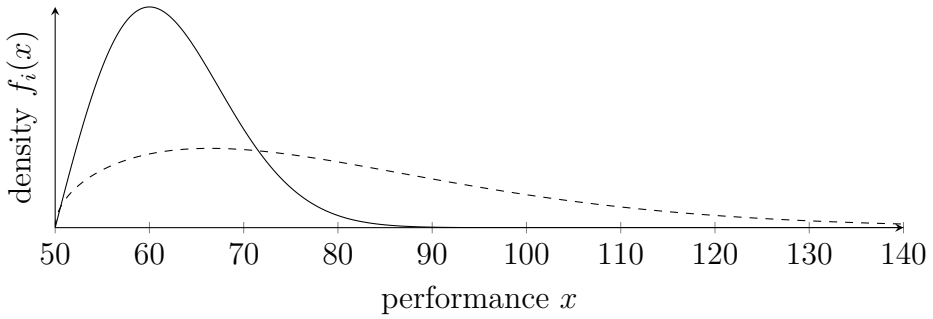


Figure 3.1: Impact of the c parameter on the shifted Weibull distribution.
$c_{\text{plain}} > c_{\text{dashed}}$, with $b = 50$, the minimal performance

Also, c dictates the progressiveness of the score attribution to the performance.

- $c > 1$, performance improvements become increasingly difficult. It implies the progressiveness in the scoring.
- $c = 1$, reduces to exponential law. No progressiveness.
- $c < 1$, implies very easy improvements. Regressiveness, I try to avoid this behavior in performance assessment, for plausible performances at least.

In the three-parameter Weibull distribution, a' controls the spread of the distribution. Higher values of a' result in a narrower distribution, causing performances to cluster near the mode. On the other hand, a lower a' leads to a wider distribution, with performances being more spread out across the scale.

In terms of scoring, a' directly influences the overall point distribution by calibrating how many athletes fall into high-score or low-score bins.

3.2.4 Original Weibull model

With the knowledge of the parameters used by World Athletics, I can see how the current scoring system combined with (2.5) fits the true distribution of the performances, for each event, at least in France. Indeed comparing, the empirical distribution, $f_n(x)$ of the performances with the theoretical distributions implied by World Athletics would suggest whether the parameters used by World Athletics are appropriate to compare two performances made in two different disciplines. After a simple look at the (3.5), I directly see that I would need a' to compute the theoretical PDF of the performance. Indeed, the value of a is known from the WA scoring rule, but the scale parameter a' (or equivalently λ) is not known. Therefore, to test whether the observed data follows the WA-implied Weibull shape, hence I must estimate a' from data while keeping b and c fixed.

To estimate a', I must use the relation between the Weibull and its reduction a exponential distribution. Given a sample of performances $x_1, \ldots, x_n$, I define the transformation:

$$Z_i = (x_i - b)^c, \quad x_i \geq b.$$

Under the Weibull model, these values should follow an exponential distribution, (Johnson et al., 1994):

$$Z \sim \text{Exponential}(a'). \tag{3.6}$$

This previous result also gives us an estimate, $\hat{a}'$ for a' using maximum likelihood estimator (MLE). Given a sample $x_1, \ldots, x_n$, I can compute:

$$\hat{a}' = \frac{1}{\bar{Z}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}(x_i - b)^c}.$$

This provides an efficient and scale-independent way to estimate a' without needing to refit the full Weibull model, under the assumption that b and c are fixed.

Fit of the World Athletics Weibull model

Now that I have all the parameters available, I can perform tests in order to see the fit of the WA-derived distribution of performances and hence see whether the proposed model by WA (IAAF) follows the requirements to assign points fairly across all disciplines.

First let's assess visually whether the WA-implied Weibull model (with estimated $\hat{a}'$) fits the empirical distribution of performances, I compare:

- The empirical density, obtained using a histogram,
- The empirical density, obtained using a kernel density estimate (KDE),
- The theoretical PDF given by (3.5)

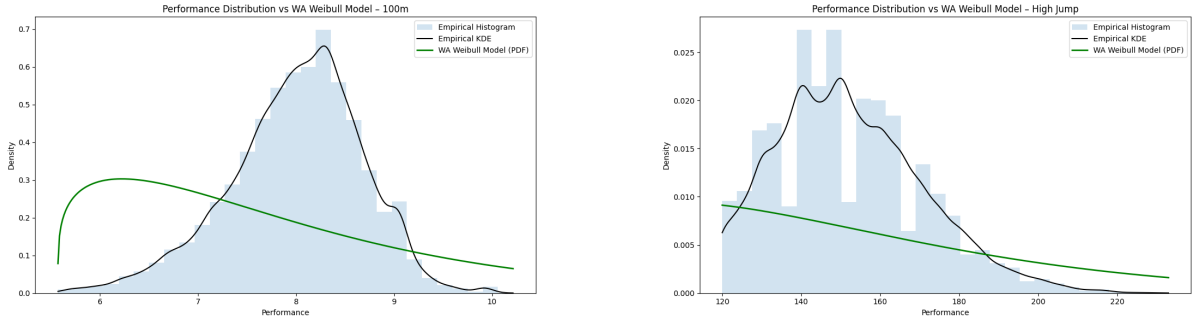


Figure 3.2: Comparison of event-wise KDEs vs fitted Weibull PDFs for 100m and HJ.

Here I present the fits of the model proposed by World Athletics for the 100m and High Jump events, compared to the collected data. The remaining figures can be found in Appendix 1. Clearly, I can see that the model suggested by World Athletics is not reflecting the empirical distribution of performance in any part of the distribution. Indeed it misses the mode and the tail does not align either.

Statistically it remain interesting to perform a Kolmogorov–Smirnov test to compare :

- The empirical cumulative distribution function based on the data,$\hat{F}_n(x)$.
- To a known cumulative distribution function, in this case,$F(x)$.

It measures the maximum absolute distance D between the two:

$$D = \sup_x |\hat{F}_n(x) - F(x)|.$$

- If the distance D, is small, $\hat{F}_n(x) \simeq F(x)$, the data $x_1, \ldots, x_n$ could possibly come from the f(x).

In this context, I have previously derived $Z \sim Exp(\hat{a}')$ from $X \sim Wei(a', b, c)$. I can hence perform the K-S test comparing the empirical CDF, $\hat{F}_n(Z)$ to $F(z) = 1 - e^{-\hat{a}'z}$. This tests the model with parameter b, c and $\hat{a}'$.

Also, even though I have estimated $\hat{a}'$, I don't know its true value. I can instead of comparing Z_i to an $Exp(\hat{a}')$, normalize Z;

$$\tilde{Z} = \frac{Z}{\bar{Z}} = \frac{(x_i - b)^c}{\frac{1}{n}\sum_{i=1}^{n}(x_i - b)^c}.$$

This transformation turns the data into exponential distribution with mean 1:

$$\tilde{Z} \sim Exponential(1) \quad .$$

By construction, it turns out that both methods produce the exact same results.

Event	a′	$Z \sim \text{Exp}(a')$	
		D-statistic	**p-value**
100m	0.3284	0.3286	0.0000
Long Jump	0.0003	0.3861	0.0000
Shot Put	0.1257	0.3875	0.0000
High Jump	0.0021	0.3696	0.0000
400m	0.3689	0.3497	0.0000
110m Hurdles	0.3464	0.2852	0.0000
Discus Throw	0.0351	0.2987	0.0000
Pole Vault	0.0008	0.2510	0.0000
Javelin Throw	0.0305	0.2840	0.0000
1500m	0.3914	0.3695	0.0000

Table 3.2: D-statistic and p-values for different events based on the Exponential distribution model.

I see that both statistically, using a global KS-test, and visually I have no evidence to show that the data gathered follow a shifted Weibull distribution parametrized such as World athletics did.

3.3 New distribution Model to approximate athletic performances

In this section I will present potential distribution models that would be good candidates to approximate the distributions of the data I gathered. I'll first describe the candidates and detail reasons why they would theoretically be good candidates. After this, I'll describe a methodology to estimates their parametrization, and asses their fit to the real data. To conclude this section, I'll set up a ranking system among those different parameterizations.

3.3.1 Model candidates

The potential distribution candidates must be continuous distribution defined on a semi-positive support $x \in [0; +\infty[$ or at least $x \in]0; +\infty[$.

Weibull

I decided to refine the Weibull model for each event's performance distribution $f_i(x)$:

$$f_i(x) = \begin{cases} \frac{c_i}{s_i}\left(\frac{x}{s_i}\right)^{c_i-1} \exp\left[-\left(\frac{x}{s_i}\right)^{c_i}\right], & x \geq 0 \\ 0, & x < 0 \end{cases}, \tag{3.7}$$

where:

- $s > 0$: The scale parameter.
- $c > 0$: The shape parameter controlling tail and skewness.

The corresponding cumulative and the complementary cumulative distribution functions are:

$$F(x_i) = \begin{cases} 1 - \exp\left[-\left(\frac{x}{s_i}\right)^{c_i}\right], & x \geq 0 \\ 0, & x < 0 \end{cases} \qquad \text{(CDF)}, \tag{3.8}$$

$$G_i(x) = 1 - F_i(x) = \exp\left[-\left(\frac{x}{s_i}\right)^{c_i}\right], \quad x \geq 0 \qquad \text{(CCDF)}. \tag{3.9}$$

The quantile of such a distribution can be computed by inverting the Cumulative Density Function :

$$x_{i,q} = s_i \left[-\ln(1-q)\right]^{1/c_i}, \quad \text{for } 0 < q < 1. \tag{3.10}$$

The parametrization has changed compared to the WA model. This change is made to simplify the coding later.

Also, I decided to remove the location parameter b in order to allow all levels of performances. I've defined previously in Section 3.2.3 the reasons why this distribution is theoretically ideal for performance distribution.

The mean and the variance of this a law are given by:

$$E[X_i] = s_i\,\Gamma\Big(1 + \tfrac{1}{c_i}\Big), \qquad \mathrm{Var}(X_i) = s_i^2\Big[\Gamma\Big(1 + \tfrac{2}{c_i}\Big) \;-\; \Gamma^2\Big(1 + \tfrac{1}{c_i}\Big)\Big].$$

Because the Weibull is a two parameters distribution law, this implies a one-to-one map $(s, c) \leftrightarrow (\mu, \sigma^2)$, any two Weibull laws with the same (s, c) must of course have the same mean and variance and conversely.

Log-Normal

Another model that could be a good candidate to fit our sport data is the Log-Normal distribution. Allen Downey in his book 'Probably Overthinking it: How to Use Data to Answer Questions, Avoid Statistical Traps, and Make Better Decisions' has reflected that running speeds could most likely be parametrized using Log-Normal distribution Law. His interpretation of 'Why a Log-Normal distribution' is well motivated, as it highlights the fact that running and, in this case, track and field event performances are influenced by multiple factors that interact in a multiplicative way. Among these factors, he includes physiology, environment, training, and others. When combined, these multiplicative effects can be modeled using a Log-Normal distribution. (Downey, 2023) [1]

Using a Log-Normal distribution, each event's performance distribution $f_i(x)$[2] is modeled:

$$f_i(x) = \begin{cases} \dfrac{1}{x s_i \sqrt{2\pi}} \exp\left(-\dfrac{1}{2s_i^2}\left[\log\left(\dfrac{x}{\theta_i}\right)\right]^2\right), & x > 0 \\ 0, & x \leq 0 \end{cases}, \tag{3.11}$$

[1] You can see some of his work in this GitHub `https://github.com/AllenDowney/ProbablyOverthinkingIt`

[2] This parametrization is set according to **scipy** package from python //`https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.lognorm.html`

where:

- $s > 0$: The shape parameter.
- $\theta > 0$: The scale parameter.

The corresponding cumulative and complementary cumulative distribution functions are:

$$F_i(x) = \begin{cases} \Phi\left(\frac{\log\left(\frac{x}{\theta_i}\right)}{s_i}\right), & x > 0 \\ 0, & x \leq 0 \end{cases} \quad \text{(CDF)}, \tag{3.12}$$

$$G_i(x) = 1 - F_i(x) = 1 - \Phi\left(\frac{\log\left(\frac{x}{\theta_i}\right)}{s_i}\right), \quad x > 0 \quad \text{(CCDF)}. \tag{3.13}$$

The quantile function is obtained by inverting the CDF:

$$x_{i,q} = \theta_i \cdot \exp\left(s_i \cdot \Phi^{-1}(q)\right), \quad \text{for } 0 < q < 1. \tag{3.14}$$

In addition to the arguments given above, other reasons would make the Log-Normal distribution a good candidate to model athletic performances.

- The support of such a distribution is set on $]0; +\infty[$, which reflects the possible values of athletic performances being strictly positive.
- On a linear scale the increment from poor performance to a medium performance (from 20 to 22 km/h) would be similar to an increment from good to elite performance (from 30 to 32 km/h). This behavior might misrepresent the effort put in by the elite athlete. He most likely trained much more than the median athlete to perform that increment. Using a logarithmic scale, these differences become more proportionally represented, allowing for a fairer and more intuitive comparison across performance levels. This is correctly representing the vertical comparison that is required to award points to performances.
- The high end of a Log-normal distribution can be much larger than one of a Gaussian distribution, meaning that extreme high values are much more probable than they would be under a Normal distribution.

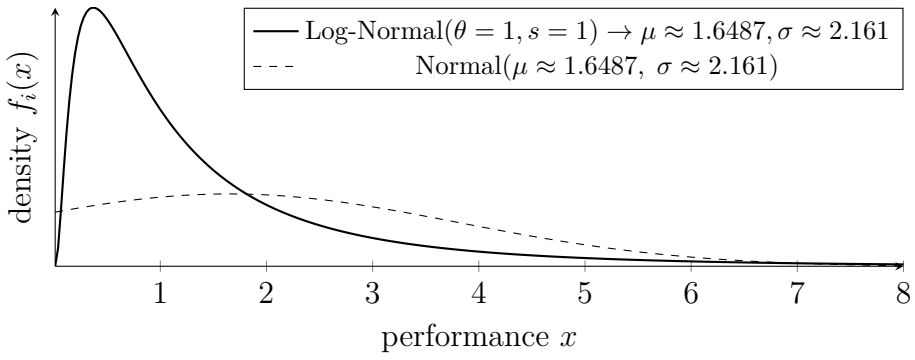


Figure 3.3: Log-Normal and Normal densities sharing the same mean and variance.

The mean and the variance of such a Log-Normal distribution law are given by:

$$E[X_i] = \theta_i \exp\Big(\tfrac{1}{2}\, s_i^2\Big), \qquad \mathrm{Var}(X_i) = \theta_i^2 \exp\Big(s_i^2\Big)\Big(\exp\Big(s_i^2\Big) - 1\Big).$$

Using the same intuition as the Weibull, two Log-Normal models that share the same parameters (θ_i, c_i), share the same mean and variance.

Gamma

Another distribution that I will examine for modeling performances is the Gamma distribution model. This type of distribution is frequently used in meteorology, particularly for modeling wind speed (Pobočíková et al., 2017). In this article, the Gamma distribution is compared to the Weibull and Log-Normal distributions, which I have mentioned earlier. Like the Weibull and Log-Normal distributions, the Gamma distribution is also derived from the exponential distribution.

There are several reasons why the Gamma distribution is particularly well-suited for modeling decathlon performance data:

- Positive support: As previously mentioned, decathlon marks are non-negative by nature. For example, one cannot run a race with a negative speed or jump a negative distance. The Gamma distribution naturally respects this constraint by having support on $]0, \infty[$.
- Flexible shape: The Gamma family includes both skewed and nearly symmetric distributions depending on the shape parameter k:
  - When $k < 1$, the distribution is highly right-skewed, capturing situations with a heavier tail than a Normal Law. The density is unbounded, a large part of the population would indeed get a performance of 0. For example, if everyone on the planet had one shot to perform a pole vault or throw a discus. It is likely that the vast majority of the participants would fail to get a valid result. In that scenario, a $k < 1$ would be reasonable
  - When $k \approx 1$, it behaves like an exponential distribution.
  - When $k > 1$, the distribution becomes increasingly bell-shaped and approaches normality.

  This flexibility allows the Gamma model to accommodate various performance distributions, from highly skewed to quasi-normal events in track and field. Even further, potentially to many sport using distances, weights, time or non negative metrics. One can think about swimming, short track cycling, weightlifting ...

The probability density function of an event, i in track and field $f_i(x)$ is modeled :

$$f_i(x) = \begin{cases} \dfrac{1}{\Gamma(k_i) \cdot \theta^{k_i}} (x)^{k_i - 1} e^{-\frac{x}{\theta_i}}, & x > 0 \\ 0, & x \leq 0 \end{cases}, \tag{3.15}$$

Once again, I use the parametrization from the scipy.stat package in python [3] with:

- $k > 0$: Shape parameter
- $\theta > 0$: Scale parameter

The introductory theoretical ground of the Gamma distribution has been found in "A brief introduction to the gamma distribution" (Harman, 2022).

The corresponding cumulative and the complementary cumulative distribution functions are:

$$F_i(x) = \begin{cases} \frac{1}{\Gamma(k_i)}\gamma\left(k_i, \frac{x}{\theta_i}\right), & x > 0 \\ 0, & x \leq 0 \end{cases} \qquad \text{(CDF)}, \tag{3.16}$$

$$G_i(x) = 1 - F_i(x) = 1 - \frac{1}{\Gamma(k_i)}\gamma\left(k_i, \frac{x}{\theta_i}\right), \quad x > 0 \qquad \text{(CCDF)}. \tag{3.17}$$

The $\gamma(k_i, \frac{x}{\theta_i})$ is the lower incomplete gamma function where, in a general context, $\gamma(k, x) = \int_0^x t^{k-1}e^{-t}dt$ is called the lower gamma function that compute the cumulative probability up to x.

The quantile function is obtained by inverting the CDF:

$$x_{q,i} = \theta_i \cdot \gamma^{-1}(q; k_i), \quad \text{for } 0 < q < 1, \tag{3.18}$$

where γ^{-1} is the inverse of the standard gamma function CDF with shape $= k$ and scale $= 1$.

Its first two moments are defined by :

$$E[X_i] = k_i\,\theta_i, \qquad \mathrm{Var}(X_i) = k_i\,\theta_i^2.$$

Once again as two Gamma laws share the same parameters, they will share the same mean and variance.

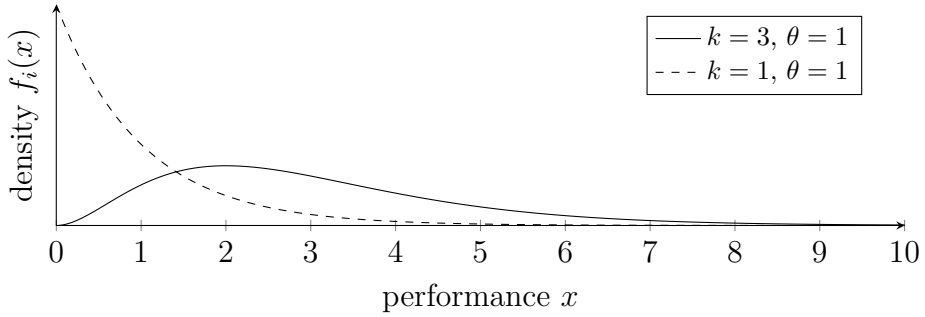


Figure 3.4: Gamma distributions with different shape parameters k and scale $\theta = 1$

[3] scipy.stats.gamma: `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.gamma.html`

Generalized Gamma

Another interesting distribution to examine is the Generalized Gamma distribution. The Generalized Gamma distribution's probability function is described such as every performance in event i can be modeled as $f_i(x)$:

$$f_i(x) = \begin{cases} \frac{p_i}{a_i^{d_i}\Gamma\left(\frac{d_i}{p_i}\right)} x^{d_i-1} \exp\left[-\left(\frac{x}{a_i}\right)^{p_i}\right], & x > 0 \\ 0, & x \leq 0 \end{cases}, \tag{3.19}$$

with parameters:

- $a > 0$: Scale parameter
- $d > 0$: Shape parameter. Higher values of d result in a sharper peak at the origin
- $p > 0$: Shape parameter. Higher values of p lead to a steeper decay for the upper tail of the distribution.

This distribution makes sense in this framework due to its relation with the Weibull and the Gamma distribution. Indeed, depending on the parametrization, the Generalized Gamma distribution can become a simple Gamma or a Weibull distribution.

- $p = 1$: reduces to Gamma distribution.
- $d = p$: reduces to Weibull distribution.
- $d = p = 1$: reduces to Exponential distribution.

This gives me the opportunity to fine-tune the distribution closer to the reality.

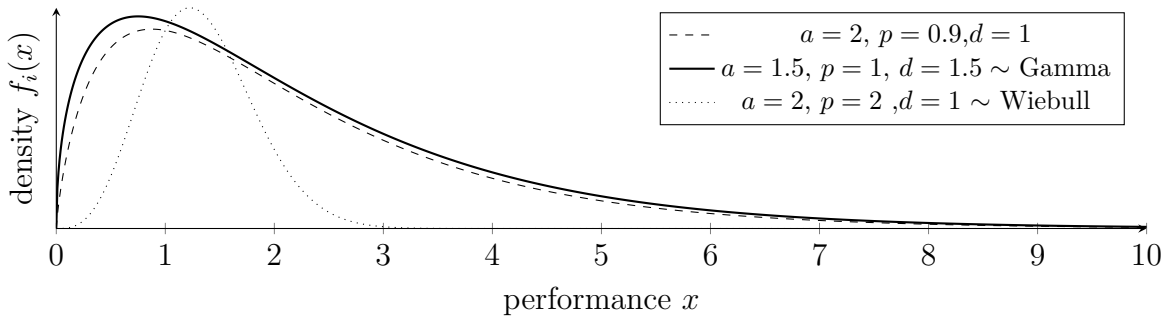


Figure 3.5: Generalized Gamma distributions

The corresponding cumulative density function and survival functions are :

$$F_i(x) = \begin{cases} \frac{\gamma\left(\frac{d_i}{p_i}, \left(\frac{x}{a_i}\right)^{p_i}\right)}{\Gamma(d_i/p_i)}, & x > 0 \\ 0, & x \leq 0 \end{cases} \quad \text{(CDF)}, \tag{3.20}$$

$$G_i(x) = 1 - F_i(x) = 1 - \frac{\gamma\left(\frac{d_i}{p_i}, \left(\frac{x}{a_i}\right)^{p_i}\right)}{\Gamma(d_i/p_i)}, \quad x > 0 \quad \text{(CCDF)}. \tag{3.21}$$

The quantile function is obtained by inverting the CDF:

$$x_{i,q} = a_i \left[\gamma^{-1} \left(\frac{d_i}{p_i}, q \cdot \Gamma(\frac{d}{p}) \right) \right]^{1/p}, \quad \text{for } 0 < q < 1. \tag{3.22}$$

The mean and the variance are defined :

$$E[X_i] = a_i \frac{\Gamma\left(\frac{d_i+1}{p_i}\right)}{\Gamma\left(\frac{d_i}{p_i}\right)}, \qquad \text{Var}(X_i) = a_i^2 \left[\frac{\Gamma\left(\frac{d_i+2}{p_i}\right)}{\Gamma\left(\frac{d_i}{p_i}\right)} - \left(\frac{\Gamma\left(\frac{d_i+1}{p_i}\right)}{\Gamma\left(\frac{d_i}{p_i}\right)} \right)^2 \right].$$

As highlighted by the graph below, two Generalized Gamma distribution can share the same moments even though they are parametrized differently.

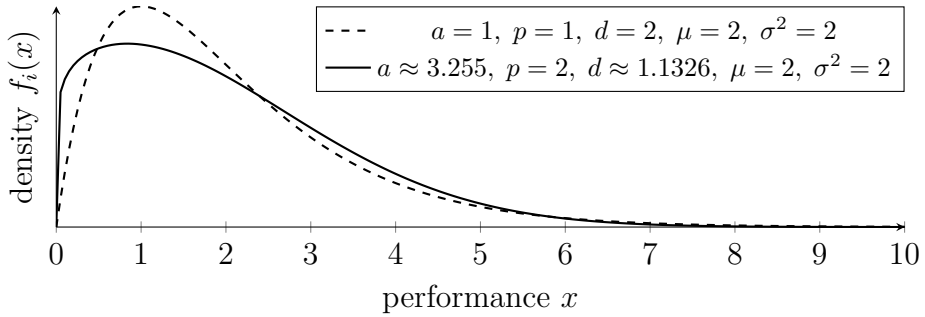


Figure 3.6: Two Generalized Gamma densities with different (a, p, d) but identical $E[X] = 2$ and $\text{Var}(X) = 2$.

Burr

The Burr distribution Type XII can be a very effective law when modeling sport performances, particularly for speed and distance events. The Burr Type XII distribution is also able to model heavy right tail. Indeed it can capture the extreme high elite performance. Nevertheless, the Burr distribution can also fit the bulk really well. This twofold property is very useful when modeling performance across the whole population of athletes. Also being parametrized by two shape parameters, the distribution can capture more complex distributions, compared to Weibull or Gamma which only have one shape parameter. It can be shown that using specific parametrization, the Burr Type XII distribution can be reduced to a simpler related distribution (Tadikamalla, 1980) among which the Weibull distribution that I examined earlier when $k \to \infty$ or the Lomax distribution when $c = 1$.

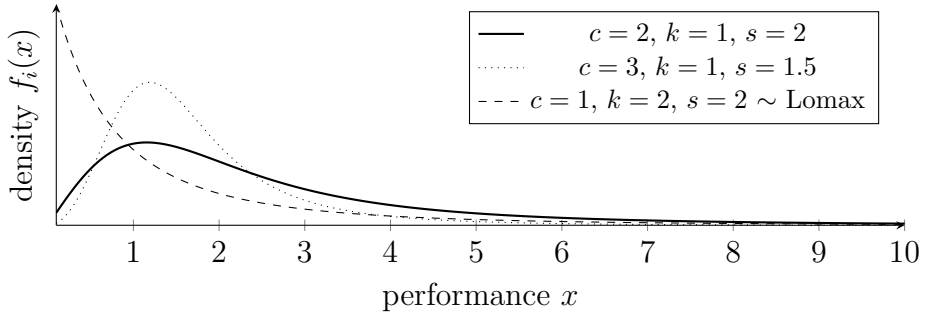


Figure 3.7: Burr distributions with varying shape parameters

The Burr Distribution's probability function is described such that the density of every performance in event i can be modeled by $f_i(x)$:

$$f_i(x) = \begin{cases} \dfrac{c_i k_i \left(\frac{x}{s_i}\right)^{c_i - 1}}{s_i \left(1 + (\frac{x}{s_i})^{c_i}\right)^{k_i+1}}, & x \geq 0 \\ 0, & x < 0 \end{cases}, \tag{3.23}$$

with parameters:

- $s > 0$: Scale parameter
- $c > 0$: Shape parameter that controls the distribution's peak and the growth rate near the origin $x = 0$. While c increases, the curve becomes sharper at the peak and slower in tail decay, meaning it takes longer for the probability to drop off.
- $k > 0$: Shape parameter that influences the decay rate of the distribution and the tail behavior. If k increases, the curve becomes taller and narrower around the peak, with the tail decaying faster.

The corresponding CDF and CCDF are respectively :

$$F_i(x) = \begin{cases} 1 - \left(1 + \left(\frac{x}{s_i}\right)^{c_i}\right)^{-k_i}, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad \text{(CDF)}, \tag{3.24}$$

$$G_i(x) = 1 - F_i(x) = \left(1 + \left(\frac{x}{s_i}\right)^{c_i}\right)^{-k_i}, \quad x \geq 0 \quad \text{(CCDF)}. \tag{3.25}$$

The quantile function is obtained by inverting the CDF:

$$x_{i,q} = s_i \left[\left(\frac{1}{1-q}\right)^{1/k_i} - 1\right]^{1/c_i}, \quad \text{for } 0 < q < 1. \tag{3.26}$$

The mean and variance of the Burr type XII are given by:

$$E[X_i] = s_i \, \frac{\Gamma\left(1 + \frac{1}{c_i}\right) \Gamma\left(k_i - \frac{1}{c_i}\right)}{\Gamma(k_i)},$$

$$\mathrm{Var}(X_i) = s_i^2 \left[\frac{\Gamma\left(1 + \frac{2}{c_i}\right) \Gamma\left(k_i - \frac{2}{c_i}\right)}{\Gamma(k_i)} - \left(\frac{\Gamma\left(1 + \frac{1}{c_i}\right) \Gamma\left(k_i - \frac{1}{c_i}\right)}{\Gamma(k_i)}\right)^2\right].$$

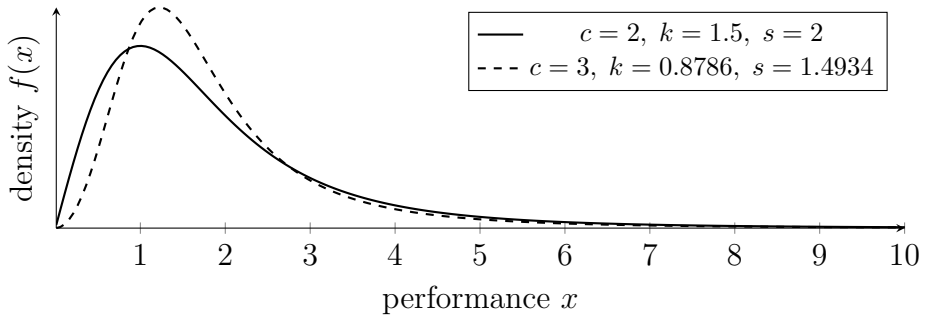


Figure 3.8: Two Burr XII densities sharing $E[X] = 2$ and $\mathrm{Var}(X) = 4$ but with different (c, k, s).

The graph shows that two models may share the same moments even though they do not share the parameters (c, k, s).

Burr Inverse

Finally, the Inverse Burr distribution is defined on the domain of interest R_0^+, and powerful around the tails, see e.g. (Hafner et al., 2025). The Inverse Burr distribution's probability function is described such that every performance in event i can be modeled using $f_i(x)$:

$$f_i(x) = \begin{cases} \dfrac{\alpha_i \tau_i \left(\left(\frac{x}{\theta_i}\right)^{\tau_i}\right)^{\alpha_i}}{x\left(1 + (\frac{x}{\theta_i})^{\tau_i}\right)^{\alpha_i+1}}, & x > 0 \\ 0, & x \leq 0 \end{cases}, \tag{3.27}$$

with parameters:

- $\theta > 0$: Scale parameter, that controls the stretch.
- $\tau > 0$: Shape parameter, a larger τ leads to a more bell-shaped curve and can reduce the skewness of the distribution.
- $\alpha > 0$: Shape parameter. When large, the distribution decays faster, and the tails become lighter.

The corresponding CDF and CCDF are respectively :

$$F_i(x) = \begin{cases} \left(\frac{(x/\theta_i)^{\tau_i}}{1+(x/\theta_i)^{\tau_i}}\right)^{\alpha_i}, & x > 0 \\ 0, & x \leq 0 \end{cases} \quad \text{(CDF)}, \tag{3.28}$$

$$G_i(x) = 1 - F_i(x) = 1 - \left(\frac{(x/\theta_i)^{\tau_i}}{1 + (x/\theta_i)^{\tau_i}}\right)^{\alpha_i} \quad x > 0 \quad \text{(CCDF)}. \tag{3.29}$$

The quantile function is obtained by inverting the CDF:

$$x_{i,q} = \theta_i \left(\frac{q^{1/\alpha_i}}{1 - q^{1/\alpha_i}}\right)^{1/\tau_i}, \quad \text{for } 0 < q < 1. \tag{3.30}$$

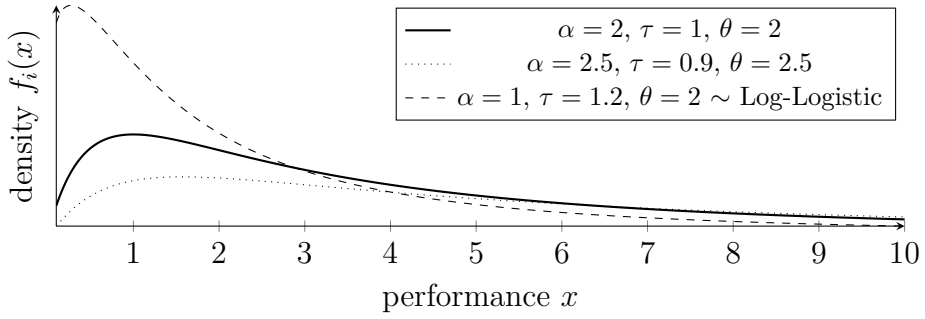


Figure 3.9: Inverse Burr distributions with varying shape parameters

The r-th moment of a Inverse Burr exists for $r < \tau_i$ and is

$$E[X_i^r] = \theta_i^r \, \frac{\Gamma\left(\alpha_i + \frac{r}{\tau_i}\right) \Gamma\left(1 - \frac{r}{\tau_i}\right)}{\Gamma(\alpha_i)}.$$

In particular, provided $\tau_i > 1$, the mean is:

$$E[X_i] = \theta_i \, \frac{\Gamma\left(\alpha_i + \frac{1}{\tau_i}\right) \Gamma\left(1 - \frac{1}{\tau_i}\right)}{\Gamma(\alpha_i)},$$

and, provided $\tau_i > 2$, the second moment is:

$$E[X_i^2] = \theta_i^2 \, \frac{\Gamma\left(\alpha_i + \frac{2}{\tau_i}\right) \Gamma\left(1 - \frac{2}{\tau_i}\right)}{\Gamma(\alpha_i)}.$$

Hence the variance (for $\tau_i > 2$) is

$$\mathrm{Var}(X_i) = \theta_i^2 \left[\frac{\Gamma\left(\alpha_i + \frac{2}{\tau_i}\right) \Gamma\left(1 - \frac{2}{\tau_i}\right)}{\Gamma(\alpha_i)} \; - \; \left(\frac{\Gamma\left(\alpha_i + \frac{1}{\tau_i}\right) \Gamma\left(1 - \frac{1}{\tau_i}\right)}{\Gamma(\alpha_i)} \right)^2 \right].$$

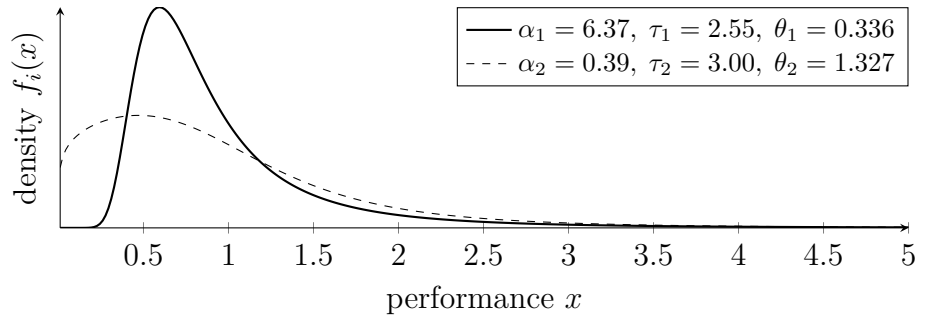


Figure 3.10: Two Inverse-Burr densities with $E[X_i] \approx 1$ and $\mathrm{Var}(X_i) \approx 1$ with different parametrizations $(\alpha_i, \tau_i, \theta_i)$.

Final note on models' parametrization

In order to allow the maximum people to take part in all the disciplines, I decided to set the location parameters ℓ to zero. Nevertheless, setting $\ell = 0$ means that it remains possible to set a score of 0.

3.3.2 Parameters estimation and sampling

To estimate our parameters and fit our models to the track and field data I gathered, I must define one or multiple methods.

In order to have multiple points of comparison, I defined three different estimation methods.

- Maximum Likelihood Estimation (MLE),
- Quantile Matching Estimation (QME),
- Bayesian inference using Markov chain Monte Carlo (MCMC).

Each method produces both point estimates and samples from the estimated distribution of the parameters. These outputs are later used to compute confidence intervals for model predictions, compare distributional fits, and assess fairness and calibration across events.

Frequentist approach

Maximum likelihood estimation

The maximum likelihood estimation (MLE) is a widely-used approach where the parameters are obtained by maximizing the likelihood function. The goal is to find the set of parameters that makes the observed data the most probable under the assumed distribution.

Mathematically, let $x_1, x_2, ...x_n$ be the sample of observed performances, assumed *iid* from a distribution $f(x; \boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ the vector of parameters. The likelihood of observing the data give $\boldsymbol{\theta}$ is:

$$\mathcal{L}(\boldsymbol{\theta}; x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x; \boldsymbol{\theta}).$$

For convenience, it is common to derive the log-likelihood given some parameters. The problem then becomes:

$$\ell(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta}; x_1, x_2, ..., x_n) = \sum_{i=1}^{n} \log f(x_i; \boldsymbol{\theta}). \tag{3.31}$$

The set of estimates are the arguments that maximize the log-likelihood :

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell(\theta). \tag{3.32}$$

Ultimately, deriving the log-likelihood with respect to each and every parameters and equalizing these expressions to zero will help me find the optimal values for those parameters.

The (3.31) becomes for our six candidates :

- Weibull: Using (3.7), the log-likelihood is defined :

$$\ell(c, s) = n \log c - n \log s + (c-1) \sum_{i=1}^{n} \log(x_i) - \sum_{i=1}^{n} \left(\frac{x_i}{s}\right)^c \quad ,$$

with shape $c > 0$, scale $s > 0$:

$$\begin{aligned}
\frac{\partial \ell}{\partial c} &= \frac{n}{c} - n \log s + \sum_{i=1}^{n} \log x_i - \sum_{i=1}^{n} \left(\frac{x_i}{s}\right)^c \log\left(\frac{x_i}{s}\right), \\
\frac{\partial \ell}{\partial s} &= -\frac{nc}{s} + \frac{c}{s} \sum_{i=1}^{n} \left(\frac{x_i}{s}\right)^c.
\end{aligned}$$

There are no analytical solutions for the parameters because they are non-linear in both parameters.

- Log-Normal: The log-likelihood, using (3.11), is defined :

$$\ell(\theta, s) = -n \log(s\sqrt{2\pi}) - \sum_{i=1}^{n} \log(x_i) - \frac{1}{2s^2} \sum_{i=1}^{n} \left[\log\left(\frac{x_i}{\theta}\right)\right]^2 \quad ,$$

with location $\theta \in R$, scale > 0

$$\begin{aligned}
\frac{\partial \ell}{\partial \theta} &= \frac{1}{s^2} \sum_{i=1}^{n} (\log x_i - \theta), \\
\frac{\partial \ell}{\partial s} &= -\frac{n}{s} + \frac{1}{s^3} \sum_{i=1}^{n} (\log x_i - \theta)^2.
\end{aligned}$$

The analytical solution for the parameters is :

$$\begin{aligned}
\hat{\theta} &= \frac{1}{n} \sum_{i=1}^{n} \log(x_i) \\
\hat{s} &= \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\log(x_i) - \hat{\theta}\right)^2}
\end{aligned}$$

- **Gamma**: With (3.15), the log-likelihood is defined :

$$\ell(k, \theta) = -n \log \Gamma(k) - nk \log \theta + (k-1) \sum_{i=1}^{n} \log(x_i) - \frac{1}{\theta} \sum_{i=1}^{n} (x_i) \quad ,$$

with shape $k > 0$, scale $\theta > 0$

$$\begin{aligned}
\frac{\partial \ell}{\partial k} &= -n\psi(k) - n \log \theta + \sum_{i=1}^{n} \log x_i, \\
\frac{\partial \ell}{\partial \theta} &= -\frac{nk}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} x_i.
\end{aligned}$$

Analytically, I find that $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i$. However, because of the digamma function $\psi(k)$, $\hat{k}$ needs to be numerically solved.

- Generalized Gamma Using (3.19) the log-likelihood is defined :

$$\ell(d,p,a) = n\log p - nd\log a - n\log\Gamma\left(\frac{d}{p}\right) + (d-1)\sum_{i=1}^{n}\log x_i - \sum_{i=1}^{n}\left(\frac{x_i}{a}\right)^p \quad ,$$

with scale $\theta > 0$, shape $a > 0$, power $p > 0$

$$\begin{aligned}
\frac{\partial \ell}{\partial d} &= -n\log a - n\frac{1}{p}\psi\left(\frac{d}{p}\right) + \sum_{i=1}^{n}\log x_i, \\
\frac{\partial \ell}{\partial p} &= \frac{n}{p} + n\frac{d}{p^2}\psi\left(\frac{d}{p}\right) - \sum_{i=1}^{n}\left(\frac{x_i}{a}\right)^p \log\left(\frac{x_i}{a}\right), \\
\frac{\partial \ell}{\partial a} &= -\frac{nd}{a} + \frac{p}{a}\sum_{i=1}^{n}\left(\frac{x_i}{a}\right)^p.
\end{aligned}$$

There are no closed form for the maximum likelihood estimation for d and p as they involve a digamma function $\psi(\frac{d}{p})$ and a requires p in its MLE :$\hat{a} = \frac{1}{n}\sum_{i=1}^{n} x_i^p$.

- **Burr**:

The log-likelihood, with (3.23), is defined as:

$$\ell(c,k,s) = n\log c + n\log k - n\log s + (c-1)\sum_{i=1}^{n}\log\left(\frac{x_i}{s}\right) - (k+1)\sum_{i=1}^{n}\log\left(1+\left(\frac{x_i}{s}\right)^c\right).$$

Let $z_i = \frac{x_i}{s}$. The gradients are:

$$\begin{aligned}
\frac{\partial \ell}{\partial c} &= \frac{n}{c} + \sum_{i=1}^{n}\log z_i - (k+1)\sum_{i=1}^{n}\frac{z_i^c \log z_i}{1+z_i^c}, \\
\frac{\partial \ell}{\partial k} &= \frac{n}{k} - \sum_{i=1}^{n}\log\left(1+z_i^c\right), \\
\frac{\partial \ell}{\partial s} &= -\frac{cn}{s} + \frac{c(k+1)}{s}\sum_{i=1}^{n}\frac{z_i^c}{1+z_i^c}.
\end{aligned}$$

Again, here based on the derivative, I see that I would need a numerical solution as analytically the MLE have no closed form.

- **Inverse Burr**:

Combing (3.31) and (3.27), the log-likelihood is defined as:

$$\ell(\alpha,\tau,\theta) = n\log\alpha + n\log\tau + \alpha\tau\sum_{i=1}^{n}\log\left(\frac{x_i}{\theta}\right) - \sum_{i=1}^{n}\log x_i - (\alpha+1)\sum_{i=1}^{n}\log\left(1+\left(\frac{x_i}{\theta}\right)^\tau\right.$$

Let $z_i = \left(\frac{x_i}{\theta}\right)^\tau$. The gradients are:

$$\frac{\partial \ell}{\partial \alpha} = \frac{n}{\alpha} + \tau \sum_{i=1}^{n} \log\left(\frac{x_i}{\theta}\right) - \sum_{i=1}^{n} \log(1 + z_i),$$
$$\frac{\partial \ell}{\partial \tau} = \frac{n}{\tau} + \alpha \sum_{i=1}^{n} \log\left(\frac{x_i}{\theta}\right) - (\alpha + 1) \sum_{i=1}^{n} \frac{z_i \log\left(\frac{x_i}{\theta}\right)}{1 + z_i},$$
$$\frac{\partial \ell}{\partial \theta} = -\frac{\alpha\tau}{\theta} \sum_{i=1}^{n} 1 + \frac{\tau(\alpha + 1)}{\theta} \sum_{i=1}^{n} \frac{z_i}{1 + z_i}.$$

Analytical closed-form solutions for these parameters are not possible because the log-likelihood involves complex summations and non-linear relationships between the parameters and the data

The previous model's parameters do not always provide an analytical closed form using the derivation according to (Danjuma et al., 2020), Only the Log-Normal distribution law has. So values of the parameters in my MLE estimation will be found through numerical optimization using scipy.optimize[4].

For the ease of optimization, I used the module `scipty.optimize`. In the framework defined above, this optimizer minimize the function given in argument. Therefore, the argument function I have to provide is the negative log-likelihood. Provided some initial guess and bounds to look into, the optimizer return the best evaluation of the parameters, $\hat{\boldsymbol{\theta}}$.

Quantile matching estimation

Another method to estimate parameters that uses the theoretical and empirical quantiles of the pdf $f(x)$ is the quantile matching estimation (QME). It is a generalization of the least square error estimation (Sgouropoulos et al., 2015). The general idea is to find the set of optimal parameters such that the theoretical quantiles defined by a set of estimated parameters matches the best the empirical quantiles. This method is particularly useful because it allows me to emphasize specific regions of the distribution, giving greater importance to areas where a more accurate fit is desired. Indeed, it can be more robust when working with heavy-tailed data.

Mathematically, let $x_1, x_2, ...x_n$ be the sample of observed performances and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ be the vector of distribution parameters.

I define a vector of m target quantiles levels:

$$\mathbf{q} = [q_1, q_2, ..., q_m] \subset (0, 1).$$

For each quantile level q_i, I compute:

- The empirical quantile $x^{emp}(q_i)$ from the sample

[4] `scipy.optimize` package : `https://docs.scipy.org/doc/scipy/reference/optimize.html`

- The theoretical quantile from the model $x^{model}(q_i; \boldsymbol{\theta})$ from the distribution

The objective of the QME is to minimize the sum of squared differences between the empirical and theoretical quantiles:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{m} (x^{emp}(q_i) - x^{model}(q_i; \boldsymbol{\theta}))^2. \tag{3.33}$$

Using the percent point function (PPF), I can easily estimate the the values of the theoretical quantiles. Luckily, `scipy` modules in python provide build-in functions that help me find the specified quantile of a distribution given the parameters.

Application of QME for Gamma

1. Choosing the fixed quantiles.
2. Computing the empirical quantiles from the data.
3. Defining the theoretical quantile function from the parametric distribution using scipy.stats.gamma.ppf(q_i, shape, scale) or the self-defined PPF.
4. Minimizing the squared error between empirical and theoretical quantiles (3.33) using `scipy.optimize.minimize`, which uses the same intuition has in the MLE.

Compared to the MLE estimation, the QME is interesting because it is more robust to outliers as it focuses on specific quantiles, it doesn't necessarily require a PDF. However, it is rare to find a closed form for the parameters.

Using `scipy` in python, I can use this technique with Gamma, Log-Normal, Weibull, Burr and Inverse Burr. This is unfortunately not the case for this Generalized Gamma. Indeed, I would need to inverse the CDF of the Generalized Gamma CDF (3.20), being (3.22). Again, this last expression has no closed form, hence the QME is highly unstable as I would need to optimize the function inside an optimizing function.

In this context, knowing the general function linking CCDF(x) and and the score (2.5), I can derive quantile of interest. If I want the curve to fit particularly well the data at specific points. Note that from (2.5), λ is just a proportionality constant, I can suppress it.

If I want to have a fit at specific point levels, I can just find the values of $F(x) \in [0, 1]$ to some predefined scores and the quantiles will be defined. I decided to focus on the scores $p(x) = \{1, 1.5, 2 \dots 6\}$ which returned me the following percentiles : [0.632, 0.777, 0.865, 0.918, 0.95, 0.97, 0.982, 0.989, 0.993, 0.996, 0.998].

K-Fold estimation

Until now, I've defined a general frequentist method to estimate the parameters of the distributions. However, the data can be noisy, skewed and can vary across events. Also, to model the performances of each and every events, I rely on a single data set for each events. To avoid overfitting and misleading generalization, I decided to implement a method that strikes a balance between model complexity and generalization. I opted to use K-Fold

estimation.

The general principle of K-Fold estimation, is the following one:
Let $x_1, x_2, \ldots, x_n$ be the sample of observed performances. In K-fold estimation, the data is randomly partitioned into K disjoint subsets of approximately equal size. The procedure is as follows:

1. One fold is held out as a unseen set.
2. The remaining $K - 1$ folds are used as the training set.
3. The model is fitted on the training data using the selected estimation method (MLE or QME).
4. The resulting parameter estimates are stored.

This process is repeated K times so that each subset is used exactly once as a unseen set. At the end of the procedure, I obtain K parameter sets, which can be aggregated to give a mean estimates.

I choose $K = 10$ folds, which provides a high-resolution view of parameter variability while maintaining enough data per fold for stable fitting. This setup enables a more realistic evaluation of the models' generalization.

Sampling

Another important part of this work is to quantify the uncertainty of the estimates. Indeed, point estimates of the parameters do not give the full story about the distribution of the data. To assess that uncertainty, I decided to compute confidence intervals, which capture the range within which the true parameters of the model are likely to fall with a high probability (95%).

In the frequentist framework, parameters $\boldsymbol{\theta}$ are considered fixed but unknown. The variability in estimation lies in the randomness of the data sampling process. When applying bootstrap resampling, I simulate this variability by generating multiple pseudo-datasets through sampling with replacement from the original data.

This process allows me to approximate the sampling distribution of an estimator $\hat{\boldsymbol{\theta}}$ without relying on strong parametric assumptions. In particular, I implemented the bootstrap which performs the following:

1. I draw $B = 500$ bootstrap samples from the original dataset (with replacements).
2. For each sample, I estimate the model parameters using both maximum likelihood estimation (MLE) and quantile matching estimation (QME) by K-Fold CV:

$$\hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}, \ldots, \hat{\boldsymbol{\theta}}^{(B)}.$$

3. I construct a $(1-\alpha)$ confidence interval for each parameter using empirical percentiles of the bootstrap distribution:

$$[\theta_{\text{lower}}, \theta_{\text{upper}}] = \left[\text{percentile}_{\alpha/2}, \text{percentile}_{1-\alpha/2}\right].$$

4. Finally I compute the sample mean, median and the standard deviation of the bootstrap estimates:

$$E[\hat{\theta}] \approx \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}, \tag{3.34}$$

$$\text{Std}(\hat{\theta}) \approx \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^{(b)} - E[\hat{\theta}]\right)^2}, \tag{3.35}$$

$$\text{Med}[\hat{\theta}] \approx \text{median}\left(\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}\right). \tag{3.36}$$

These statistics provide point estimates and standard errors for each parameter. Both the mean and the median are relevant to extract. Indeed the mean, is analytically defined while the median is much safer when the distribution is skewed and I want to reduce the sensitivity of outliers, due to bad samplings for example.

The interpretation of the frequentist confidence interval is that if I were to repeat this entire data-generating process many times, approximately $(1-\alpha)\%$ of the resulting intervals would contain the true set of parameters $\boldsymbol{\theta}$.

Full frequentist estimate framework

Under this section, I describe step by step the full framework used to estimate my parameters for each model:

1. Bootstrap Sampling:
   (a) I generate a total of $n_{\text{bootstrap}} = 500$ bootstrap samples by sampling with replacement from the full dataset.
   (b) For each bootstrap sample, I perform the following:
      i. 10 folds estimation: The sample is split into 10 folds. For each fold, the model is fitted using the selected method (MLE or QME), and the estimated parameters are recorded.
      ii. The 10 folds estimates are aggregated using the median to produce an estimate for that bootstrap sample. This allows me to produce estimates that are not based on the full bootstrap, and mitigate the risk of overfitting.
   (c) The robust parameter estimates from all bootstrap samples are collected.
2. Summary Statistics:
   - The final distribution of parameter estimates is summarized using the sample mean, median, standard error, and the 95% confidence interval.
   - These statistics are stored per event, model, and method.

Heavy-tail distribution behavior and tail index

Using extreme value theory (EVT), I can derive a method to estimate parameters that specifically fit the tail of the distribution. Indeed, I describe how quickly the probability of extreme values decreases using the tail index α and the survival function $G(x)$ for a non-negative random variable X (McNeil et al., 2015):

$$G(x) = 1 - F(x) \approx c^{\alpha}\, x^{-\alpha} \quad \text{for } x \geq c,$$

or more simply,

$$G(x) \approx x^{-\alpha}. \tag{3.37}$$

In a parametric context, I can find each law's tail index from its survival function. For example, a Burr Type XII has:

$$G(x) = [1 + (x/s)^{c}]^{-k} \sim (x/s)^{-c\,k} \quad (x \to \infty),$$

so by (3.37) its tail index is

$$\alpha = c\,k \quad \Longleftrightarrow \quad k = \frac{\alpha}{c}.$$

To estimate α, I use the Hill estimator (Hill, 1975) and the asymptotic results of Cheng and Peng (Cheng and Peng, 2001). Given order-statistics

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)},$$

I choose m, the number of top observations (e.g. the 95% quantile). The Hill estimate is

$$\hat{\alpha}_H = \left[\frac{1}{m} \sum_{i=1}^{m} \ln X_{(n-i+1)} - \ln X_{(n-m)} \right]^{-1}. \tag{3.38}$$

Under $m \to \infty$, $m/n \to 0$,

$$\sqrt{m}\big(\hat{\alpha}_H - \alpha\big) \xrightarrow{d} N(0, \alpha^2),$$

so an approximate $(1-\beta)$ confidence interval is

$$\hat{\alpha}_H \;\pm\; z_{1-\beta/2}\, \frac{\hat{\alpha}_H}{\sqrt{m}}.$$

Event	$\hat{\boldsymbol{\alpha}}_{\boldsymbol{H}}$	**CI lower**	**CI upper**
100	33.3462	33.3100	33.3823
LJ	15.0477	15.0299	15.0656
SP	8.4090	8.3924	8.4255
HJ	16.2761	16.2544	16.2977
400	33.9735	33.9270	34.0201
110h	23.7069	23.6261	23.7878
DT	7.5509	7.5350	7.5669
PV	9.8036	9.7830	9.8241
JT	7.2095	7.1963	7.2227
1500	26.9179	26.8821	26.9538

Table 3.3: Hill tail-index estimates $\hat{\alpha}_H$ with their 95% confidence intervals for each decathlon event.

Given the Table 3.3 I see that the Hill estimate for the tail index is large for all events $\alpha_H > 2$. This places every discipline in the light-tailed regime, ensuring finite variance and a fast decay probability of extreme quantile (Feruzbek, 2022). In particular, the 100m ($\alpha_H \approx 33$) behaves almost exponentially as $x \to \infty$ when $n \to \infty$ while the jumps and throws ($\alpha_H \approx 7$–16) still decay fast but can be viewed as power-law behavior with modest to high dispersion od at the upper end.

With $\hat{\alpha}_H$ from Equation (3.38) in hand, I plug it into the Burr log-likelihood,

$$\ell(c, s) = \sum_{i=1}^{n} \ln f\Big(X_i;\, c,\, k = \hat{\alpha}_H/c,\, s\Big),$$

and optimize over (c, s) only, reducing a three-dimensional search (c, k, s) to two dimensions.

Distribution	**Survival** $G(x)$	**Asymptotic tail form**	$\boldsymbol{\alpha}$
Weibull	$e^{-(x/s)^c}$	$\ln G(x) \approx -(x/s)^c$	/
Log-Normal	$1 - \Phi\Big(\frac{\ln(x/\theta)}{s}\Big)$	$\ln G(x) \approx -\frac{(\ln(x/\theta))^2}{2s^2}$	/
Gamma	$1 - \frac{1}{\Gamma(k)}\,\gamma\Big(k, \frac{x}{\theta}\Big)$	$\ln G(x) \approx -x/\theta + O(\ln x)$	/
Gen. Gamma	$1 - \frac{\gamma(d/p, (x/a)^p)}{\Gamma(d/p)}$	$\ln G(x) \approx -(x/a)^p$	/
Burr XII	$(1 + (x/s)^c)^{-k}$	$G(x) \approx (x/s)^{-c\,k}$	$\alpha = c\,k$
Inverse Burr	$1 - \Big(\frac{(x/\theta)^\tau}{1+(x/\theta)^\tau}\Big)^\alpha$	$G(x) \approx \alpha\,\theta^\tau\,x^{-\tau}$	$\alpha = \tau$

Table 3.4: Tail behavior for candidate distributions.

This approach is theoretically efficient. However, in my framework and data:

- I estimate α via Hill's method on only the largest m observations, so small m yields high variance in $\hat{\alpha}_H$.
- Bootstrapping small subsamples drastically alters each fold's extreme order-statistics, causing $\hat{\alpha}_H$ (and thus the fitted parameters) to collapse or explode.
- By enforcing $k = \hat{\alpha}_H/c$, I collapse (c, k) to a one-dimensional curve. If the curve is not uni-modal, estimates become non-identifiable and unstable.

Probabilistic approach

Bayesian estimation method

In addition to the frequentist methods such as maximum likelihood estimation (MLE) and quantile matching estimation (QME), I will also use Bayesian inference to provide a new set of parameters. The Bayesian framework is powerful and flexible because unlike frequentist approaches, I treat the parameters as random variables with a prior distribution that is updated using the observed data.

More formally I aim to compute the posterior distribution of the sets of parameters $\boldsymbol{\theta}$ given the observed data $\mathbf{x} = \{x_1, x_2, ..., x_n\}$.

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(\mathbf{x})}, \tag{3.39}$$

where:

- $\mathbf{x}$: The observed data.
- $\boldsymbol{\theta}$: The sets of parameters of the distribution.
- $p(\boldsymbol{\theta})$: The prior distribution of the parameters, can be not extremely informative.
- $p(\mathbf{x}|\boldsymbol{\theta})$: The likelihood, representing how likely the observed data is under the model with parameters θ.
- $p(\mathbf{x})$: The marginal likelihood, which ensures that the posterior is a valid probability distribution. The marginal likelihood is computed as $p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}) d\boldsymbol{\theta}$.

To overcome the complexity of the integral $p(\mathbf{x})$ required for computing the posterior $p(\boldsymbol{\theta}|\mathbf{x})$, I use Monte Carlo Markov Chain :

$$\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(T)} \sim p(\boldsymbol{\theta}|\mathbf{x}), \tag{3.40}$$

with the following properties:

- Each sample $\boldsymbol{\theta}^{(t)}$ depends only on the previous one (Markov property):

$$p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(t-1)}) = p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)}),$$

- The chain is designed so that its stationary distribution is the posterior:

$$\boldsymbol{\theta}^{(t)} \sim p(\boldsymbol{\theta} \mid \mathbf{x}) \quad \text{for large } t,$$

In these experiments, I use the `Stan` probabilistic programming language via the `cmdstanpy` interface [5]. Stan through CmdStanPy allows users to define complex probabilistic models in Python and performs efficient sampling for posterior distributions. In other words, I implement, using this tool, a gradient-based MCMC algorithm. This algorithm efficiently explores complex, high-dimensional posterior spaces and offers fast convergence.

For each distribution (Gamma, Weibull, Log-Normal, Generalized Gamma, Burr and Inverse Burr), I configure the sampling as follows:

- Chains: 4 independent Markov chains. I use multiple chains for three main reasons: to assess that each chain converges to the same stationary distribution, to explore a large sets of values for each parameters, the chains being independent and to verify that the chain does not get stuck due to poor initialization.

[5] `https://mc-stan.org/cmdstanpy/users-guide/examples/MCMC\%20Sampling.html`

- Samples per chain: 2000 iterations, including warm-up. The warm up is used by Stan and tunes the step size as well other parameters linked to the sampling algorithm. Also, the length of the sample is once again useful to ensure the convergence of the distribution.
- Outputs: posterior samples, means, medians, Std, 95% credible intervals, and log-likelihood values

In the following tables, I present the priors used to initialize the Monte Carlo sampling. For each model, I employed "semi"-weakly informative priors, which are designed to be consistent with plausible parameter ranges while still allowing the data to influence the posterior. These priors are defined as truncated normal or Log-Normal distributions (denoted $\mathcal{N}^+$ and $\log \mathcal{N}^+$), which constrain the parameters to remain positive.

By "semi"-weakly informative priors, I mean that the model defining the distribution is not fully optimized. However, since I already have parameter estimates derived from my frequentist approach (such as MLE and QME), I use these estimates to inform the location of my (Log)Normal distributions, depending on the skewness of the frequentist samples were. In other words, my mean parameter estimation from the frequentist approach is used as hyperparameter to inform the priors for the parameters in a hierarchical structure. Specifically, for the mean of the prior, I use a weighted average of the MLE and QME estimates, which accounts for the precision of each estimate. This is computed as follows:

$$\hat{\theta}_{\text{prior}} = \frac{\hat{\theta}_{MLE}/\hat{\sigma}^2_{MLE} + \hat{\theta}_{QME}/\hat{\sigma}^2_{QME}}{1/\hat{\sigma}^2_{MLE} + 1/\hat{\sigma}^2_{QME}} \,, \tag{3.41}$$

where $\hat{\theta}_{MLE}$ and $\hat{\theta}_{QME}$ are the parameter estimates from the MLE and QME methods, and $\hat{\sigma}_{MLE}$ and $\hat{\sigma}_{QME}$ are their respective standard errors. This formula uses the precision to weight the estimates, ensuring that estimates with higher precision receive more influence in the determination of the prior mean. This also result in a determination of the prior with tighter confidence intervals.

When combining two independent, unbiased estimators :

$$\hat{\theta} = w_1\hat{\theta}_1 + w_2\hat{\theta}_2,$$

the variance of the mixture is :

$$\mathrm{Var}(\hat{\theta}) = w_1^2\,\mathrm{Var}(\hat{\theta}_1) + w_2^2\,\mathrm{Var}(\hat{\theta}_2) = w_1^2\,\sigma_1^2 + w_2^2\,\sigma_2^2,$$

subject to $w_1 + w_2 = 1$. Minimizing this leads to :

$$w_i \;=\; \frac{1/\sigma_i^2}{1/\sigma_1^2 + 1/\sigma_2^2},$$

The weights proportional to the precision $1/\sigma^2$, which yields the smaller variance for Equation (3.41).

Event	μ_{var}	σ^2_{var}	μ_{std}	σ^2_{std}	$\sigma^2_{\text{var}} < \sigma^2_{\text{std}}$
100	15.597	0.008	16.113	0.013	True
110h	9.685	0.028	9.421	0.031	True
1500	12.522	0.010	12.094	0.012	True
400	16.005	0.019	15.838	0.024	True
DT	4.823	0.001	4.712	0.002	True
HJ	13.848	0.013	12.895	0.015	True
JT	4.330	0.022	4.358	0.022	True
LJ	11.140	0.012	10.547	0.013	True
PV	4.558	0.001	4.502	0.001	True
SP	7.145	0.004	7.049	0.007	True

Table 3.5: Comparison of merged α-prior of Burr XII using inverse-variance (var) vs inverse-std (std) weighting.

Furthermore, to introduce a reasonable degree of uncertainty, I set the standard deviation of the prior to $\sigma = 1$, which defines a weakly informative prior that allows for moderate uncertainty around the mean. This choice reflects a belief that the parameter values are likely to be close to the frequentist estimates but still allows room for data to influence the posterior distribution.

While the priors provide some guidance, the likelihood approximation from the data plays a dominant role in overcoming the influence of the priors during the sampling process. The "semi"-informative priors, while not fully optimized, significantly improve convergence, sampling efficiency, and speed, especially by constraining the parameter space while still allowing the data to largely inform the posterior.

Distribution	Stan Parameters	Prior Distributions	Prior Type
Weibull	`shape c`, `scale a`	$\texttt{c} \sim \mathcal{N}^+(\hat{c}_{\text{prior}}, 1)$ $\texttt{s} \sim \mathcal{N}^+(\hat{c}_{\text{prior}}, 1)$	Event-specific
Log-Normal	`shape s`, `scale` θ	$\texttt{s} \sim \log\mathcal{N}^+(\hat{s}_{\text{prior}}, 1)$ $\theta \sim \mathcal{N}^+(\hat{\theta}_{\text{prior}}, 1)$	Event-specific
Gamma	`shape k`, `scale` θ	$\texttt{k} \sim \log\mathcal{N}^+(\log(\hat{k}_{\text{prior}}), 1)$ $\theta \sim \log\mathcal{N}^+(\log(\hat{\theta}_{\text{prior}}), 1)$	Event-specific
Gen. Gamma	`scale a`, `shape d`, `shape p`	$\texttt{a} \sim \log\mathcal{N}^+(\log(\hat{a}_{\text{MLE}}), 1)$ $\texttt{d} \sim \mathcal{N}^+(\hat{d}_{\text{MLE}}, 1)$ $\texttt{p} \sim \mathcal{N}^+(\hat{p}_{\text{MLE}}, 1)$	Event-specific
Burr	`scale s`, `shape c`, `shape k`	$\texttt{c} \sim \mathcal{N}^+(\hat{c}_{\text{prior}}, 1)$ $\texttt{k} \sim \mathcal{N}^+(\hat{k}_{\text{prior}}, 1)$ $\texttt{s} \sim \mathcal{N}^+(\hat{s}_{\text{prior}}, 1)$	Event-specific
Inv. Burr	`scale` θ, `shape` τ, `shape` α	$\alpha \sim \mathcal{N}^+(\hat{\alpha}_{\text{prior}}, 1)$ $\tau \sim \mathcal{N}^+(\hat{\tau}_{\text{prior}}, 1)$ $\theta \sim \mathcal{N}^+(\hat{\theta}_{\text{prior}}, 1)$	Event-specific

Table 3.6: Priors used in each Stan model. Truncated normals ($\mathcal{N}^+$) are used to enforce positivity where required.

Bayesian credible intervals In contrast to the frequentist view, the Bayesian framework treats parameters $\boldsymbol{\theta}$ as random variables. Prior beliefs about these parameters are encoded through a prior distribution $p(\boldsymbol{\theta})$, which is then updated with the observed data $\mathbf{x}$ to

obtain a posterior distribution as a proxy of (3.39):

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

As explained above,, to explore this posterior, I use Markov Chain Monte Carlo (MCMC) sampling defined by (3.40). From these samples, I compute a $(1-\alpha)$ credible interval using empirical quantiles:

$$[\theta_{\text{lower}}, \theta_{\text{upper}}] = \left[\text{percentile}_{\alpha/2}, \text{percentile}_{1-\alpha/2}\right]$$

In addition to credible intervals, the posterior samples naturally provide estimates of the mean, the median and the standard deviation for each parameter:

$$E[\theta] \approx \frac{1}{T}\sum_{t=1}^{T}\theta^{(t)}, \tag{3.42}$$

$$\text{Std}(\theta) \approx \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(\theta^{(t)} - E[\theta]\right)^2}, \tag{3.43}$$

$$\text{Med}[\hat{\theta}] \approx \text{median}\left(\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(T)}\right). \tag{3.44}$$

This posterior summarization was implemented using Bayesian inference tools such as `cmdstanpy`, in order to quantify the uncertainty for all model parameters across multiple distributions.

Given the observed data and the chosen prior, there is a $(1-\alpha)\%$ probability that the true parameter value lies within this interval.

Estimation results

In the following tables are presented the estimated parameters for the 100m as well as the PDF, $f(x)$, and their equivalent CDF, $F(x)$.

Weibull						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	13.218	13.217	0.067	13.083	13.342
	scale s	8.3	8.3	0.004	8.294	8.308
QME	shape c	10.0	10.0	0.0	10.0	10.0
	scale s	8.143	8.144	0.008	8.127	8.156
Bayes.	shape c	13.214	13.213	0.053	13.111	13.318
	scale s	8.3	8.3	0.004	8.293	8.308

LogNorm						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale θ	0.09	0.09	0.0	0.089	0.091
	shape s	7.962	7.963	0.004	7.955	7.97
QME	scale θ	0.068	0.068	0.001	0.066	0.07
	shape s	8.064	8.063	0.008	8.05	8.08
Bayes.	shape s	0.09	0.09	0.0	0.089	0.091
	scale θ	7.962	7.962	0.004	7.955	7.97

Gamma						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape k	126.843	126.868	1.212	124.449	129.055
	scale θ	0.063	0.063	0.001	0.062	0.064
QME	shape k	63.404	63.382	0.303	62.855	64.013
	scale θ	0.117	0.117	0.001	0.116	0.118
Bayes.	shape k	126.854	126.865	0.943	125.024	128.64
	scale θ	0.063	0.063	0.0	0.062	0.064

GenGamma						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
	scale a	6.823	6.815	0.161	6.524	7.122
MLE	shape d	22.549	22.591	0.947	20.776	24.396
	shape p	6.785	6.746	0.349	6.206	7.485
	scale a	7.506	7.507	0.046	7.413	7.596
Bayes.	shape d	18.533	18.525	0.294	17.969	19.122
	shape p	8.65	8.649	0.169	8.318	8.986

Burr						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	15.494	15.484	0.091	15.347	15.676
	shape k	3.992	3.997	0.116	3.747	4.191
	scale s	8.963	8.965	0.023	8.911	9.002
QME	shape c	20.279	20.409	0.612	18.59	21.347
	shape k	1.923	1.87	0.168	1.683	2.412
	scale s	8.428	8.413	0.05	8.352	8.571
Bayes.	shape c	15.467	15.468	0.107	15.255	15.671
	shape k	4.188	4.183	0.188	3.837	4.57
	scale s	8.996	8.996	0.035	8.929	9.065

InvBurr						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape α	0.418	0.418	0.007	0.407	0.431
	shape τ	31.129	31.151	0.304	30.516	31.686
	scale θ	8.453	8.454	0.008	8.437	8.466
QME	shape α	0.292	0.294	0.02	0.247	0.329
	shape τ	35.802	35.615	0.944	34.288	38.14
	scale θ	8.593	8.59	0.027	8.545	8.655
Bayes.	shape α	0.414	0.414	0.007	0.401	0.428
	shape τ	31.418	31.413	0.315	30.819	32.043
	scale θ	8.456	8.456	0.007	8.441	8.47

Table 3.7: Parameter estimates for the Weibull, Log-Normal, Gamma, Generalized Gamma, Burr and Inverse Burr distributions for the 100m event.

The tables containing the parameters for all events, distributions and methods are in Appendix 5

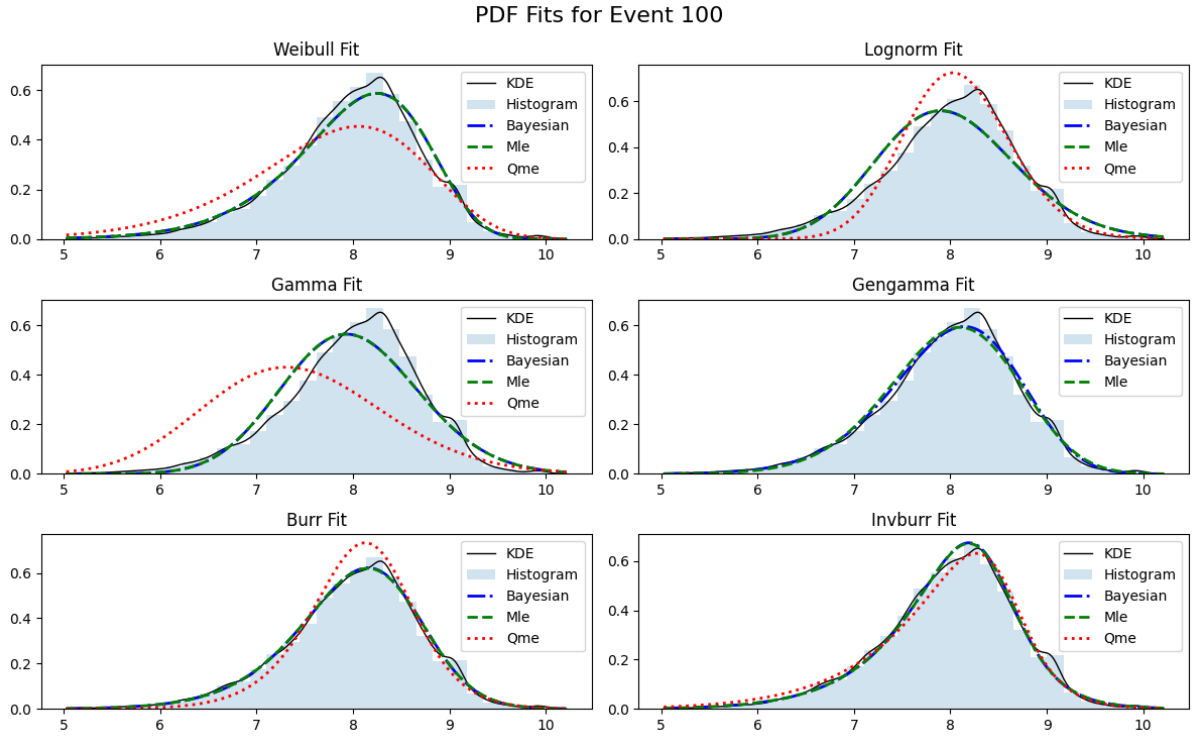


Figure 3.11: Estimated fitted distribution of the 100m

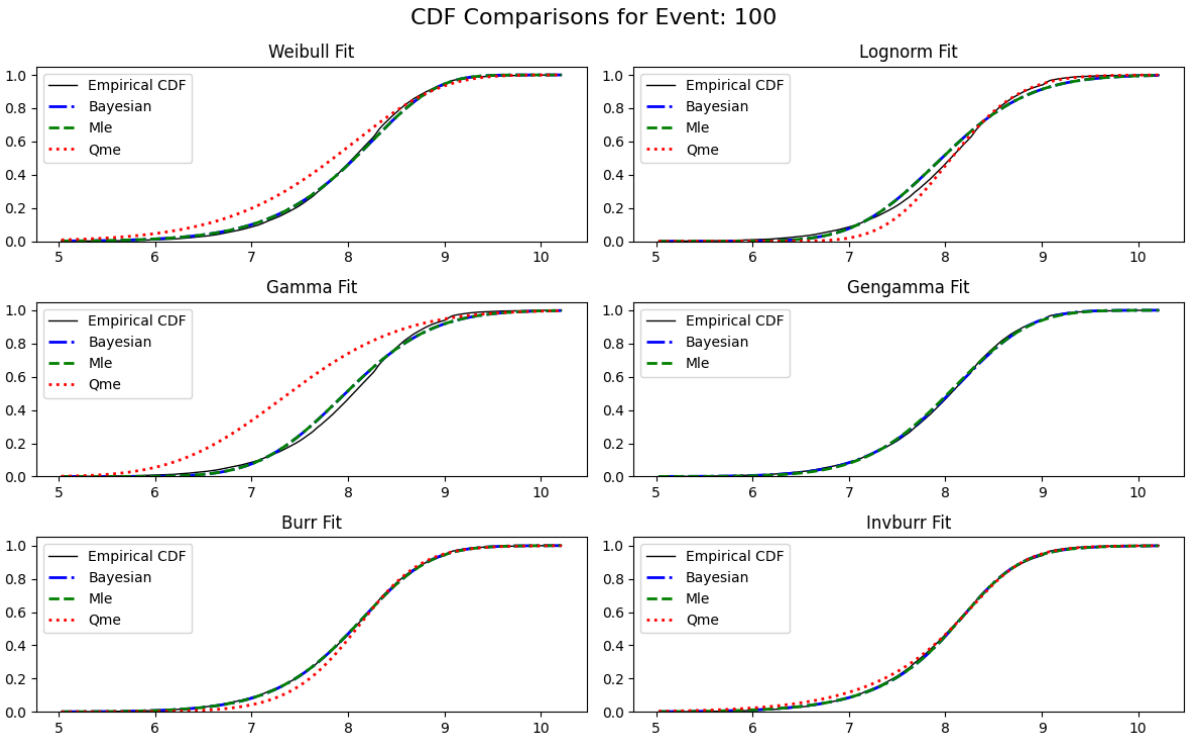


Figure 3.12: Estimated fitted cumulative distribution of the 100m

Both the PDF and CDF are available in the Appendix 5.

I can see some interesting insights that I'll be able to confirm later on. Weibull and Log-Normal distributions perform well in capturing the central part of the data but show notable discrepancies in the right tail, especially with the MLE and Bayesian, where they fail capture the extreme values.

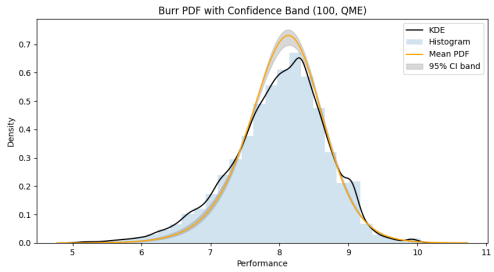
The GenGamma, Burr and Inverse Burr models shows heavier tails, suggesting that they are better suited for modeling data with extreme values or outliers. The GenGamma and Burr models, in particular, show a more gradual tail decay, which may be more appropriate for datasets with longer tails.

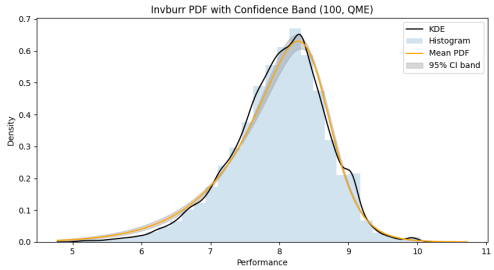
The MLE and Bayesian methods provide better fits to the overall distribution, while QME excels at capturing the tail behavior which is expected as I've designed the QME specifically on the tail quantiles.

Visual support

The different samples of the parameters are ideal to assess the stability of the density estimation of the model. Indeed, given these 500 samples, I can create a shaded interval around the estimated curve that will reflect the variability of that curve due to parameter uncertainty.

Once I have the samples, I can evaluate a linear space of values on the models generated by those parameter samples and stack those "probable" models in order to have a confidence region around the mean estimated set of parameters.





Visually, these graphs suggest that both the Burr and Inverse Burr distributions provide a good fit to the data as the non-parametric pdf using kernel density estimation and the estimated PDF are relatively close. The 95% confidence intervals offer additional information about the uncertainty in the fitted models. Indeed as it is QME parameters, it's expected that the confidence interval in the right tail is smaller than in the left tail because I used specific quantiles above 50% to fit the parameters.

3.3.3 Goodness of fit

After estimating each parameter of each distribution using MLE, QME and Bayesian inference, it is important to assess whether the theoretical model fits the observed data. In other words, I want to test the hypothesis :

Is it reasonable to assume the observed data were drawn from the fitted distribution?

Goodness of the full distribution fit

Instead of relying only on asymptotic p-values, I implemented a bootstrap-based Goodness of Fit framework that evaluates fit quality through repeated resampling. To assess this, I use multiple tests that answer the previous question. The theoretical ground behind the three first tests is coming from class notes issued by Eduardo García-Portugués (García, 2023)

- Kolmogorov–Smirnov (KS) Test: As mentioned in Section 3.2.4, measures the maximum absolute deviation between the empirical CDF and the theoretical CDF,

sensitive to global fit but not to local/tail deviations.

$$D_{\mathrm{KS}} = \sup_x \Big| F_{\mathrm{emp}}(x) - F_{\mathrm{model}}(x) \Big|.$$

- Cramér–von Mises (CvM) Test: Measures the integrated squared difference between the empirical and model CDFs, providing a smoother and more global assessment.

$$W^2 = n \int_{-\infty}^{\infty} \Big(F_{\mathrm{emp}}(x) - F_{\mathrm{model}}(x) \Big)^2 dF_{\mathrm{model}}(x).$$

  This test is widely used in survival analysis. Similar to the KS test, it compares the Empirical CDF $F_n(x)$ to the theoretical CDF $F(x)$ defined by the model. Instead of taking the maximal distance, it measures the sum squared difference over the whole set of CDF values. Similarly to the KS test, the lower the statistics of the CvM, the better the fit.

- Anderson–Darling (AD) Test: A weighted version of the CvM test that emphasizes the tails.

$$A^2 = \int_{-\infty}^{\infty} \Big(F_{\mathrm{emp}}(x) - F_{\mathrm{model}}(x) \Big)^2 w(x) dF_{\mathrm{model}}(x) \quad ,$$

  with $w(x) = 1/\Big((1 - F_{\mathrm{model}}(x)) * F_{\mathrm{model}}(x) \Big)$.

  The AD test is also an improvement of the Kolmogorov-Smirnov test (KS-test), placing more weight on the tails. Indeed, when the CDF $F_{model}(x) \to \{0; 1\}$, the numerator $w(x) \to 0$. This behavior puts a lot of weight on the squared difference near the tails. It is often revised in its discrete form as follow:

$$A^2 = -n - \sum_{i=1}^{n} \left(\frac{2i-1}{n} \right) \left(\ln(F_{model}(x)) + \ln(1 - F(x_{n+1-i})) \right).$$

- Revised Anderson–Darling (AD) Test: Taking the initial AD test and modifying the weights $w(x)$ so that it weights the right tails more. The test becomes:

$$A^2 = \int_{-\infty}^{\infty} \Big(F_{\mathrm{emp}}(x) - F_{\mathrm{model}}(x) \Big)^2 w(x) dF_{\mathrm{model}}(x),$$

  with $w(x) = 1/\Big((1 - F_{\mathrm{model}}(x)) \Big)$. It can be rewritten :

$$A^2 = -n - \sum_{i=1}^{n} \left(\frac{2i-1}{n} \right) \ln(1 - F(x_{n+1-i})) \quad .$$

  Indeed, from here, as the CDF, $F(x) \to 1$ the weighting function $w(x) \to 0$ which leads to a larger value of the test at the right tail of the distribution

- Log-likelihood: Measures fit quality from the likelihood perspective.

$$\log \mathcal{L} = \sum_{i=1}^{n} \log f(x_i \mid \theta).$$

Practically, the bootstrap procedure is implemented following a clear pipeline. First, I calculate the test statistics discussed earlier (Kolmogorov–Smirnov, Anderson–Darling, Cramér–von Mises, its revised version, and log-likelihood) comparing the observed data to the fitted model with the set of mean or the median estimates of the parameters. This provides the reference statistics under the assumed distribution.

Next, I perform parametric sampling under the null hypothesis by generating $B = 500$ synthetic datasets of the same size as the observed sample. These bootstrap samples are drawn from the fitted model distribution. For each of these bootstrap datasets, I recompute the same test statistics as in the initial step.

Finally, for each test statistic, I compute the empirical p-value based on the bootstrap distribution. The p-value is defined as:

$$\hat{p} = \frac{1}{B} \sum_{b=1}^{B} I[T_b \geq T_{\text{obs}}] \quad ,$$

where T_{obs} is the statistic computed on the observed data (AD, CvM, KS), and T_b is the statistic computed on the b-th bootstrap sample (AD, CvM, KS). This procedure allows me to generate bootstrapped p-values, avoiding over-estimating the fit when the model is fitted and tested on the same set.

Event	Model	Method	KS	p-KS	CvM	p-CvM	AD	p-AD	AD right	p-AD right	LogLik
100	Burr	Bayesian	0.009	0.0	0.635	0.018	4.68	0.002	-16711.608	0.34	-34549.927
100	Burr	QME	0.059	0.0	34.997	0.0	317.737	0.0	-18102.996	1.0	-35463.279
100	Burr	MLE	0.009	0.0	0.733	0.01	5.201	0.0	-16711.619	0.342	-34551.184
100	Gamma	MLE	0.057	0.0	37.567	0.0	224.17	0.0	-15506.299	0.0	-35962.579
100	Gamma	QME	0.337	0.0	1754.871	0.0	8314.913	0.0	989.721	0.0	-45126.195
100	Gamma	Bayesian	0.056	0.0	36.741	0.0	221.423	0.0	-15539.798	0.0	-35962.591
100	Gengamma	Bayesian	0.015	0.0	2.66	0.0	16.208	0.0	-16626.448	0.104	-34650.989
100	Gengamma	MLE	0.023	0.0	5.399	0.0	32.999	0.0	-16402.552	0.0	-34732.753
100	Invburr	Bayesian	0.014	0.0	1.414	0.0	10.654	0.0	-16903.052	0.928	-34640.477
100	Invburr	QME	0.034	0.0	8.505	0.0	82.303	0.0	-16126.183	0.0	-34952.523
100	Invburr	MLE	0.013	0.0	1.232	0.0	9.725	0.0	-16905.328	0.894	-34640.389
100	LogNorm	Bayesian	0.063	0.0	46.799	0.0	280.361	0.0	-15355.311	0.0	-36366.808
100	LogNorm	QME	0.08	0.0	45.963	0.0	804.766	0.0	-18005.575	1.0	-39953.19
100	LogNorm	MLE	0.063	0.0	46.577	0.0	279.551	0.0	-15362.115	0.0	-36366.807
100	Weibull	Bayesian	0.031	0.0	7.621	0.0	51.192	0.0	-16938.816	0.972	-34935.664
100	Weibull	QME	0.143	0.0	274.299	0.0	1412.138	0.0	-11413.581	0.0	-37109.878
100	Weibull	MLE	0.031	0.0	7.615	0.0	51.081	0.0	-16943.137	0.974	-34935.662

Table 3.8: Full distribution Goodness-of-fit for the 100m based on the median estimated parameter

Looking at this table and the tables for the other events in Appendix 5, I can clearly notice some key takeaways:

- The bootstrapped p-values for the first three tests were very small or null, leading to believe that the model does not statistically approximate the empirical data well. Only a few settings for specific events were significant in terms of p-values, and the first three tests led to the same conclusion (i.e., Shot Put at $\alpha = 5\%$ for the Bayesian and MLE estimation of Log-Normal).

- Looking at the log-likelihood, I see that the MLE and Bayesian estimates are similar. Indeed, in terms of interpretation, that makes a lot of sense. These two fittings were not constrained to fit one tail better than the other.
- On the contrary, the log-likelihood of the QME fitting is generally more negative, revealing a worse overall fit. Once again, it makes sense, as my QME estimation method was designed to enhance the fit in the upper 50% of the population. This led to poorer fitting in the lower tail, increasing the likelihood of the global fitting.
- Similarly, I observe the same behavior when looking at the statistic values of the AD, KS, and CvM tests. Indeed, all three tests reveal larger statistics for QME, while the Bayesian and MLE fittings of a distribution provide tests of smaller magnitude, yet similar among them.
- On the contrary, the modified AD test provides reasons to think that the data could have been generated by the fitted distribution. Among the three fitting methods, the Bayesian approach seems to lead to a more significant test, providing greater p-values and smaller test statistics for most of the distribution.

Limits of the Goodness-of-fit tests

As far as being interesting to use, the tests assessing the Goodness-of-fit may be too restrictive, as their output is reduced to a single scalar, the p-value. Hence, that value does not always capture all aspects of the quality of the model. For instance, the different tests are sensitive to different parts of the distribution. KS emphasizes global maximum deviation, AD emphasizes the tails. To prevent that, it remains interesting to assess the quality of the fit using a visual tool. Graphical tools can be used for instance:

- PP plots (probability–probability plots) are used to compare empirical and theoretical quantiles. A good fit would result in a straight diagonal line. In my scoring framework, probability–probability plots make perfect sense, as the scoring function is evaluated using $1 - F(x)$, which yields values in the range [0, 1].

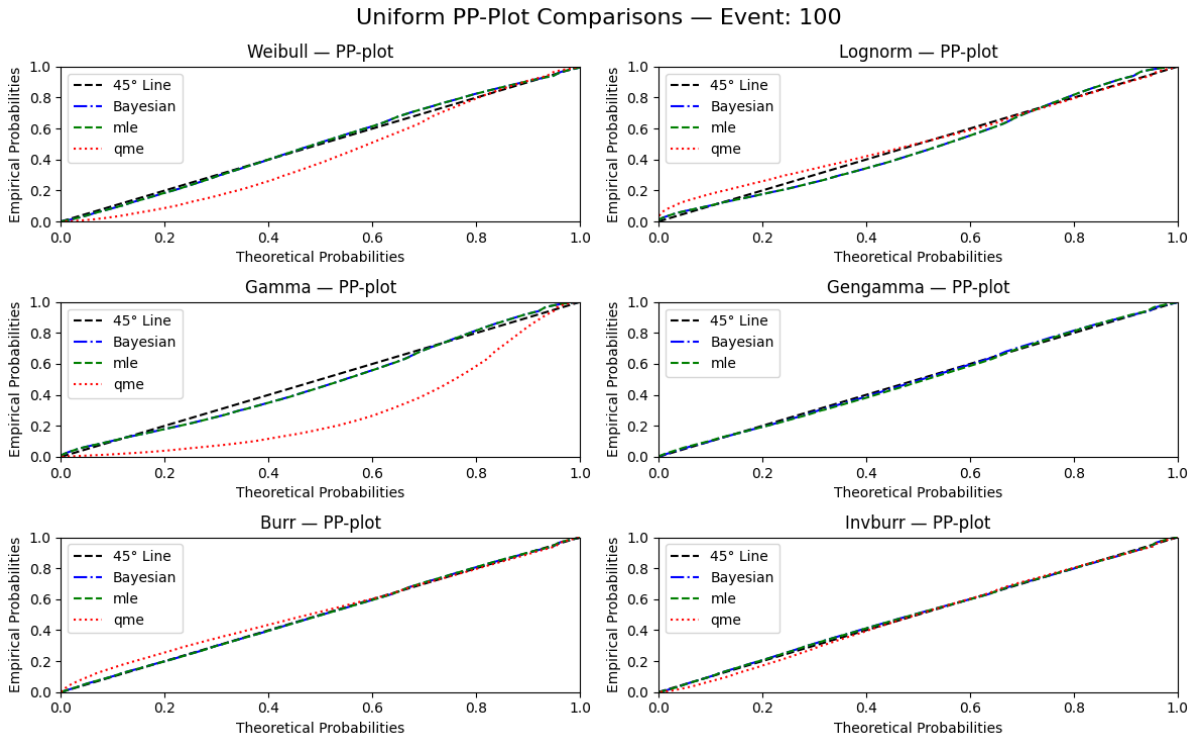


Figure 3.14: PP-plot for the fitted 100m

Looking at the above set of graphs and in Appendix 5, I can highlight some key insights.

- Two distribution models are indeed performing better than the others across the entire distribution. Specifically, the Burr and GenGamma models have fitting lines that almost merge with the diagonal line, which represents a perfect fit. This is particularly true for the Bayesian and MLE estimation methods. This highlights that these two estimation methods were not specifically designed to fit a particular region of the distribution but remain overall efficient. In contrast, the QME line starts to align with the straight line beginning at the ∼50% percentile, indicating a focus on the upper tail.

Goodness of the Tail Distribution Fit

Up to this point, I have used common tests to assess the fit across the entire distribution. However, in my context, the scoring formula $p(x) = -\ln(1 - F(x))$ (see Equation (2.5)) suggests that the quantiles contributing most to the points are located in the upper tail. Specifically, the quantile that gives a performance of more than one point is determined by:

$$1 = -\ln(1 - F(x)) \quad \Leftrightarrow \quad F(x) = 1 - e^{-1} = 0.632.$$

For values below this percentage, any performance will yield a score lower than one. Conversely, for quantiles above this threshold, the assigned points increase exponentially. This aligns with the intuition used in Section 3.3.2, where the model was fitted using QME

on the upper quantiles.

To assess how well the distributions fit the upper quantiles, I designed a test inspired by "A Goodness-of-Fit Test for the Distribution Tail" (Diebolt et al., 2023). In this article, the author presents a method for comparing quantile estimates to those derived from a semi-parametric approach based on Extreme Value Theory (EVT). EVT is particularly useful for modeling the tail behavior of distributions, which corresponds to the quantiles I am interested in.

Instead of using a semi-parametric model, which would require an initial estimation, I chose to directly compare the quantiles calculated from the data with the quantiles from the fitted model. The difference between these two sets of quantiles, denoted as $\delta_{\text{obs}} = q_{\text{empirical}} - q_{\text{fitted}}$, quantifies how well the model fits the observed data in the tail. The intuition here is that if the model fits the observed data, the differences between the observed and fitted quantiles should be small in the tail. Large differences, on the other hand, indicate a poor fit in the tail.

This leads to the central question:

Could the data in the tail plausibly have been generated from the fitted distribution?

To answer this, I employed a bootstrap resampling strategy to compute the variability of the quantile differences and construct confidence intervals for these differences.

For each bootstrap sample, quantiles are computed in the same manner as the observed quantiles, and the quantile differences are calculated as follows:

$$\delta_{\text{sim}} = q_{\text{sim}} - q_{\text{model}},$$

where q_{sim} represents the quantiles of the simulated data. These differences are then stored to compute a confidence interval.

The confidence interval for the quantile differences is derived from the 2.5th and 97.5th percentiles of the bootstrap differences:

$$\delta_{\text{ci\_lower}} = \text{percentile}(\text{bootstrap\_deltas}, 2.5),$$
$$\delta_{\text{ci\_upper}} = \text{percentile}(\text{bootstrap\_deltas}, 97.5).$$

Finally, the p-value to assess the fit of the empirical upper quantiles to the fitted model is computed by comparing the observed quantile differences to the bootstrap distribution:

$$p_{\text{val}} = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{B} \sum_{b=1}^{B} I\left[|\delta_b(q)| \geq |\delta_{\text{obs}}(q)|\right],$$

where $\delta_{\text{obs}}(q)$ is the observed quantile difference for quantile q, and $\delta_b(q)$ represents the bootstrap quantile difference for the same quantile.

Additionally, I calculate a score to define the magnitude of the test, which is the Mean Squared Error (MSE) of the observed quantile differences across all evaluated quantiles:

$$\delta_{\text{score}} = \frac{1}{Q} \sum_{q=1}^{Q} \left(\delta_{\text{obs}}(q)\right)^2,$$

where Q is the number of quantiles evaluated.

Event	Model	Method	δ_{score}	p-value
100	Burr	QME	0.002	0.015
100	Burr	Bayesian	0.004	0.069
100	Burr	MLE	0.004	0.126
100	Gamma	MLE	0.051	0.001
100	Gamma	QME	0.15	0.016
100	Gamma	Bayesian	0.052	0.001
100	Gengamma	Bayesian	0.009	0.118
100	Gengamma	MLE	0.006	0.083
100	Invburr	MLE	0.006	0.024
100	Invburr	QME	0.001	0.128
100	Invburr	Bayesian	0.005	0.028
100	LogNorm	MLE	0.083	0.0
100	LogNorm	QME	0.003	0.068
100	LogNorm	Bayesian	0.083	0.0
100	Weibull	MLE	0.02	0.021
100	Weibull	QME	0.011	0.064
100	Weibull	Bayesian	0.02	0.02

Figure 3.15: Tail distribution Goodness of fit for the 100m based on the Median estimated parameter

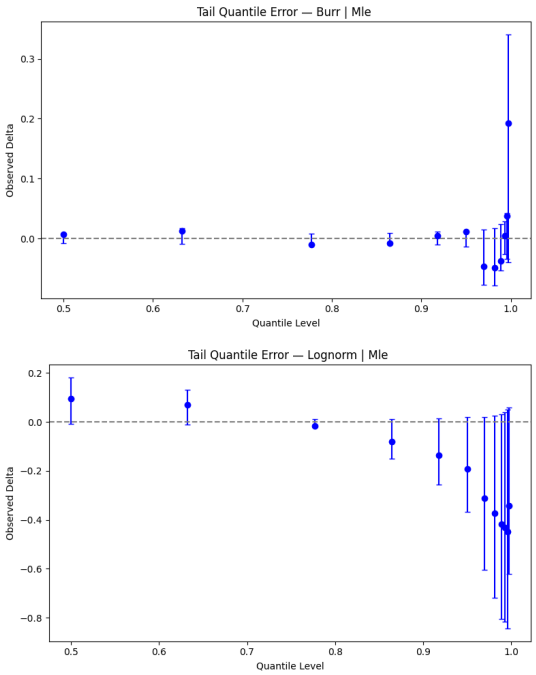


Figure 3.16: Tail's Plot Evaluation of the MLE's fit for Burr and Log-Normal

On the graph above, I compare the tail's goodness of fit using the sets of median parameter estimates by MLE for both the Burr and Log-Normal distributions. When compared with the δ_{score} in the previous table 3.15, these two settings are respectively the best and the worst performing models regarding the tail's fit. I can derive several key and interpretable takeaways from these graphs.

- The Burr model with the MLE settings clearly provides a good visual fit. The delta between the empirical quantiles and the fitted quantiles remains very close to zero for quantile levels ranging from 50% to 99.6%. The most extreme quantile tested, at the 99.75% level (which corresponds to 6 points in this scoring system defined in Equation (2.5)), shows an observed difference of less than 0.2 m/s. This difference is still reasonable, as it represents less than 0.1 seconds over 100 meters at this level of 6 points.

- In contrast, the Log-Normal model for the 100m shows a poorer fit. Across all quantile levels, the absolute delta is consistently larger, ranging from 0.1 m/s at the 50% quantile to 0.45 m/s at the 99.5% quantile.

- A notable observation is that while the Burr model excels in fitting the upper quantiles, it appears to maintain a more stable fit across the entire distribution, indicating that it is a robust model for the range of performance levels observed in the 100m. On the other hand, the Log-Normal model shows greater variability in the delta values, suggesting that it might be more sensitive to the tail behavior but less consistent across the full distribution.

Goodness of Fit Conclusion

At this stage, given the extensive data regarding the fit of each model, it remains challenging to definitively state that one model outperforms the others. Nevertheless, I can confidently assert that some of the settings provide an adequate fit for the tail. This allows me to proceed with the analysis of event scoring.

3.4 Model rankings

In my hands, I now have 34 different models. When combining our six distributions with our three (only two for Generalized Gamma) estimation methods, and two sets of estimates (mean and median), I am left with thirty-four different models.
To transform the distribution into a scoring system according to Equation (2.5), I need to evaluate which one of the above would be best suited for converting distributions to scores. To complete this task, I can use Goodness of Fit metrics to determine which of these models best fits the data. Recall that I have computed six metrics on observed data: AD, Right AD, CvM, KS and δ_{score} statistics, as well as the Log Likelihood (see section 3.3.3). In addition to these four metrics, I could also compute two other commonly used metrics, the BIC and the AIC.

In a nutshell, I will compute different rankings among these 34 models to see which one performs the best across the 10 events of the decathlon using these metrics. Then, I will evaluate the best-ranked model using sport knowledge supported by statistics.

I will construct multiple sets of rankings:

- One ranking based on likelihood scores.
- Two based on test statistics
    1. One based on δ_{score}, which focuses solely on the tail.
    2. One based on Right AD, which evaluates the entire fit but places more weight on the right tail.
- One ranking based on the set of World Records or other performance sets, for instance.

To create the first three rankings, we must first aggregate statistical scores per distribution and per method with Model $m \in \{1, \dots, 34\}$ and event $i \in \{\text{decathlon events}\}$:

$$s_m = \sum_{i=1}^{10} s_{m,i}. \tag{3.45}$$

3.4.1 Log likelihood ranking

To begin this ranking, it is important to address the log-likelihood. It is interesting to consider whether other information criteria such as BIC or AIC should be introduced in our scoring.

While maximizing the log-likelihood gives the best possible fit under a given model, it does not account for model complexity. More flexible models (like Generalized Gamma) tend to fit better simply because they have more degrees of freedom. To address this, two commonly used penalized criteria are:

- AIC (Akaike Information Criterion):

$$\text{AIC} = 2k - 2\log \mathcal{L}, \tag{3.46}$$

  where k is the number of free parameters in the model. AIC balances fit and complexity by penalizing each added parameter with a constant factor.

- BIC (Bayesian Information Criterion):

$$\text{BIC} = k\log n - 2\log \mathcal{L}, \tag{3.47}$$

  where n is the number of observations. BIC imposes a stronger penalty for complexity, especially when n is large.

In my framework, to obtain a total BIC/AIC for a specific model, I sum the log-likelihood across all ten events according to Equation (3.45), and compute AIC/BIC using Equations (3.46) and (3.47) accordingly:

$$\text{BIC}_m = k \cdot \log(N) - 2 \cdot \sum_{i=1}^{10} \log \mathcal{L}_i,$$
$$\text{AIC}_m = 2k - 2 \cdot \sum_{i=1}^{10} \log \mathcal{L}_i.$$

Thus, I compute a global AIC and BIC for the entire model rather than summing all individual AIC/BIC values. In a multi-event setting like this one, where each event $i \in \{1, \ldots, 10\}$ has its own number of observations n_i, summing individual BIC would introduce bias:

- When summing per-event BIC:

$$\sum_{i=1}^{10} (k \cdot \log(n_e) - 2\log \mathcal{L}_i)$$

  I treat each event equally, regardless of the sample size n_i. Events with smaller sample sizes n_i are penalized less per parameter, but this lower penalty is weighted just as heavily in the total sum.

- This can distort the overall model evaluation, since events with small n contribute to little penalty relative to their information content, such as hurdles and discus in our framework.

However, computing a global BIC treats the total number of samples as fixed. Since the global log-likelihood sums all data points, it also accounts for model complexity more consistently across events. Also, the global k, $\sum_i k$, is similar in my models and often small compared to the sum of the log-likelihood. After computing a global AIC, BIC, and Log Likelihood, I see that the ranking is the same. Using this intuition, I then use only the global log-Likelihood as part of my ranking system.

TOP 5 Likelihood Scores								
Model	**Method**	**Est**	**LL**	**AIC**	**BIC**	**Rank LL**	**Rank AIC**	**Rank BIC**
Gengamma	Bayesian	Median	-479877.9	959761.79	959791.92	1	1	1
Gengamma	Bayesian	Mean	-479970.48	959946.97	959977.1	2	2	2
Invburr	MLE	Mean	-480550.43	961106.85	961136.98	3	3	3
Invburr	MLE	Median	-480553.3	961112.6	961142.72	4	4	4
Burr	MLE	Mean	-480771.47	961548.93	961579.06	5	5	5

Table 3.9: Top 5 Models Based on Log-Likelihood Scores

As mentioned earlier, the log-likelihood ranking coincides with both the AIC and BIC rankings. It is interesting to note that this ranking is based on the fit of the entire distribution and not just the right tail. Therefore, this top 5 reflects the overall fit of the distribution, and it is normal not to find any "QME" fits since I focused on the right tail when designing the Quantile Matching Estimation. The best QME fit is ranked 21st.

3.4.2 AD right and delta score ranking

To rank the models based on these test scores, I use the Z-score to standardize my measure of test scores across all events. The Z-score for a given metric m is defined as:

$$z_m = \frac{s_m - \mu_s}{\sigma_s}, \tag{3.48}$$

where:

- s_m : The score for model m.
- μ_s : The mean score across all models.
- σ_s : The corresponding standard deviation.

Concretely, using the (3.45) Equation, I need to:

1. Calculate the z-score for each event using the statistics and comparing it to the same event for all models using (3.48).
2. Sum the z-scores of a model over the 10 events using (3.45).

Lower values of these statistics indicate better fits, so my ranking will be based on the lowest Z-scores to reflect better performance.

TOP 5 Delta Scores				
Model	**Method**	**Est.**	**Z-Delta Score**	**Rank**
Burr	QME	Median	-6.97	1
Invburr	QME	Median	-6.86	2
Invburr	QME	Mean	-6.85	3
Lognorm	QME	Median	-6.79	4
Lognorm	QME	Mean	-6.79	5

TOP 5 AD Right Score				
Model	**Method**	**Est.**	**Z-AD Right**	**Rank**
Lognorm	QME	Median	-9.11	1
Lognorm	QME	Mean	-9.08	2
Gengamma	Bayesian	Mean	-5.86	3
Weibull	MLE	Mean	-5.02	4
Weibull	MLE	Median	-5.01	5

Table 3.10: Top 5 Models Based on Statistics Scores

The ranking based on the Delta scores reveals that the Burr distribution fitted using QME provides a better fit at certain levels. It makes a lot of sense that the top five is solely composed of QME-fitted models, since QME was designed to provide an accurate fit specifically on the tail. The first non-QME fitting comes in 8th position, with the Generalized Gamma fitted using a Bayesian approach.

On the other hand, the Right AD score ranking provides more moderate results with models fitted using different methods. Indeed, even though this Right AD test gives more weight to the tail, it still considers the entire distribution.

3.4.3 Fairness score Ranking

Another intuitive ranking, that can be derived, is based on a set of known performance. Indeed, one could argue that a scoring model should award the same amount of points to the set of World record. To this end, I decide to create a fourth ranking that will highlight the models with the smallest standard deviation in terms of points among the world record performance.

Using the performance-score relation (2.5) (without the scaling constant λ):

$$p(x) = -log(1 - F(x)),$$

where $F(x)$ is the CDF of the model evaluated at the athlete's performance x.

I calculate for a set of given performance, the world record of the 10 events of decathlon $\{x_{100}, \ldots, x_{1500}\}$, the scores that each performance would receive under all the defined settings. For each settings, I then have 10 specific $p(x_i)$, one for each World Record. From these score values I calculate the standard deviation for each settings $m = 1...34$:

$$\sigma_m = \sqrt{\frac{1}{10}\sum_{i=1}^{10}(p(x_i) - \mu_m)},$$

with $\mu_m = \frac{1}{10}\sum_{i=1}^{10} p(x_i)$ and $x_i \in \{x_{100}, \ldots, x_{1500}\}$, the set of World record. I can finally rank my models based on these standard deviations.

Top 5 Std WR Ranking				
Model	**Method**	**Est.**	**Std**	**Rank** Std
Invburr	QME	Mean	0.745	1
Invburr	QME	Median	0.752	2
Invburr	Bayesian	Mean	0.808	3
Invburr	Bayesian	Median	0.809	4
Burr	QME	Median	0.912	5

Table 3.11: Top 5 Models based on lowest Std among WR performance

Using this ranking I ensure that the points awarded to the World Record are of the same magnitude. This relies to our initial principle : "The same number of points should be attributed to performance reached by the same amount of people". Even tough the World record of each event is subject to changes and the density around it varies. It seems reasonable to adopt this ranking. The ranking classifies the InvBurr Distribution estimated with QME and Bayesian as distributions awarding the set of World Record with Score of the same scale.

3.4.4 Model Selection

Using those previous rankings, I see clearly four models coming on top with multiple appearances in the top contestant of each ranking.

- Burr distribution using Quantile Matching Estimation (QME), a model that approximates predefined fixed quantiles of interest but also provides a small variance among World Record Scoring.
- Log-Normal distribution using QME a model that is emphasizing the tail's fit well, considering not only predefined quantiles but on the full CDF domain.
- Inverse Burr Distribution on QME. A distribution that rewards the World Record with the same level of points.
- GenGamma Bayesian, a model that approximates the full empirical distribution the best, regardless weights on specific regions

In section 2.3.1 I mentioned that for the set of plausible results, the scoring should be slightly progressive. This means that the scoring function should be convex for all set of plausible values.

This helps us selecting two models based solely on the graph of each event. Indeed I can see when plotting the scores of each event using the parameters of the previously mentioned model that the Invburr distribution model tends to provide slightly regressive plots near the end which indicates that a trigger value is reached and that behind this performance, the score is increasing less rapidly at the upper tail than expected. This can occur due to overfitting the model to the data, especially if the tail is not properly captured. A concave curve typically means that the model's tail behavior is not steep enough to match the expected theoretical behavior, and the CDF might not be increasing

quickly enough near the upper quantiles. The score should, past that inflection point, be higher. The distribution seems in reality to under-estimate the population at larger quantile.

On the other hand, the Generalized Gamma have a behavior that changes across events. Sometimes over-estimating the density, for example in Pole Vault or on the 110mh, the scoring is not progressive enough, sometimes it under-estimates the density for example for the Javelin, the Long Jump and the Shot Put, are, on the other hand, showing an inverse behavior.

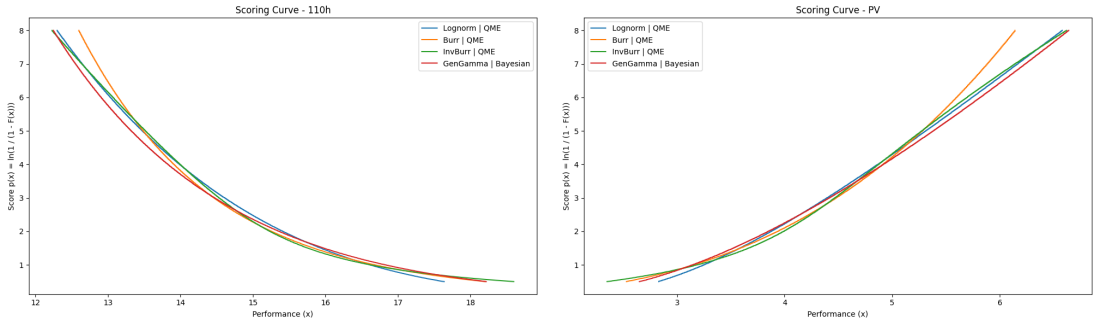


Figure 3.17: Scoring function For 110m Hurdle and Pole Vault: underestimating the score for GenGamma and showing regressiveness for InvBurr on the upper quantile

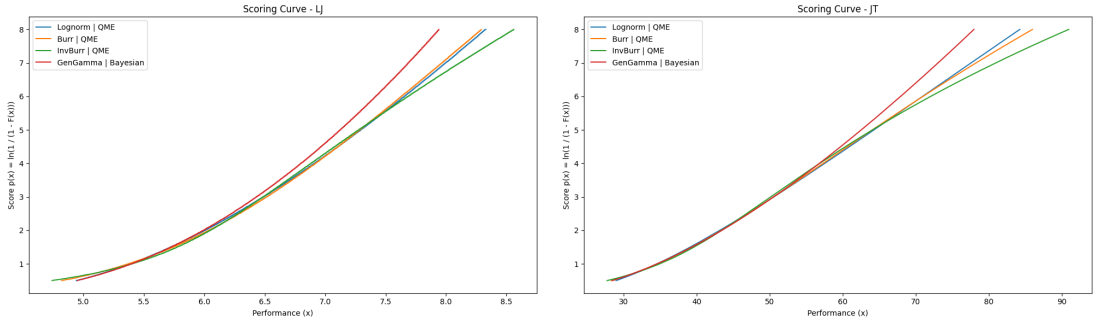


Figure 3.18: Scoring function For Long Jump and Javelin Throw: overestimating the score for GenGamma and showing regressiveness for InvBurr on the upper quantile

After discarding those two models, I remain with two design. One that performs well on the full distribution, still putting more importance on the tail, the Log-Normal model, and one that fits the upper quantile really well without caring about the lower tail of the distribution at all and that rewards the sets of World Record with the same amount of points, the Burr distribution.

The selection of the model is not objective anymore. Indeed, one could argue that fitting the tail is the only prerequisite while other could state that a design should be selected considering its evaluation on the whole empirical distribution.

Scoring Function

For the LogNormal distribution the scoring function is defined by:

$$p_i(x) = -\ln\left(1 - F_i(x)\right) = -\ln\left(1 - \Phi\left(\frac{\log\left(\frac{x}{\theta_i}\right)}{s_i}\right)\right).$$

With θ_i, the scale parameters and s_i the shape parameter of a given event i. Φ is the CDF of a standard Normal distribution.

On the other hand, for the Burr distribution, the score is defined by.

$$p_i(x) = -\ln\left(1 - F_i(x)\right) = -\ln\left(\left(1 + \left(\frac{x}{s_i}\right)^{c_i}\right)^{-k_i}\right).$$

In the following table, I computed the performance needed in order to obtain specific amount of points using those two models.

LogNorm (QME)										
Score	**100 (s)**	**LJ (m)**	**SP (m)**	**HJ (m)**	**400 (s)**	**110h (s)**	**DT (m)**	**PV (m)**	**JT (m)**	**1500 (mm:ss)**
0.5	12.63	4.94	7.72	1.46	56.70	17.64	22.33	2.83	28.97	4:39.15
1	12.12	5.39	9.02	1.57	54.51	16.61	26.79	3.25	34.57	4:25.54
2	11.50	6.00	10.98	1.73	51.87	15.41	33.69	3.88	43.17	4:09.36
3	11.08	6.49	12.64	1.85	50.07	14.61	39.67	4.39	50.59	3:58.42
4	10.75	6.91	14.16	1.96	48.65	13.99	45.31	4.87	57.54	3:49.89
5	10.47	7.29	15.62	2.05	47.47	13.48	50.79	5.31	64.29	3:42.79
6	10.23	7.65	17.05	2.14	46.43	13.03	56.23	5.74	70.95	3:36.67
7	10.02	8.00	18.45	2.23	45.52	12.65	61.66	6.16	77.59	3:31.25
8	9.83	8.33	19.85	2.31	44.69	12.30	67.14	6.58	84.27	3:26.38

Table 3.12: Scoring Using Log-Normal settings

Burr (QME)										
Score	**100 (s)**	**LJ (m)**	**SP (m)**	**HJ (m)**	**400 (s)**	**110h (s)**	**DT (m)**	**PV (m)**	**JT (m)**	**1500 (mm:ss)**
0.5	12.60	4.82	7.87	1.43	57.36	18.21	21.75	2.53	28.31	4:46.39
1	12.09	5.37	9.15	1.57	54.58	16.68	26.82	3.15	34.61	4:26.68
2	11.51	6.04	10.94	1.74	51.72	15.26	33.90	3.94	43.36	4:07.81
3	11.11	6.52	12.48	1.86	49.96	14.46	39.76	4.48	50.56	3:56.96
4	10.77	6.92	13.98	1.96	48.63	13.91	45.29	4.91	57.33	3:49.26
5	10.46	7.29	15.54	2.05	47.52	13.49	50.82	5.28	64.07	3:43.24
6	10.18	7.63	17.20	2.14	46.56	13.14	56.52	5.60	71.00	3:38.24
7	9.91	7.96	18.99	2.23	45.69	12.85	62.52	5.88	78.28	3:33.95
8	9.65	8.29	20.94	2.31	44.88	12.60	68.90	6.14	86.00	3:30.16

Table 3.13: Scoring Using Burr settings

It is well established in the track and field community that selecting the best model requires domain-specific knowledge. As shown in the lower range of the scoring, under 6 points, both models yield similar performances. However, when examining values above 6 points, performance estimations begin to diverge.

For instance, the Log-Normal model provides a noticeably worse fit in certain events, particularly in the 110m hurdles and in the pole vault. It tends to overestimate the cumulative distribution function $F(x)$, which results in over-attributing points. Concretely, it awards 8 points to a pole vault height of 6.58m even though the world record is 6.28m

and 8 points to a 110m hurdles time of 12.30 seconds, while the world record stands at 12.80 seconds.

Event	**Athlete**	**World Record**	**Burr**	**Log-Normal**
100m	Usain Bolt	9.58	8.26	9.42
LJ	Mike Powell	895	10.00	9.98
SP	Ryan Crouser	23.56	9.22	10.65
HJ	Javier Sotomayor	2.45	9.58	9.81
400m	Wayde van Niekerk	43.03	10.57	10.29
110mH	Aries Merritt	12.80	7.19	6.59
DT	Jürgen Schult	74.08	8.76	9.25
PV	Armand Duplantis	6.24	8.40	7.18
JT	Jan Železný	98.48	9.48	10.10
1500m	Hicham El Guerrouj	3:26.00	9.24	8.08
Std (across events)			0.91	1.27

Table 3.14: World Record Assessment: Point attribution comparison between Burr and Log-Normal models

Given that the lower quantiles yield similar point values across models, the key differences emerge in the extreme right tail of the distribution. Therefore, I believe it is more important to prioritize the quality of the fit in this upper tail than to optimize the overall fit across the full distribution.

Also there is something else important to mention. This is the ease of utilization. Indeed these performance-score tables should be used by non-statistician. It feels only right not to give a very complicated formula with numerical approximation due to the Φ function. An relevant argument in the world of sport favoring the use of Burr modeling is that the mathematical formula should be straight forward to use given a set of parameters. Hence the Burr CDF is easier to calculate because it has a closed form without the need to integrate anything.

Chapter 4

New Scoring model in action

4.1 Scoring system overview

After selecting, fitting, and evaluating models to score track-and-field performance, I chose a distribution and developed a scoring system that transforms performance into scores for all decathlon events. The model I selected fits exceptionally well with the second half of the distribution, making it particularly effective for distinguishing performance rankings in that range.

The new scoring system attributing a number of points to a given performance is based out of a Burr distribution of type XII defined in Section 3.3.1. Thanks to Equation (2.5) that can be modified removing the constant lambda that just scales the scores, the new system becomes :

$$p_i(x) = -\ln\left(1 - F_i(x)\right) = -\ln\left(\left(1 + \left(\frac{x}{s_i}\right)^{c_i}\right)^{-k_i}\right), \quad \text{with event } i = 1, \ldots, 10.$$

The best fit is obtained with the median parameter that I gathered using sampling detailed in Section 3.3.2 using Equations (3.36) and (3.44). Here is a summary of the estimated parameters. Note that for running events, 100m, 400m, 110m Hurdles and 1500m, the performance x is considered in speed being the $\frac{\text{distance}}{\text{time}}$ in m/s

Parameters	**100m**	**LJ**	**SP**	**HJ**	**400m**	**110h**	**DT**	**PV**	**JT**	**1500m**
shape c	20.409	6.984	5.55	8.297	14.987	8.044	3.723	3.139	3.884	10.013
shape k	1.87	4.288	1.88	3.659	4.752	20.294	3.015	262.259	2.977	12.172
scale s	8.413	650.396	9.751	180.342	8.079	9.557	34.455	1858.234	43.844	7.189

Table 4.1: Estimated Burr parameters for each event

4.1.1 Additional notes on the Burr model

Before assessing this model under specific condition, for example in a decathlon framework, I believe it is important to give some additional notes about what this model reveals for individual events. Indeed the Table 3.14 associating different points to the set of World

records provides some interesting insights that deserve some additional notes. For the Burr distribution, I can see that the 110m Hurdles' WR is awarded with the fewer points. First looking at the data, it can be explained by the density of French specialists in hurdles. Indeed France is renowned for having many specialists in that discipline with 6 athletes in the Top 40 in the world since 2015. Additionally, the complexity of the hurdles event plays a significant role in this underestimation. Hurdles is a highly specialized discipline, which results in a smaller pool of athletes compared to other events. This smaller dataset complicates the model's ability to estimate the distribution at the higher end, causing the Burr model to underestimate the performance density in the upper tail (such as world record performances). Consequently, elite performances in the hurdles are assigned fewer points than one might expect, as the model struggles to capture the true density of these top-level performances. This is reflected by the following figure, while the KDE captures the high concentration of elite performances in the tail, the Burr model seems to underestimate this density, leading to fewer points being assigned to top performances.

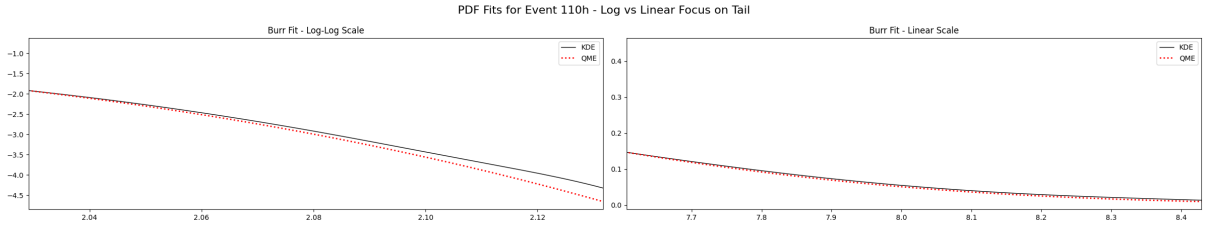


Figure 4.1: Fit of the 110m Hurdles' 5% Upper tail on a log-log and normal scale

On the other hand, I observe that the 400m world record (WR) is assigned a relatively high number of points. Given that the French top performance in the 400m is significantly below the world record, the Burr model seems to assign points appropriately, reflecting the overall distribution without inflating the points for performances at the world-record level. However, this phenomenon would require further verification by fitting the model to a larger dataset, ideally one with more data points closer to the world record. This would adjust the slope of the PDF, leading to a revision of the point allocation. As a result, the world-record performance might receive a lower number of points, but still within an appropriate range that accurately reflects its place in the distribution.

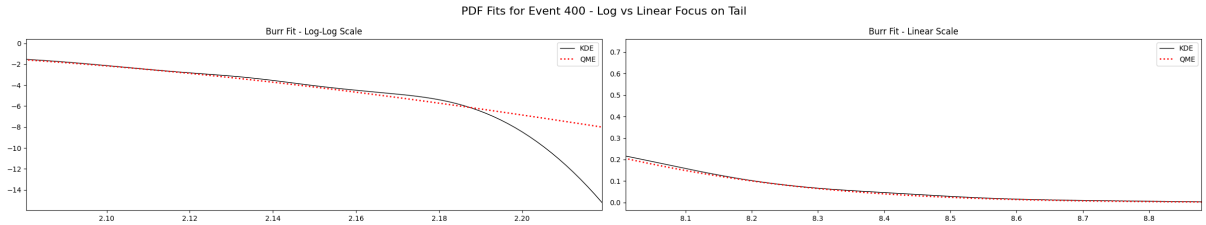


Figure 4.2: Fit of the 400m's 5% Upper tail on a log-log and linear scale

Using the French best performances, the scores assigned to each event are generally more consistent, with the exception of the 110m hurdles, where the scoring seems less aligned with expectations.

Event	Athlete	French Best (2015-2025)	Burr
100	OMANYALA Ferdinand	9.79 s	7.44
LJ	GOMIS Kafetien	826 cm	7.90
SP	BERTEMES Bob	21.36 m	8.21
HJ	GASCH Loic	233 cm	8.19
400	BIRON Gilles	45.05 s	7.79
110h	BADAMASSI Saguirou	13.05 s	6.30
DT	SOSNA Mika	68.96 m	8.01
PV	LAVILLENIE Renaud	605 cm	7.63
JT	CHOPRA Neeraj	87.58 m	8.20
1500	HABZ Azeddine	3:29.26	8.26
Std (across events)			0.56

Table 4.2: French Best assessment: Comparison of Burr vs. Log-Normal point attributions

To improve the fit of the model, it would be valuable to expand the dataset by including a broader range of performance data from around the world. Specifically, adding more data from athletes with top-tier performances, would help the model more accurately capture the behavior of extreme performances. This would refine the tail estimation and results in a more robust and representative model that better reflects the distribution of world-class performances.

4.2 Analysis of Olympic performance data

In this section, I will dive into the analysis of the new scoring system within the decathlon framework, which I have developed and applied to the performance data from the last three Olympic Games. The goal of this analysis is to explore how the new point attribution system might have impacted the rankings of athletes if it had been used instead of the traditional one.

To assess the potential impact of this new scoring system, I have gathered the performance data and corresponding scores from the last three Olympic Games. Using this data, I will perform several analyses, including Principal Component Analysis and correlation analysis of the athlete rankings under both the old and new scoring systems.

First, I need to mention that Olympic decathletes are the ideal target for such a scoring system. These athletes are highly versatile and typically perform above average in most events. However, they do not perform at the same level as Olympic athletes who specialize in individual events. For most events, I expect the decathletes to score within a four- to five-point range, with occasional exceptional performances. On the other hand, for the 1500m, I expect the score to be much lower, as this event is quite different from the others in the decathlon. While the previous events require explosiveness, power, strength, and raw speed, the 1500m demands more endurance. Unfortunately, speed and endurance do not complement each other very well; training for one often comes at the expense of the other. As a result, decathletes tend to prioritize the first 9 events, which means the 1500m is likely to yield a relatively lower score. This could potentially lead to a further refinement of the scores formula for the 1500m.

- Correlation of Rankings
- Principal Component Analysis (PCA)

Through these analyses, I can gain insights into how the new scoring system might have impacted the outcome of the decathlon events at the Olympics, offering a fresh perspective on athlete performance and potentially leading to a fairer or more representative ranking system.

4.2.1 Correlation analysis

I will first calculate the correlation between the athlete rankings under the old scoring system and the new scoring system. This will allow me to quantify how much the rankings change with the introduction of the new point attribution system.

The first thing that comes to mind after setting up the new system is to compare the correlation between the old ranking and the new ranking implied by the new scores. To achieve this, Spearman's rank correlation is ideal, as it allows to assess how well the relationship between two variables can be described using a monotonic function. In other words, I compute the correlation between the ranks of the old and new rankings.

Mathematically, Spearman's rank correlation coefficient ρ is defined as (McAleer, 2022):

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

where:

- d_i is the difference between the ranks of the corresponding values in the old and new rankings,
- n is the number of items being ranked.

A high correlation would naturally suggest that the new system doesn't drastically alter the relative rankings of athletes, while a low correlation might indicate that the new system offers a significantly different perspective on athlete performance.

Merged Olympics					
Athlete	**Old Points**	**Old Rank**	**New points**	**New Rank**	**Delta Rank**
Damian Warner	9018	1	47.63	1	0
Ashton Eaton	8893	2	46.04	2	0
Ashley Moloney	8649	9	44.03	3	-6
Markus Rooth	8796	4	43.98	4	0
Leo Neugebauer	8748	5	43.82	5	0
...	...	...	...	...	...
Sven Roosen	8607	11	42.53	10	-1
...	...	...	...	...	...
ρ_{rank}					0.977

Table 4.3: Recomputed ranking on the last 3 Olympics Games

The Spearman correlation between the old and new ranks is very high for the merged Ranking, $\rho = 0.977$. This suggests that, overall, the ranking in the new scoring system is highly consistent with the old one. Indeed, the introduction of a new point attribution system, the general ranking order of athletes remains mostly unchanged. Nevertheless some athletes like Ashley Moloney really benefits from the new ranking, gaining 6 spots

with the new system while Kevin Mayer is not beneficial from the new system, loosing 3 spots.

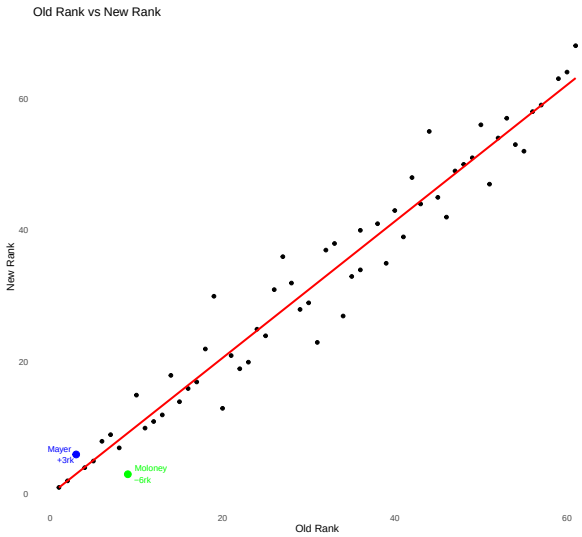


Figure 4.3: Rank correlation between old and new scoring system

Looking at the merged rankings, I can gain additional insights. For example, I compute the Spearman correlation between each event's rank and the overall rank under both the old and new scoring systems, as well as the correlation between each event's rank and the change in overall rank, Delta Rank (Rank(new)-Rank(Old)).

Event	Old Rank Correlation	New Rank Correlation	Delta Rank Correlation
100	0.491	0.608	0.628
LJ	0.723	0.736	0.286
SP	0.426	0.429	0.065
HJ	0.469	0.448	0.096
400	0.632	0.701	0.475
110h	0.574	0.659	0.530
DT	0.258	0.216	-0.125
PV	0.472	0.385	-0.250
JT	0.329	0.279	-0.095
1500	0.420	0.338	-0.282

Table 4.4: Spearman Correlations between Ranked Event Scores and Ranks

A positive correlation between the ranked individual event scores and the overall rankings indicates that as athletes perform better in specific events, they are likely to achieve better overall rankings. The stronger the correlation, the more closely an event's performance aligns with the global ranking. It is expected that there should be no negative correlation between the old and new rankings, as it would be unrealistic for strong performance in an event to result in a lower overall rank.

For both the old ranking and the new ranking correlation, I observe that strong performances in the Long Jump or 400m lead to higher rankings, as seen in the case of Ashton Eaton or Damian Warner, whose excellent performance in these events gave them an advantage in the old ranking system. Indeed as they perform as well as the top performer in those events these events bring them major increment in terms of points

and ultimately leads for both ranking to a higher final ranking. Conversely, both the old and new rankings show that the Javelin and Discus have the weakest correlation. This suggests that performance in those events has a moderate influence on an athlete's rank, whether under the old or new system.

The correlation between delta rank and event's rank represents how performance in an event impacts the change in an athlete's overall ranking between the old and new scoring systems. A positive correlation indicates that athletes who improve in specific events are likely to see a greater improvement in their overall rank. For example, improvements in the 100m are strongly associated with a better rank change, as indicated by the correlation of 0.628. This suggests that athletes who perform well in this event would likely gain in terms of ranking with a new system. This is also the case with the 400m and the 110m Hurldes. This suggests that the new ranking favors athletes that are among the best individual sprinters. On the other hand, a negative correlation means that improvements in certain events may not lead to a better overall rank. For instance, both the 1500m and Pole Vault (PV) events show negative correlations of -0.282 and -0.250, respectively. This indicates that even if an athlete improves in these events, it may not translate into an overall ranking improvement. The lower correlation for these events could be due to the endurance-based nature of the 1500m and the specialized nature of the Pole Vault, where decathletes often focus more on sprinting and explosive events. Additionally, some events like the High Jump or Shot Put and Javelin show a weaker correlation with delta rank, suggesting that improvements in this event have a less significant effect on the overall ranking change compared to events like the 400m or 100m, which are more directly linked to rank changes in both systems.

In the Figure 4.3, I highlighted two athletes, Ashley Moloney, in green, who benefited from the new system and Kevin Mayer, in blue, who did not.

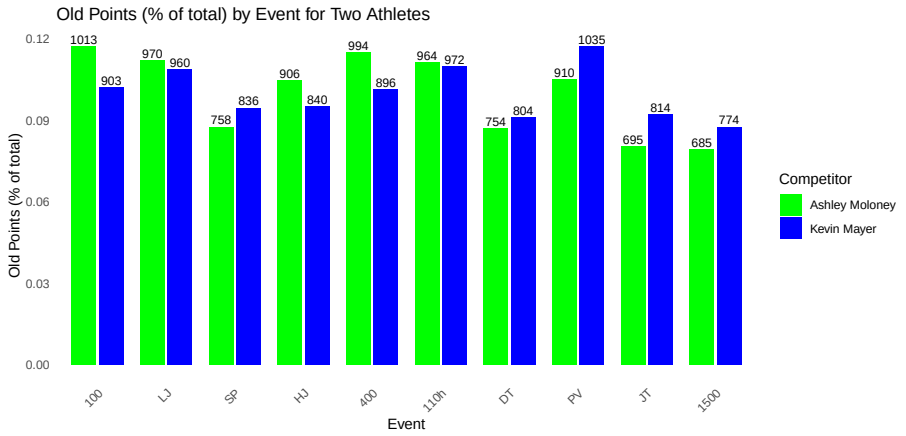


Figure 4.4: Score repartition across all ten events under the old scoring system

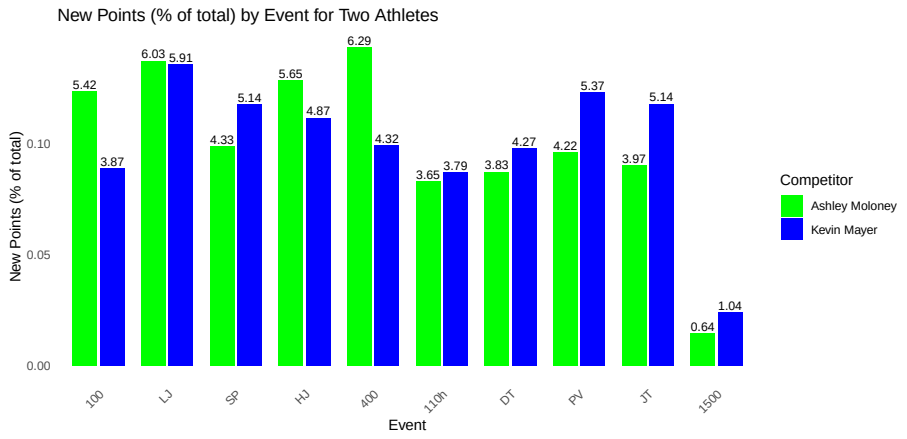


Figure 4.5: Score repartition across all ten events under the new scoring system

Using the two previous figures, several key insights emerge. Under the old scoring system, Moloney's 100 m and Mayer's pole-vault performances were worth the same number of points. In the revised table, however, Mayer's pole-vault score drops relative to Moloney's 100 m, giving Moloney a clear advantage in power events versus Mayer's pure speed. I also observe that Moloney's 400 m performance now carries significantly more weight, further boosting his total. Considering just the 100 m and the javelin throw: under the old rules Moloney's 100 m lead is fully erased by Mayer's javelin, leaving Mayer ahead by 9 points. Under the new scoring, Moloney gains 1.55 points in the 100 m but Mayer only recovers 1.17 points in the javelin, so Moloney now leads by 0.38 points across those two events. The same illustration could be done with the Pole Vault. This perfectly illustrates how the re-weighted table favors the 100m and the 400m and reduces Mayer's Pole Vault and Javelin advantage, explaining the switch in their overall rankings.

4.2.2 Principal component analysis

To better understand the changes brought by the new scoring system, I performed a Principal Component Analysis (PCA) on both the old and new scoring systems. By doing so, this approach identifies the main factors that explain the differences in event performances across athletes for both old and new ranking system. For that my data set is powered with the scores attributed to each athletes both for old and new systems.

First, I will not be scaling the variables to preserve the original variance structure and the rewards given by the new system.

Under the IAAF scoring system, the first two principal components explain a combined 48.5% of the total variance, with PC1 alone accounting for 26.2%. In contrast, the new scoring system shows a more concentrated variance structure: PC1 explains 37.6% of the variance and PC1+PC2 together account for 55.6%. This suggests that the new scoring system produces a more dominant first axis of variation, indicating stronger athlete separation in fewer dimensions. This is consistent with the tail-aware design of the new scores, which emphasize exceptional performances and create greater spread among top

athletes.

Now, to examine how each event contributes to the main dimensions of variation in athlete scores, the variable loadings on the first two principal components from the PCA are insightful. Under the old scoring system, the 1st component shows high positive loadings from Long Jump, 400m, and 110m Hurdles, indicating that these events heavily influence the main axis of score variation. PC2 is, on the other hand dominated by large positive loadings from the throwing events, suggesting that these strength-oriented field events form a distinct secondary dimension in athlete profiles. In the new scoring system, PC1 is most strongly aligned with 400m, LJ, and 100m, all of which exhibit large positive loadings. This reflects a shift in emphasis toward short sprints and explosive events. PC2 is again dominated by the JT. These differences in event loadings reveal that the new scoring system alters the underlying geometry of performance space, reinforcing the impact of elite-level sprinting events, particularly on the first axis, while events with a small contribution to that axis are less impactful for the final ranking .

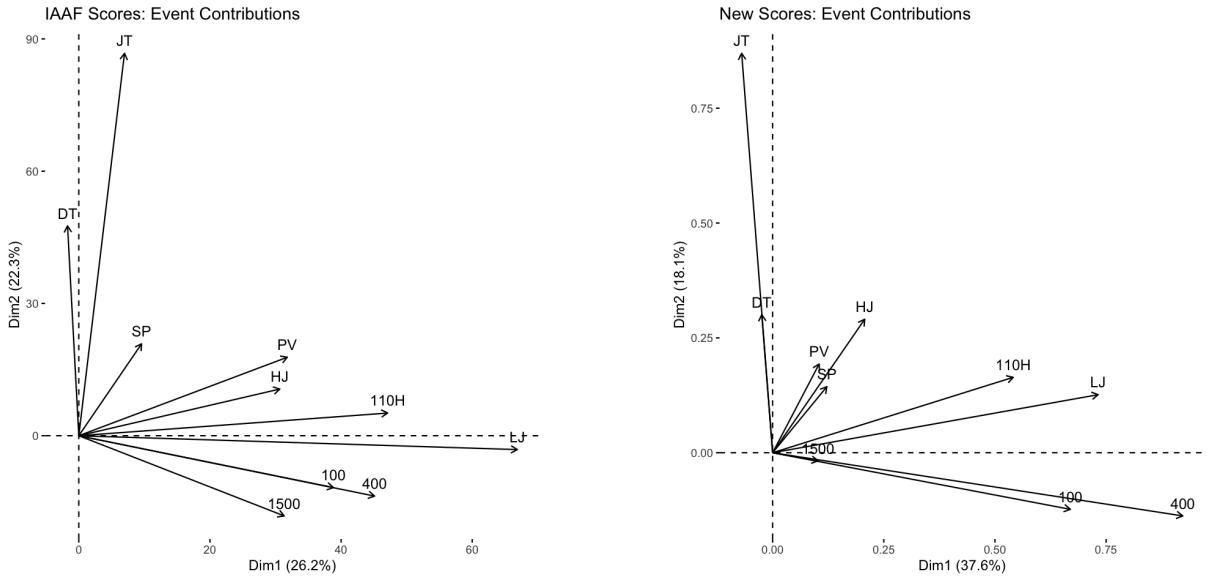


Figure 4.6: Variable plot on the first two Principal components

In both systems, the first principal component aligns strongly with athlete ranking: top-ranked athletes are located on the left end of the axis. The PC1 effectively captures global performance level.

Under the IAAF (old) system, the distribution of athletes is broader and less structured along PC2, with top athletes more dispersed. In contrast, the new scoring system exhibits a tighter grouping of elite performers such as Warner, Neugebauer, and Eaton. This pattern reflects the sensitive nature to the tail of the new scoring function, which amplifies differences among exceptional performances.

From an individual perspective, Damian Warner benefits notably from events like the

Long Jump and 110m Hurdles, where he ranks among the world's best. These events also exhibit stronger influence in the PCA space, contributing to his elevated position. Conversely, athletes like Kevin Mayer, who previously gained an advantage through events like Pole Vault under the IAAF model, appear less favored in the new framework.

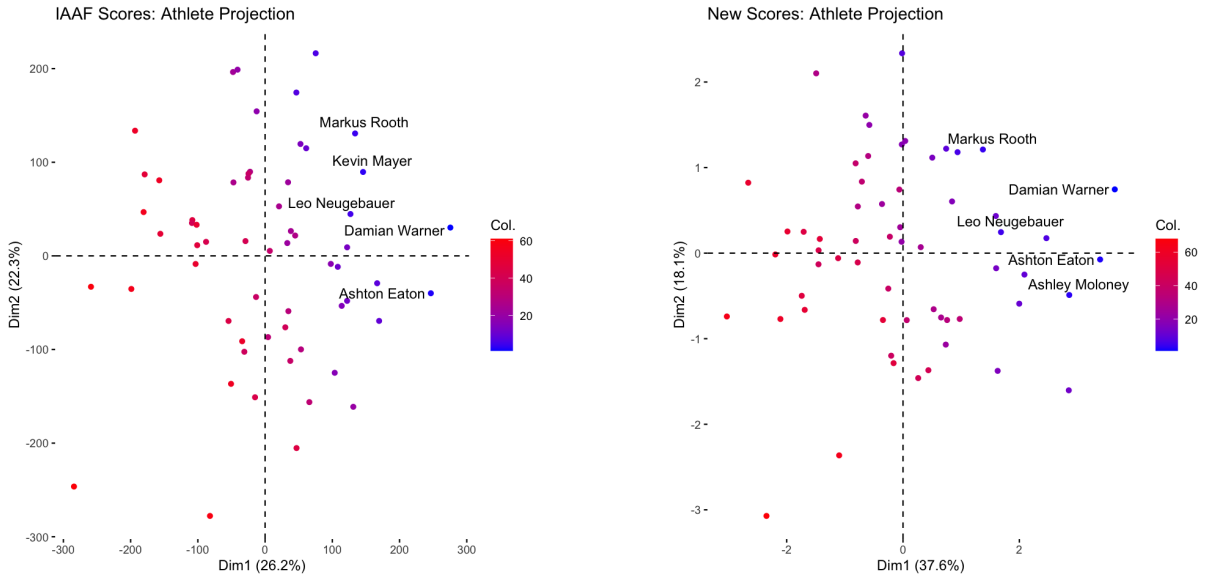


Figure 4.7: Individuals plot on the first two Principal components (Top 5 are named). The darker the blue, the better the individual is ranked

Finally to understand how the principal components of the decathlon scores relate to the final athlete rankings, I regressed both the IAAF and new scoring system rankings on the first two principal components from PCA. The results are summarized below.

Metric	**IAAF (Old)**	**New System**
Abs. Correlation (PC1 + PC2)	0.948	0.955
R^2 (PC1 + PC2)	0.899	0.911
Adjusted R^2 (PC1 + PC2)	0.896	0.908
Residual Std. Error	5.66	5.64

Table 4.5: Regression of athlete ranks on PC1 and PC2

Both scoring systems show a strong relationship between the two principal components and the final ranking. In particular, over 89% of the variation in rank is explained by just the first two principal components under both systems. Notably, the new system performs slightly better, with an R^2 of 0.911 compared to 0.899 for the IAAF scores.

The PCA analysis confirms that the new scoring system improves the structural alignment between athlete scores and final rankings. By amplifying the contribution of elite-level performances, especially in sprint and explosive events, the new system creates a more dominant and interpretable first principal component. Overall, PCA highlights how the new scoring framework re-balances event influence.

Chapter 5

Conclusion

The question of who the best athlete between Usain Bolt and Javier Sotomayor remains subjective in more than one point, technicality in the event, physiology, new training techniques and many more points deserve to be examined. Nevertheless, this thesis has demonstrated that creating and refining a scoring system based on statistics and distribution laws may provide a fair, comparable and objective metric to assess the value of performances in different disciplines.

After examining a decade of French track-and-field results, I demonstrated the limitations of the current model whose shifted-Weibull assumptions misaligned with real-world data. Through systematic comparison of six alternative distributions, I revealed that the Burr XII family emerged as particularly adept at capturing tail behaviors while still mitigating central tendencies. With those Burr XII fitted functions, I created tables using a new performance-score formula (2.5). This new system better rewards exceptional efforts and reduce arbitrariness in cross-event comparisons, all while maintaining strong correlation ($\rho_s = 0.977$ with existing decathlon rankings.

Even with these advantages, there are several considerations that deserve notice. The assessment relies mainly on performances from the French federation and the Olympics. Worldwide datasets might show varying statistical characteristics. As I mentioned in Section 4.1. It would be ideal to enhance our dataset with performance from different countries in order to avoid under-representation of some performance in the tail. (e.g. taking Jamaican data for sprint would lead to a smaller point allocation to the World record of Usain Bolt as it would make his record more probable).

Also, further improvement can be made in the pipeline itself. For example, beyond the examined distributions, there are other candidate models that can be examined. Also, in the quantile-matching method, experimenting with different sets of quantiles. Finally, within the Bayesian framework, different priors could influence the posterior distribution and help regularize estimates.

Looking ahead, the same methodology, possibly refined with improvements, could be applied to other sports such as weightlifting, swimming, indoor cycling, or virtually, provided some minor changes and sport knowledge, any sport where measurable and standardized outcomes exist, enabling meaningful cross-sport comparisons.

Bibliography

S. Cheng and L. Peng. Confidence intervals for the tail index. *Bernoulli*, 7(5):751–760, 2001. doi: 10.2307/3318540.

I. Danjuma, A. Yahaya, and O. E. Asiribo. Fitting and comparing gamma, lognormal and weibull distributions using nigeria rainfall intensity data. *Quest Journals: Journal of Research in Applied Mathematics*, 6(5):18–24, 2020. ISSN 2394-0743. ISSN (Print): 2394-0735.

J. Diebolt, J. Garrido, and A. Girard. A goodness-of-fit test for the distribution tail. *Journal of Statistical Theory and Practice*, 17(3):345–368, 2023. doi: 10.1080/15326349.2023.2183964.

A. B. Downey. *Probably overthinking it: A book of questions*. University of Chicago Press, 2023. ISBN 978-0226822583. See Chapter 4, "Running speeds".

D. Feruzbek. Estimating the tail index of conditional distribution of asset returns. *International Journal of Financial Research*, 13(2), 2022. doi: 10.5430/ijfr.v13n2p14.

E. García. *Notes for nonparametric statistics*. Universidad Carlos III de Madrid, Madrid, 2023. ISBN 978-84-09-29537-1.

B. Grammaticos. The physical basis of scoring the athletic performance. *New Studies in Athletics*, 22(3):47, 2007.

B. Grammaticos, J. Meloun, and J. G. Purdy. Distribution of performances and scoring in athletics. *Mathematics and Sports*, 3:1, 2022.

C. M. Hafner, O. B. Linton, and L. Wang. Supplementary material for "dynamic autoregressive liquidity (darliq)". Technical report, Louvain Institute of Data Analysis and Modelling in Economics and Statistics (LIDAM), UCLouvain and Faculty of Economics, University of Cambridge, April 2025.

D. Harder. *Sports comparisons: You can compare apples to oranges*. Education Plus, Castro Valley, CA, 2001.

R. Harman. A brief introduction to the gamma distribution. `http://www.iam.fmph.uniba.sk/ospm/Harman/gamma.pdf`, May 2022. Unpublished manuscript.

B. M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, September 1975. doi: 10.1214/aos/1176343247.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions. Vol. 1*. Wiley series in probability and mathematical statistics: Applied probability and statistics. John Wiley & Sons, New York, 2nd edition, 1994. ISBN 978-0-471-58495-7.

C. D. Lai. Weibull distribution. In *Generalized Weibull distributions*, SpringerBriefs in Statistics. Springer, Berlin, Heidelberg, 2014. doi: 10.1007/978-3-642-39106-4_1. URL `https://doi.org/10.1007/978-3-642-39106-4_1`.

J. Leonard. The value of athletic glory in ancient greece, 2016. URL `https://www.greece-is.com/athletic-glory-shadow-zeus/`.

P. McAleer. A handy workbook for research methods & statistics, 2022.

A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management*. Princeton University Press, Princeton, revised edition edition, 2015.

G. S. Mudholkar, D. K. Srivastava, and G. D. Kollia. A generalization of the weibull distribution with application to the analysis of survival data. *Journal of the American Statistical Association*, 91(436):1575–1583, 1996. doi: 10.1080/01621459.1996.10476725.

I. Pobočíková, Z. Sedliačková, and M. Michalková. Application of four probability distributions for wind speed modeling. *Procedia Engineering*, 192:713–718, 2017.

J. G. Purdy. Computer generated track and field scoring tables, in three parts: I. historical development. *Medicine and Science in Sports*, 6, 1974.

J. G. Purdy. Computer generated track and field scoring tables, in three parts: Ii. theoretical foundation and development of a model. *Medicine and Science in Sports*, 7, 1975.

J. G. Purdy. Computer generated track and field scoring tables, in three parts: Iii. model evaluation and analysis. *Medicine and Science in Sports*, 9, 1977.

N. Sgouropoulos, Q. Yao, and C. Yastremiz. Matching a distribution by matching quantiles estimation. *Journal of the American Statistical Association*, 110(510):742–759, 2015. doi: 10.1080/01621459.2014.929522.

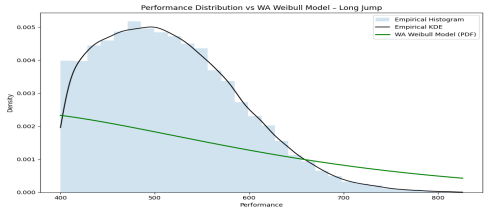
B. Spiriev. *IAAF scoring tables of athletics*, pages 11–18. World Athletics, Monaco, 2022.

P. R. Tadikamalla. A look at the burr and related distributions. *International Statistical Review / Revue Internationale de Statistique*, 48(3):337–344, Dec 1980. URL `https://www.jstor.org/stable/1402945`.

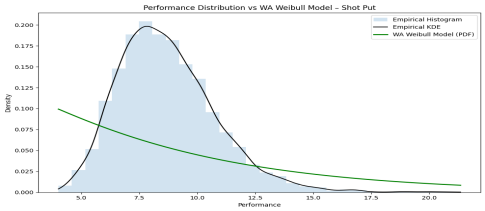
V. Trkal. The development of combined events scoring tables and implications for the training of decathletes. *New Studies in Athletics*, 18(4):7–14, 2003.

Additional Figures

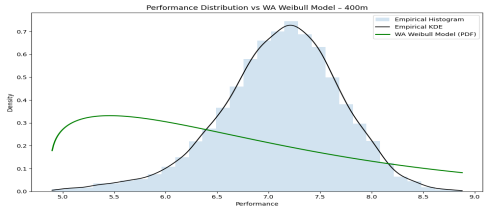
Visual fit of current Modeling



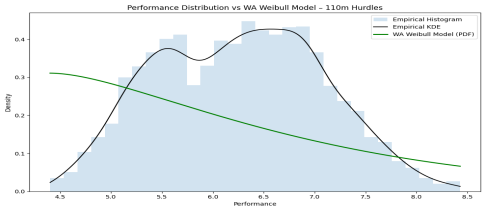
Long Jump



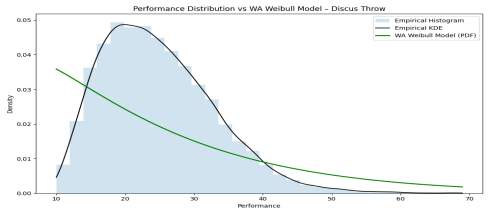
Shot Put



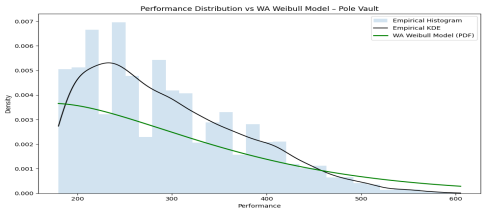
400m



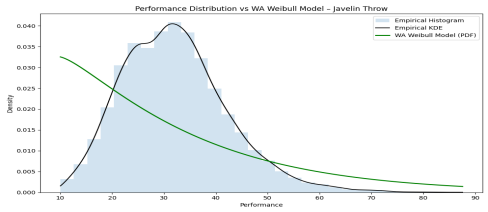
110mh



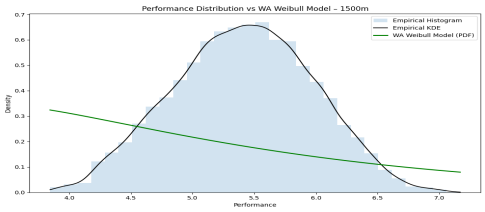
Discus Throw



Pole Vault



Javelin Throw



1500m

Figure 1: Comparison of event-wise KDEs vs fitted Weibull PDFs across 8 decathlon events.

Visual Fit of the new models

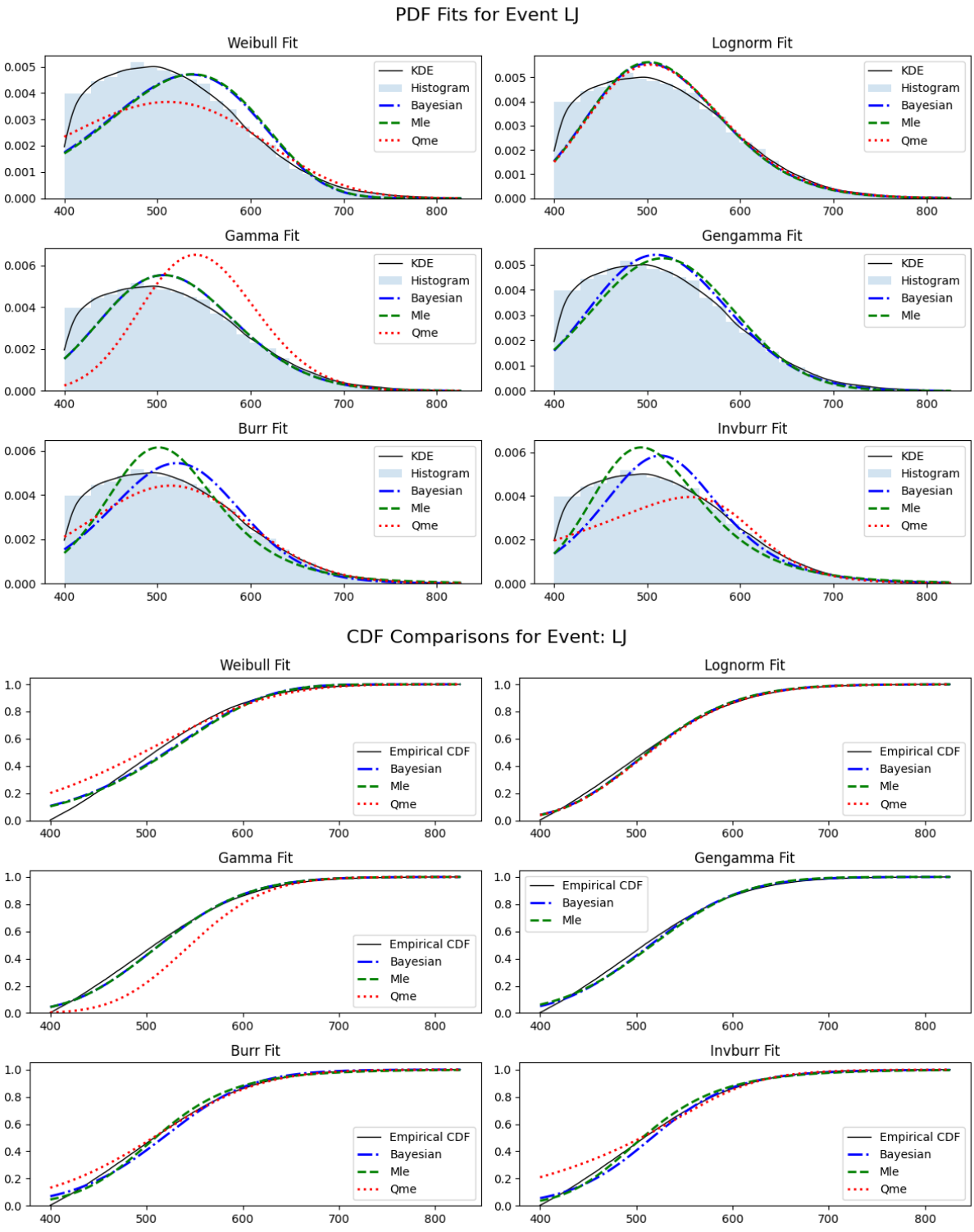


Figure 2: PDF and CDF of Long Jump

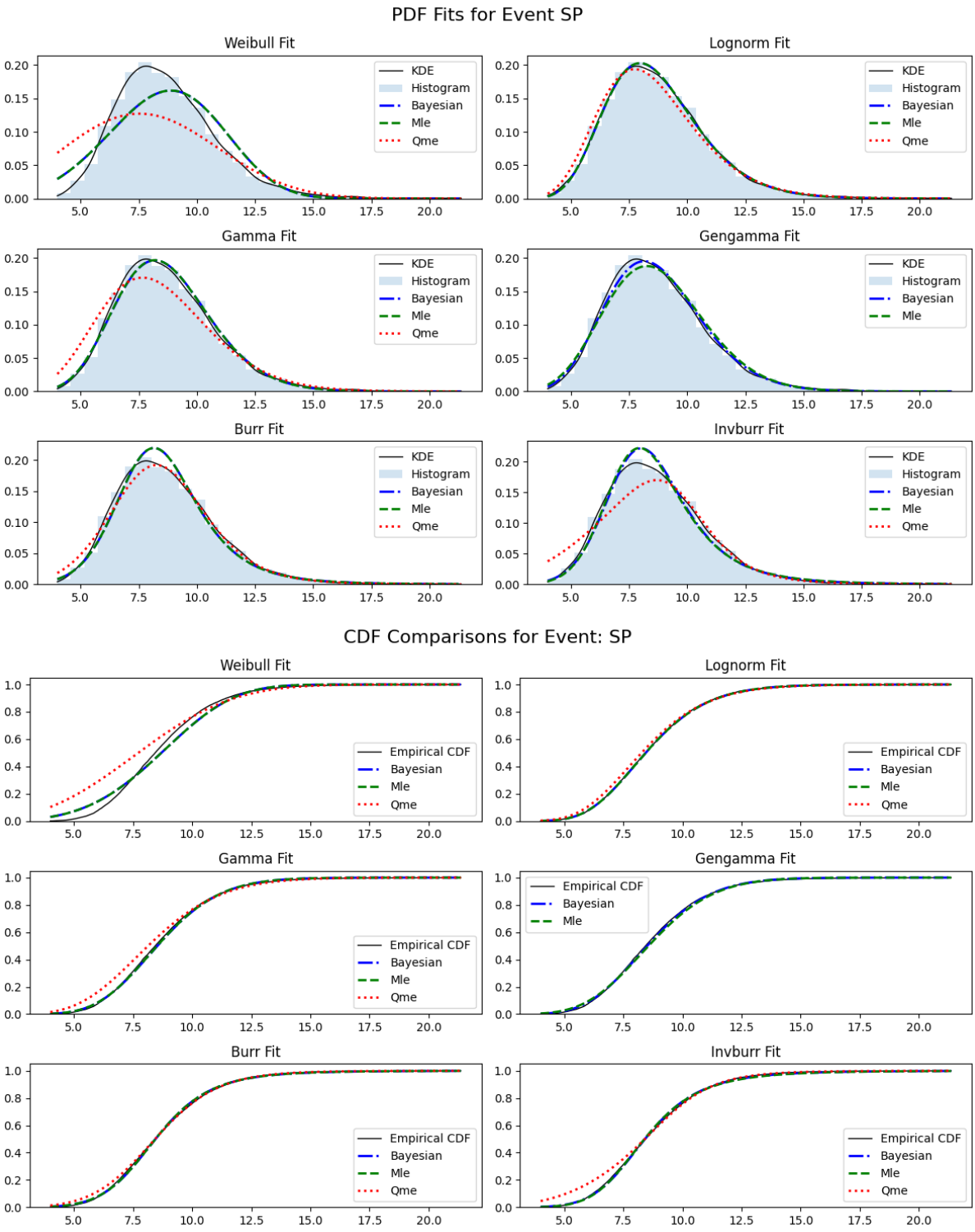


Figure 3: PDF and CDF of Shot Put

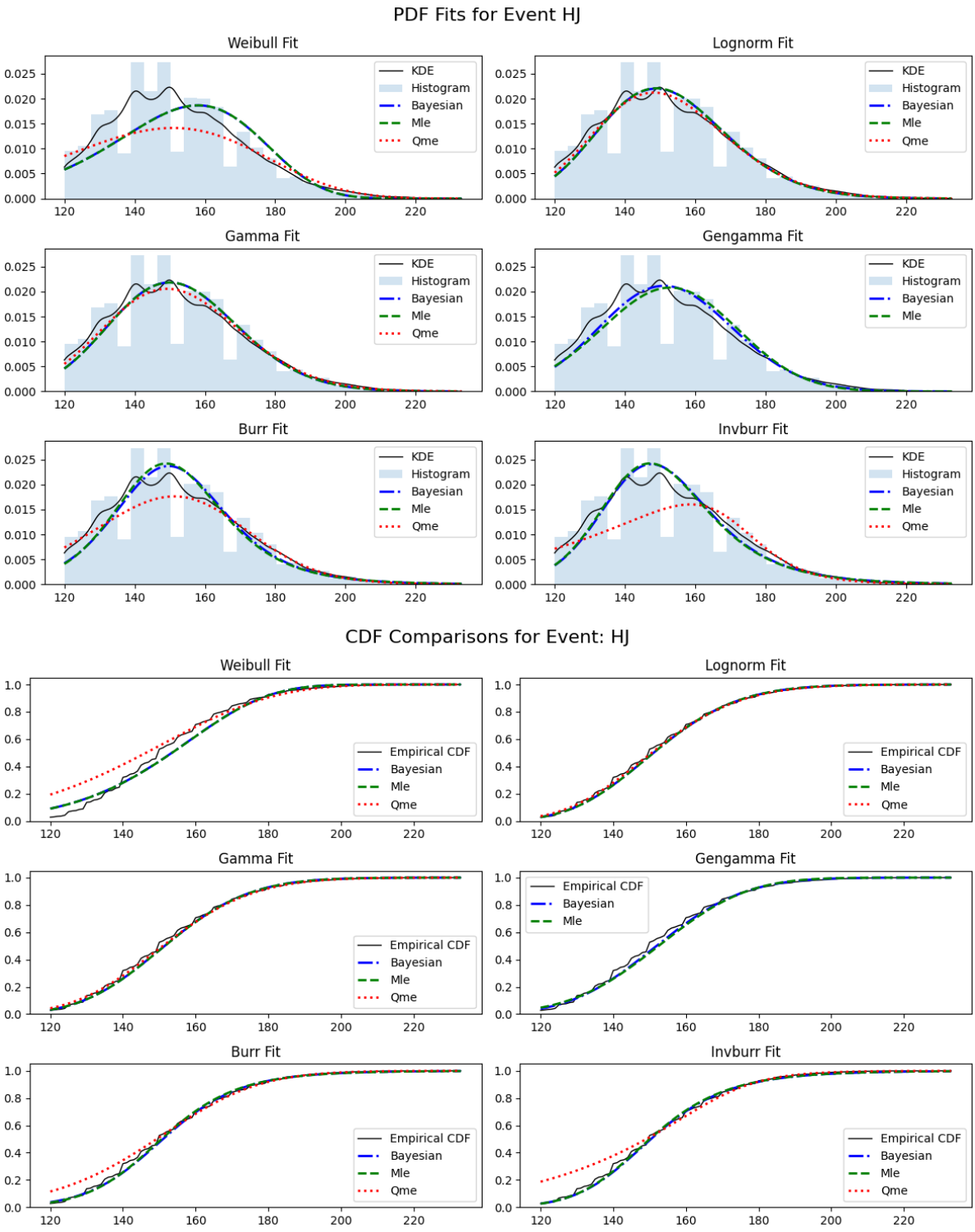


Figure 4: PDF and CDF of High Jump

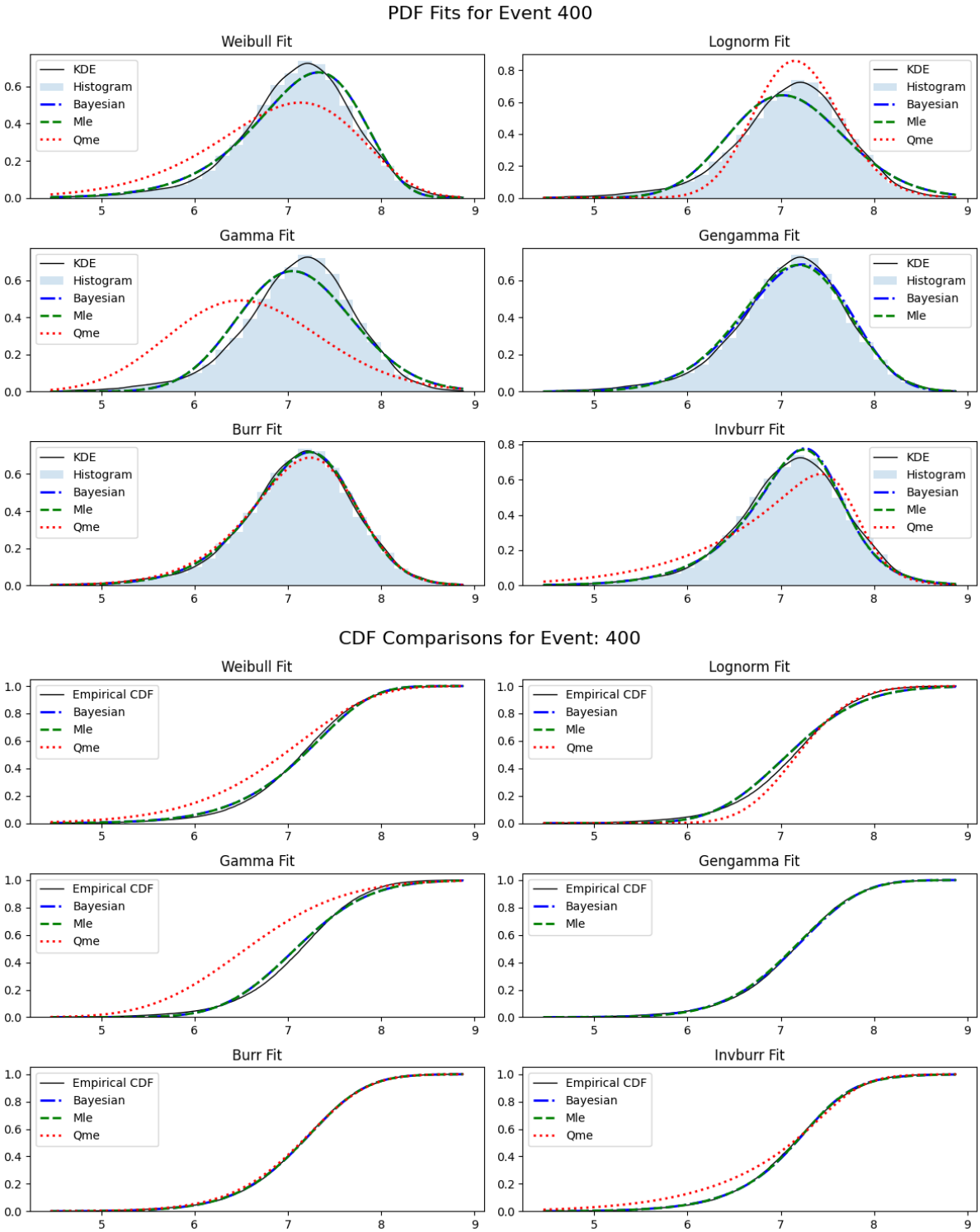


Figure 5: PDF and CDF of 400

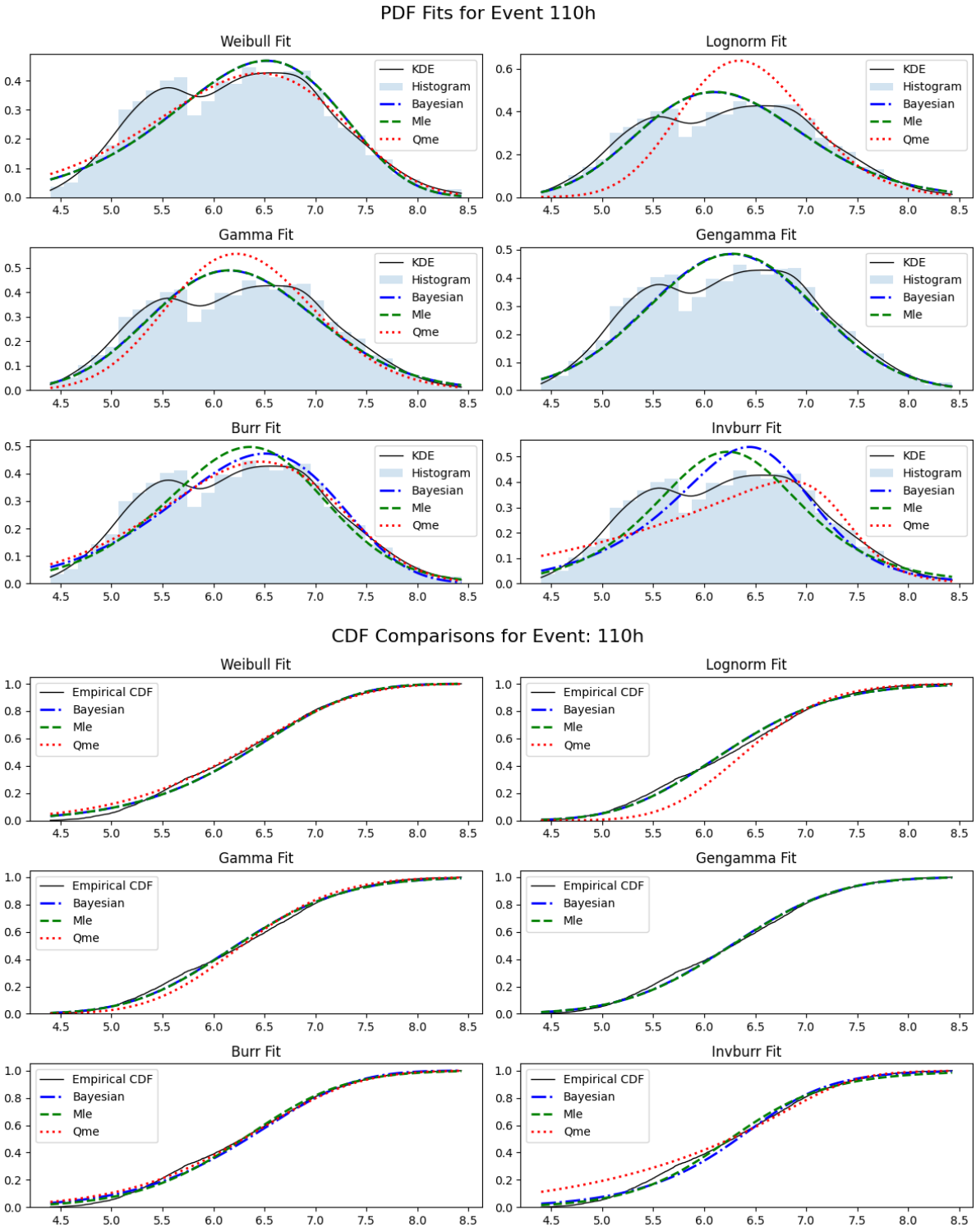


Figure 6: PDF and CDF of 110m Hurdles

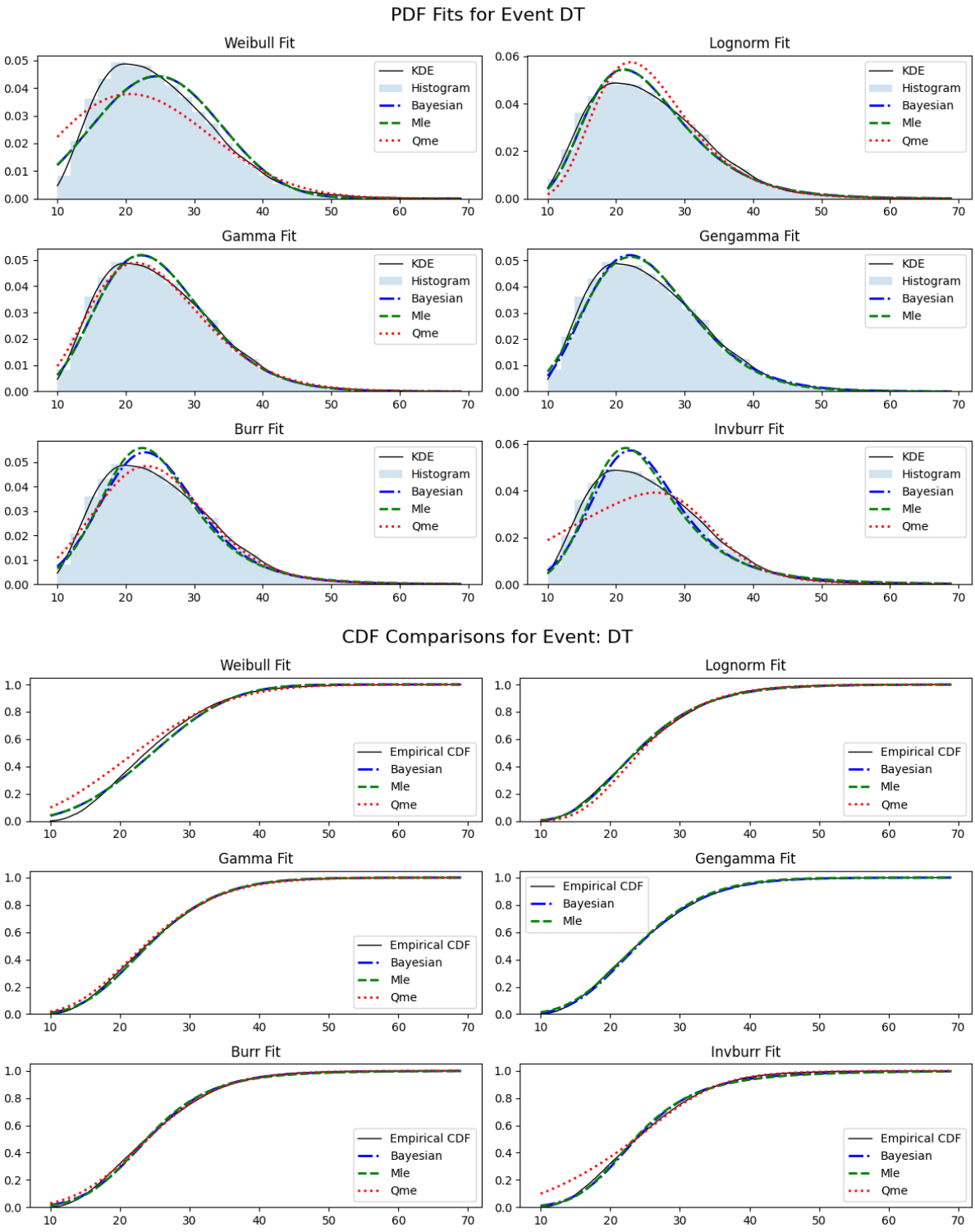


Figure 7: PDF and CDF of Discus Throw

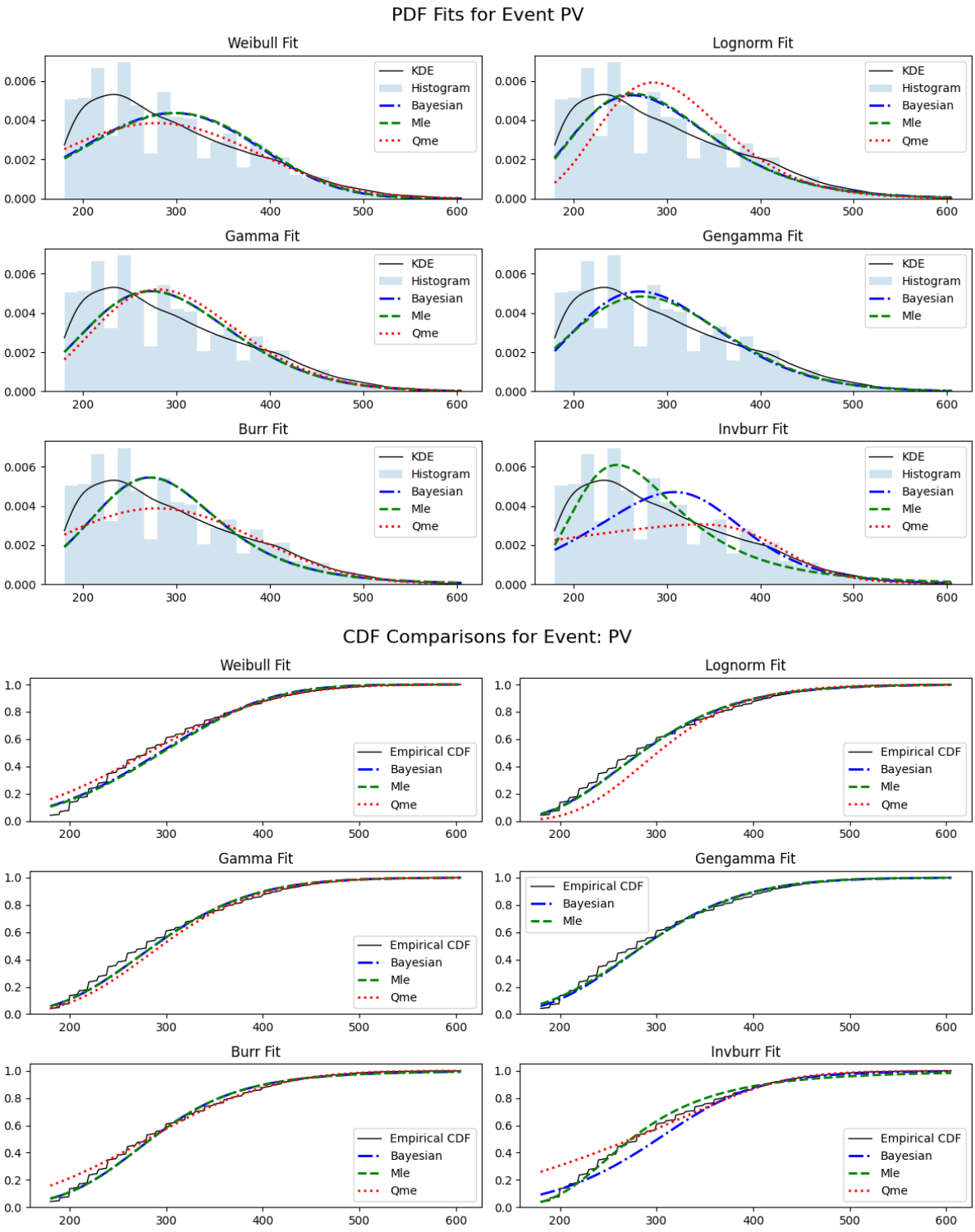


Figure 8: PDF and CDF of Pole Vault

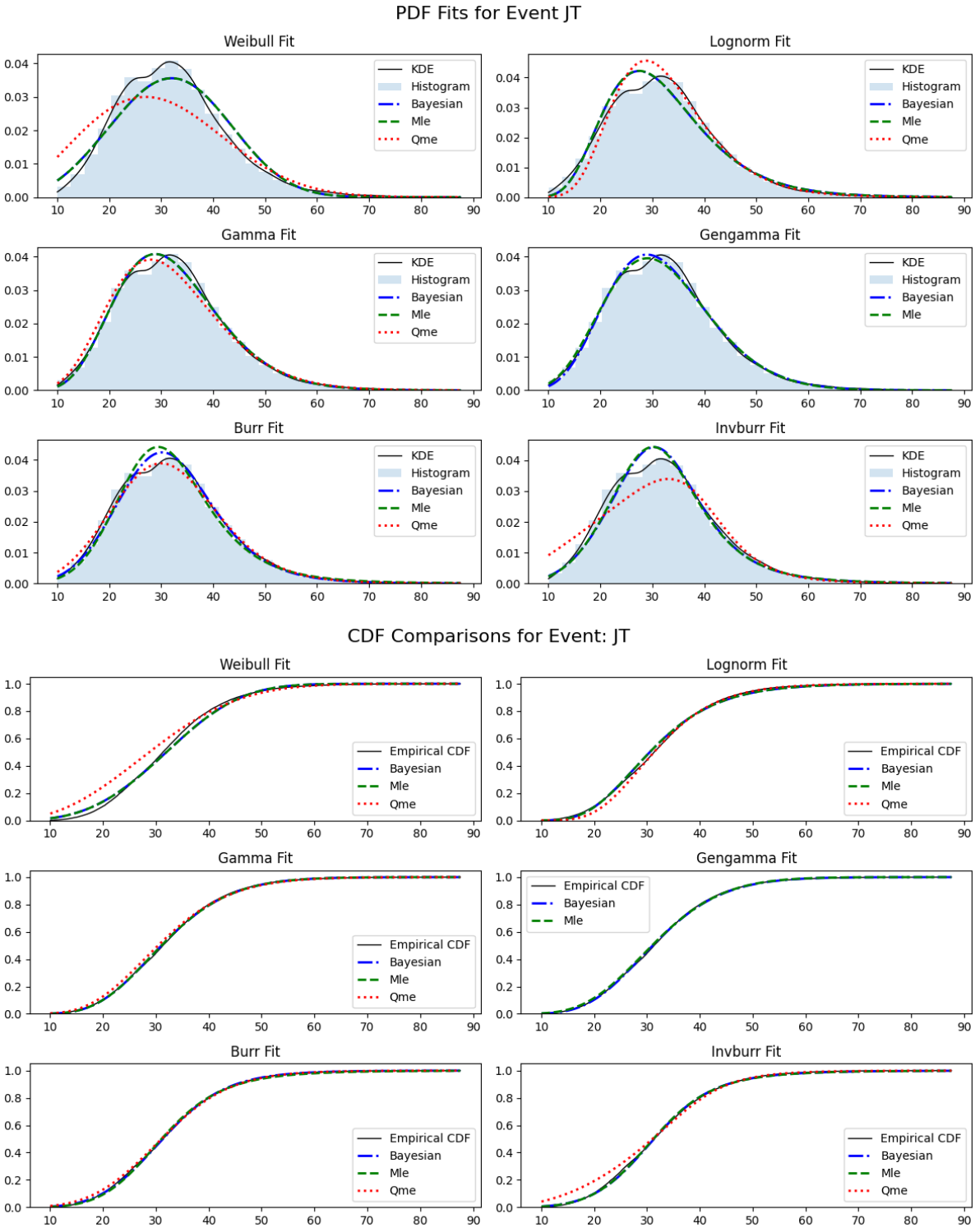


Figure 9: PDF and CDF of Javelin Throw

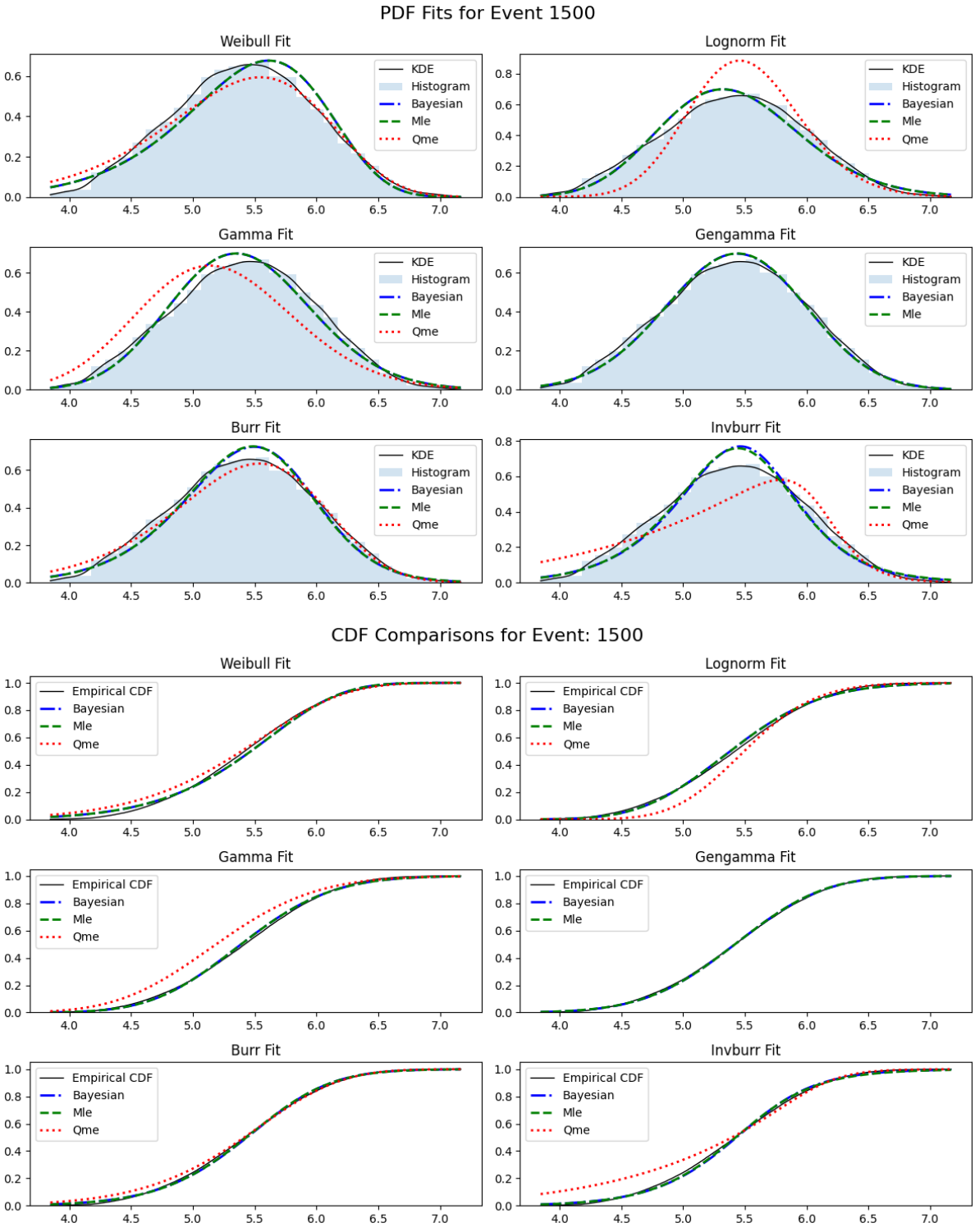


Figure 10: PDF and CDF of 1500m

PP-plot the new models

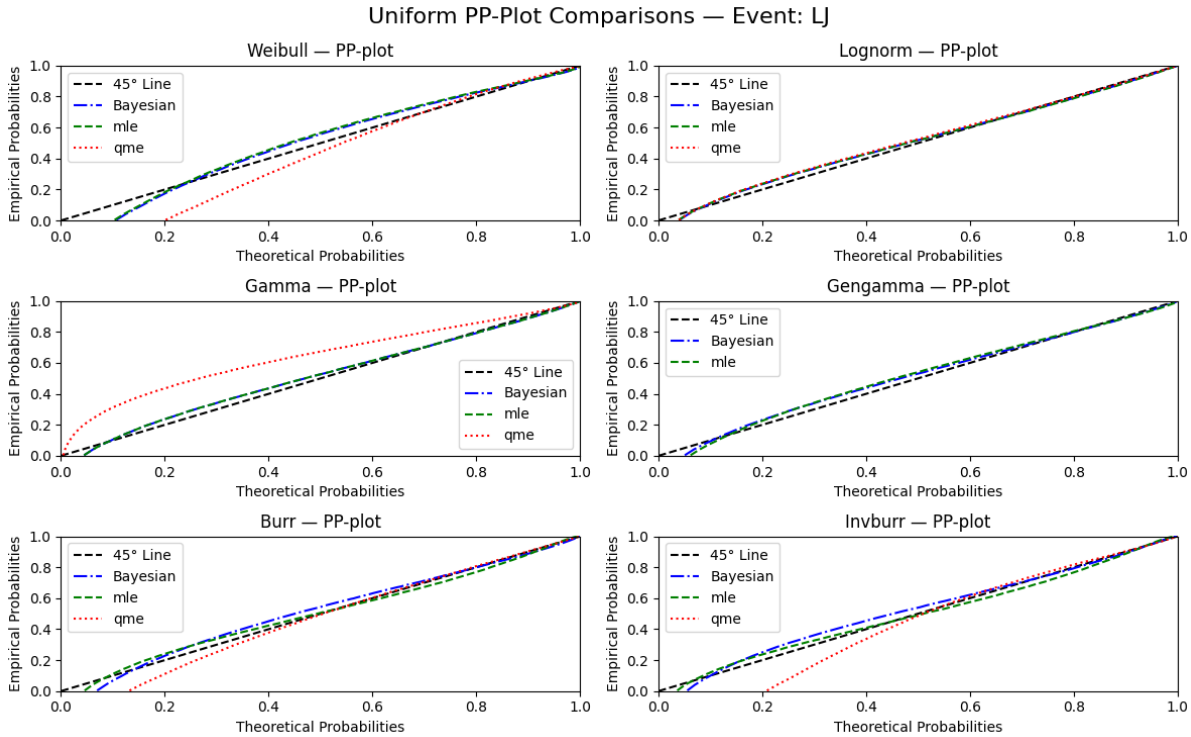


Figure 11: PP-Plot of Long Jump

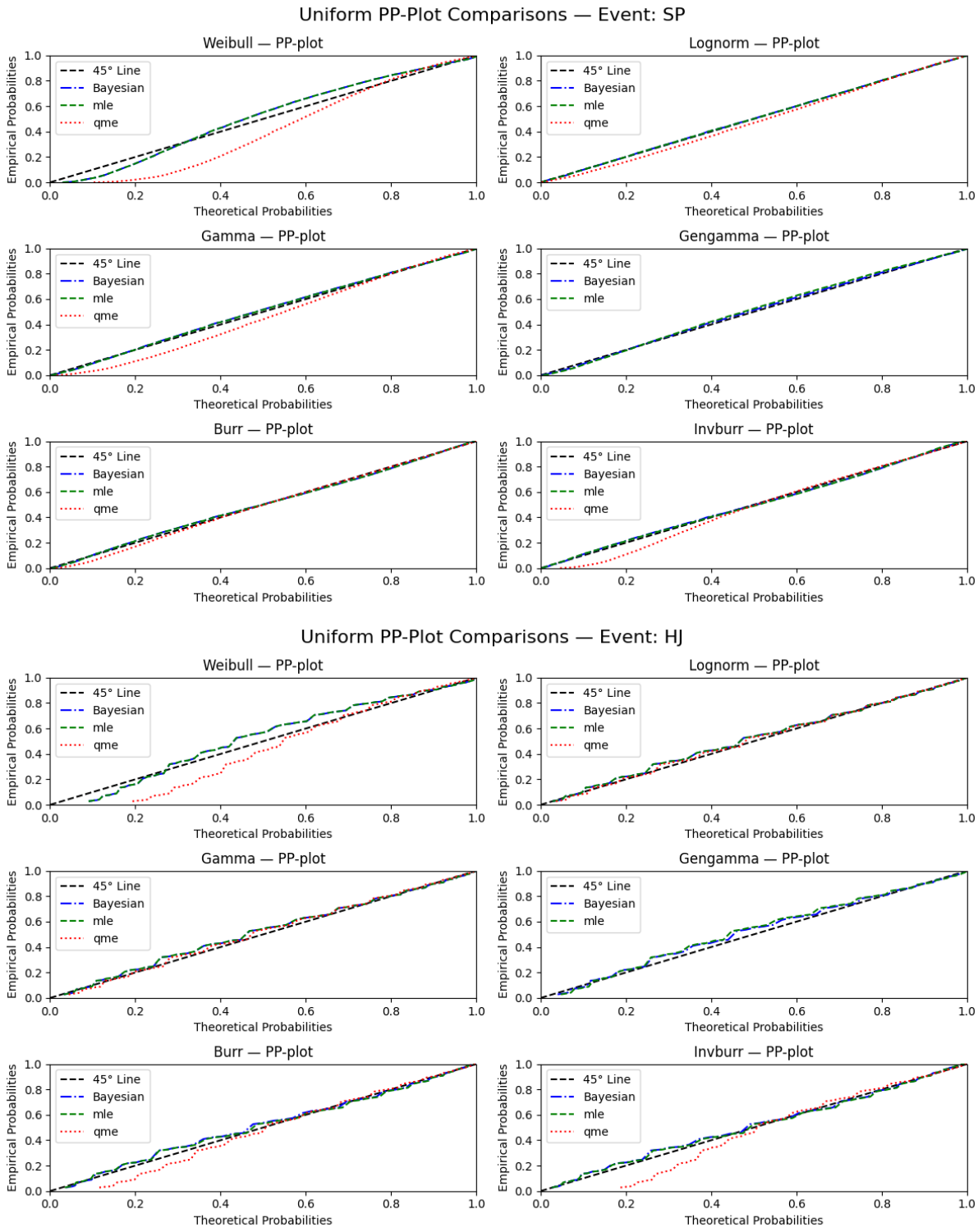


Figure 12: PP-Plot of Shot Put and High Jump

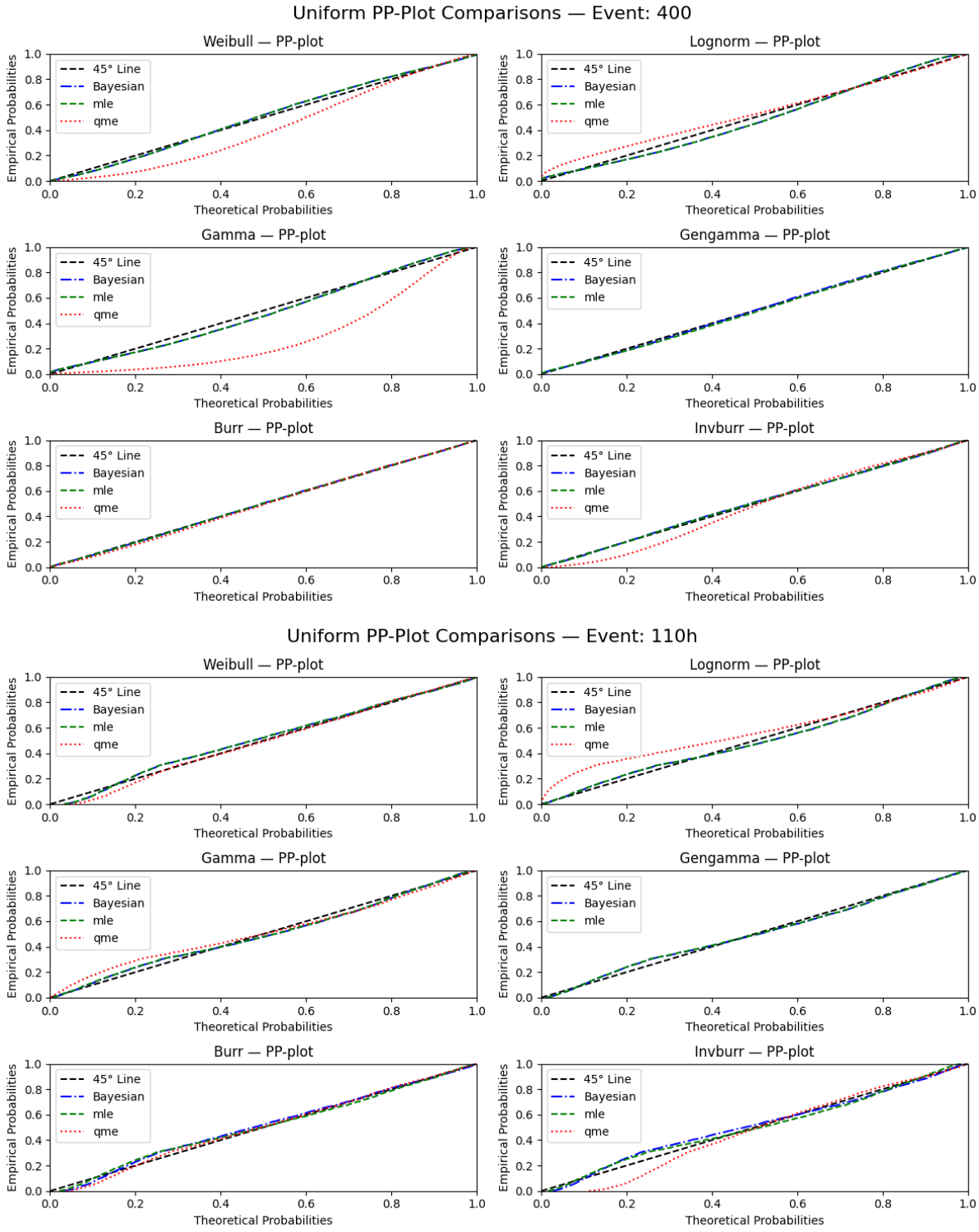


Figure 13: PP-Plot of 400m and 110 m Hurdles

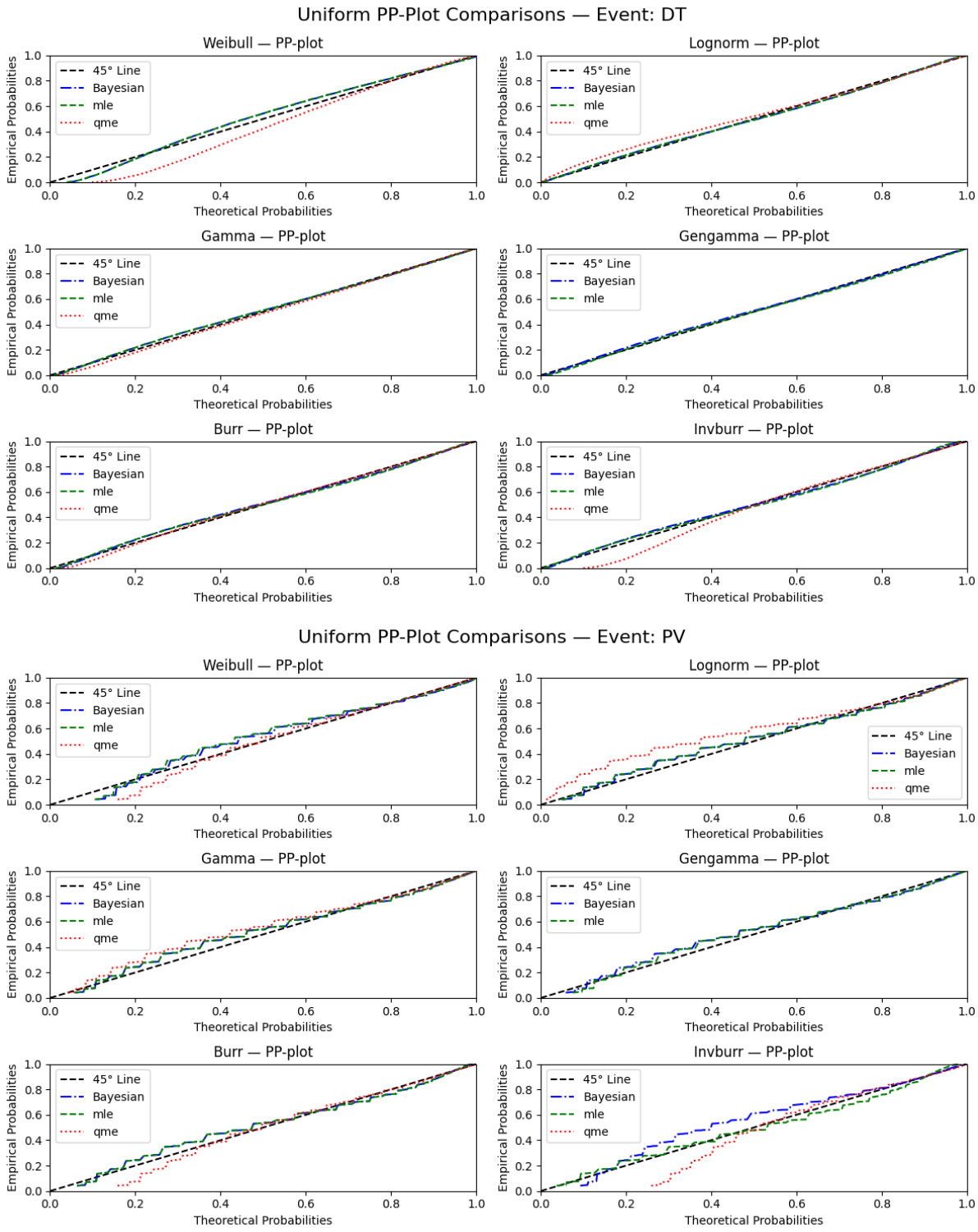


Figure 14: PP-Plot of Discus Throw and Pole Vault

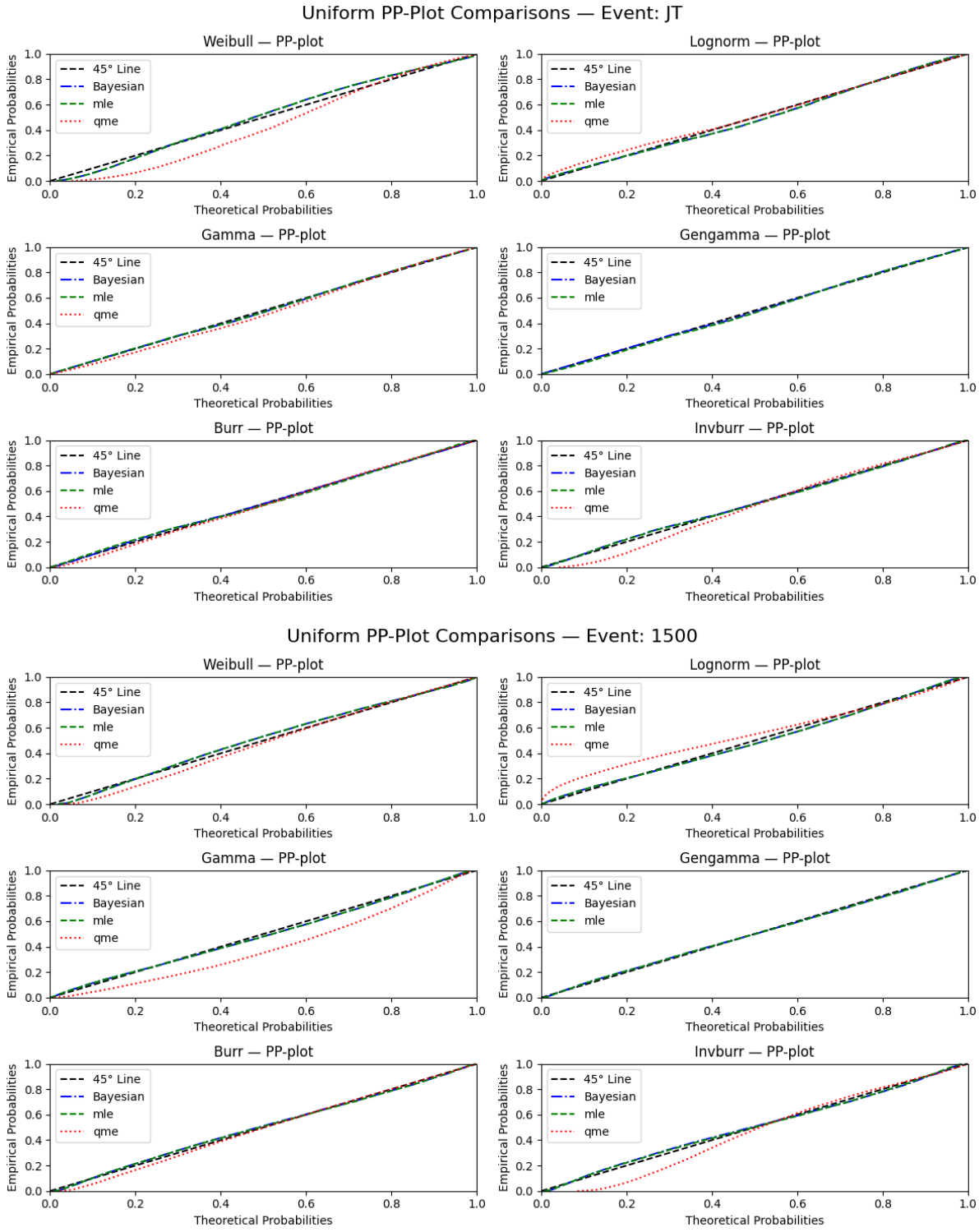
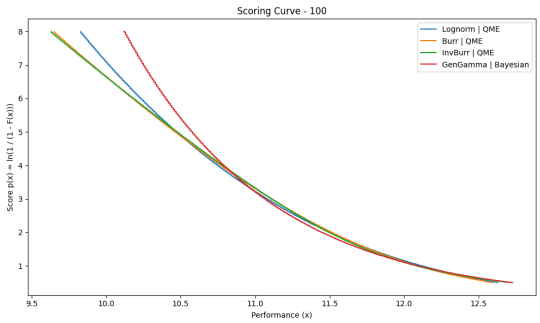


Figure 15: PP-Plot of Javelin Throw and 1500m

Scores for all new models



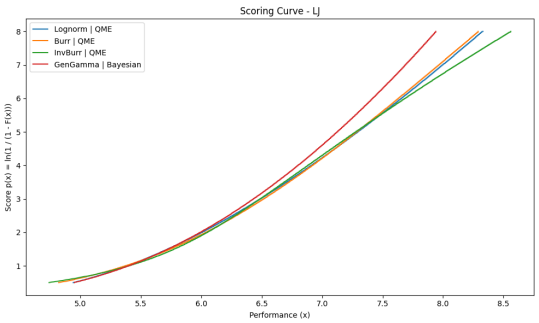
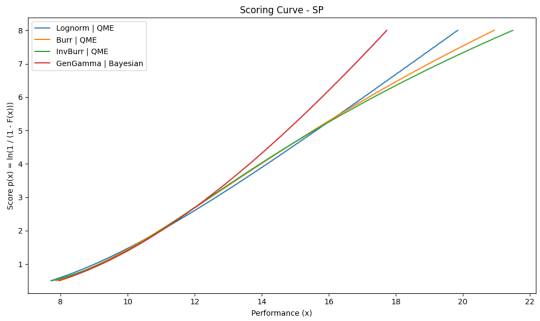


Figure 16: Scores of 100m and Long Jump



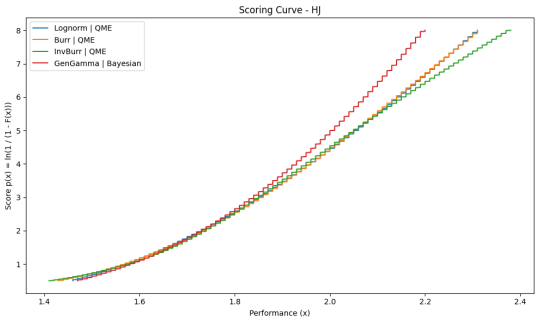
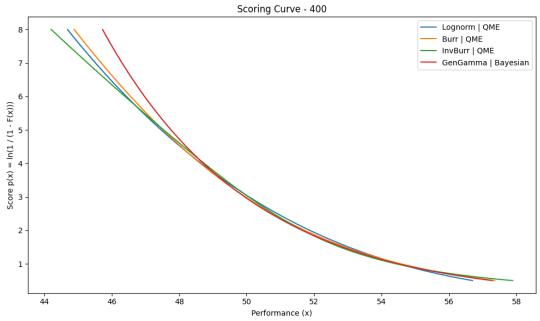


Figure 17: Scores of Shot Put and High Jump



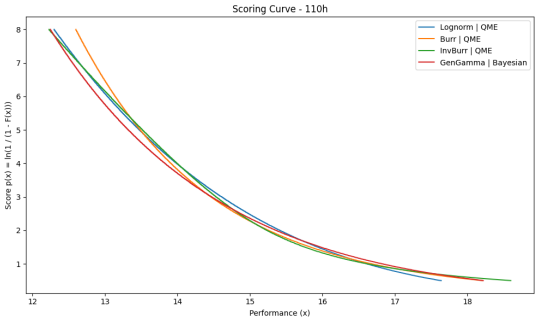
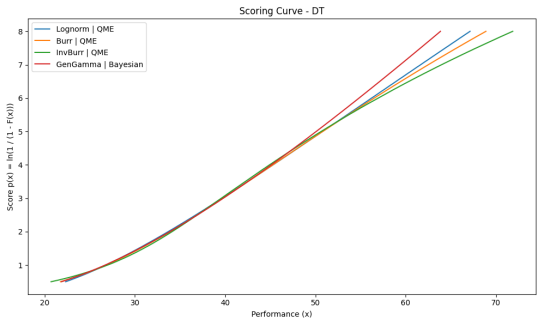


Figure 18: Scores of 400m and 110 m Hurdles



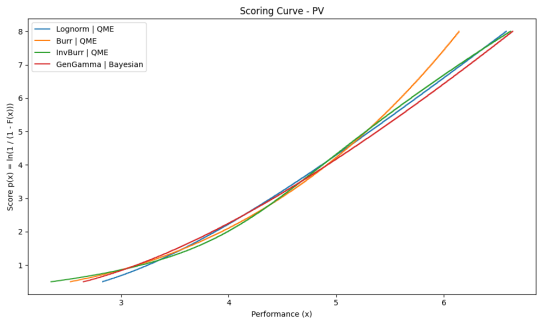
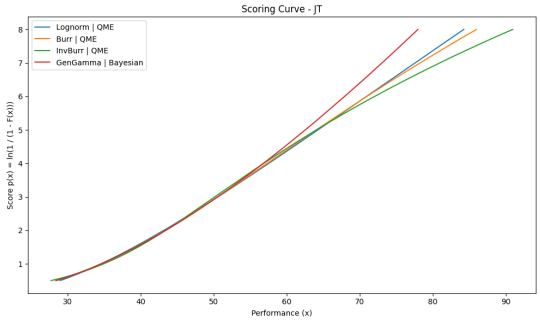


Figure 19: Scores of Discus Throw and Pole Vault



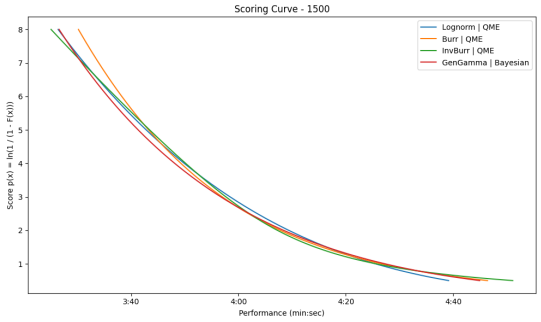


Figure 20: Scores of Javelin Throw and 1500m

Additional Tables

New Parameter estimation

Weibull

Table 1: Parameter estimates for the Weibull model across events (MLE, QME, Bayesian)

100						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	13.218	13.217	0.067	13.083	13.342
MLE	scale s	8.3	8.3	0.004	8.294	8.308
QME	shape c	10.0	10.0	0.0	10.0	10.0
QME	scale s	8.143	8.144	0.008	8.127	8.156
Bayesian	shape c	13.214	13.213	0.053	13.111	13.318
Bayesian	scale s	8.3	8.3	0.004	8.293	8.308

LJ						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	6.967	6.967	0.035	6.899	7.037
MLE	scale s	549.841	549.856	0.499	548.909	550.742
QME	shape c	5.2	5.198	0.059	5.081	5.307
QME	scale s	532.907	532.906	0.754	531.444	534.29
Bayesian	shape c	6.931	6.931	0.03	6.872	6.988
Bayesian	scale s	548.161	548.171	0.488	547.291	549.024

SP						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	4.051	4.051	0.044	3.965	4.136
MLE	scale s	9.526	9.526	0.024	9.48	9.571
QME	shape c	2.814	2.829	0.092	2.587	2.964
QME	scale s	8.781	8.793	0.071	8.615	8.889
Bayesian	shape c	4.049	4.049	0.028	3.993	4.104
Bayesian	scale s	9.526	9.526	0.025	9.476	9.575

HJ						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	8.09	8.088	0.047	7.995	8.19
MLE	scale s	160.665	160.659	0.139	160.411	160.935
QME	shape c	5.915	5.898	0.082	5.771	6.072
QME	scale s	155.789	155.775	0.23	155.365	156.213
Bayesian	shape c	8.085	8.085	0.039	8.008	8.162
Bayesian	scale s	160.616	160.615	0.139	160.347	160.89

400						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	13.511	13.511	0.082	13.362	13.676
MLE	scale s	7.369	7.37	0.004	7.361	7.377
QME	shape c	10.0	10.0	0.0	10.0	10.0
QME	scale s	7.208	7.209	0.006	7.197	7.222
Bayesian	shape c	13.5	13.5	0.069	13.365	13.637
Bayesian	scale s	7.369	7.369	0.004	7.361	7.377

110h						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	8.375	8.372	0.097	8.177	8.557
MLE	scale s	6.62	6.621	0.014	6.593	6.646
QME	shape c	7.536	7.533	0.199	7.168	7.949
QME	scale s	6.574	6.576	0.02	6.536	6.614
Bayesian	shape c	8.364	8.365	0.107	8.151	8.572
Bayesian	scale s	6.62	6.62	0.015	6.591	6.649

DT						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	3.158	3.157	0.029	3.105	3.215
MLE	scale s	27.747	27.745	0.097	27.573	27.941
QME	shape c	2.396	2.394	0.065	2.282	2.534
QME	scale s	25.816	25.812	0.223	25.388	26.263
Bayesian	shape c	3.154	3.154	0.024	3.108	3.201
Bayesian	scale s	27.742	27.743	0.099	27.549	27.937

PV						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	3.71	3.71	0.025	3.659	3.755
MLE	scale s	325.994	325.954	0.989	324.069	327.913
QME	shape c	3.122	3.12	0.045	3.047	3.217
QME	scale s	316.205	316.131	1.215	314.241	318.789
Bayesian	shape c	3.684	3.684	0.028	3.629	3.739
Bayesian	scale s	323.684	323.687	0.708	322.283	325.074

JT						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	3.289	3.287	0.027	3.24	3.34
MLE	scale s	35.785	35.785	0.107	35.579	35.995
QME	shape c	2.476	2.472	0.048	2.395	2.573
QME	scale s	33.369	33.375	0.196	32.996	33.735
Bayesian	shape c	3.286	3.286	0.022	3.244	3.33
Bayesian	scale s	35.773	35.773	0.105	35.566	35.981

1500						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	10.391	10.39	0.051	10.294	10.483
MLE	scale s	5.667	5.667	0.004	5.66	5.674
QME	shape c	9.031	9.028	0.107	8.828	9.231
QME	scale s	5.618	5.618	0.005	5.608	5.629
Bayesian	shape c	10.39	10.389	0.051	10.291	10.495
Bayesian	scale s	5.667	5.667	0.004	5.659	5.675

Lognormal

Table 2: Parameter estimates for the Lognormal model across events (MLE, QME, Bayesian)

100						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale θ	0.09	0.09	0.0	0.089	0.091
MLE	shape s	7.962	7.963	0.004	7.955	7.97
QME	scale θ	0.068	0.068	0.001	0.066	0.07
QME	shape s	8.064	8.063	0.008	8.05	8.08
Bayesian	shape s	0.09	0.09	0.0	0.089	0.091
Bayesian	scale θ	7.962	7.962	0.004	7.955	7.97

LJ						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale θ	0.14	0.14	0.001	0.139	0.141
MLE	shape s	511.833	511.83	0.42	511.045	512.632
QME	scale θ	0.142	0.142	0.002	0.139	0.145
QME	shape s	513.417	513.431	0.942	511.563	515.177
Bayesian	shape s	0.141	0.141	0.001	0.14	0.142
Bayesian	scale θ	511.81	511.805	0.426	510.98	512.64

SP						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale θ	0.24	0.24	0.002	0.236	0.243
MLE	shape s	8.444	8.444	0.02	8.407	8.482
QME	scale θ	0.259	0.257	0.009	0.246	0.282
QME	shape s	8.257	8.274	0.086	8.05	8.38
Bayesian	shape s	0.24	0.24	0.002	0.237	0.243
Bayesian	scale θ	8.445	8.445	0.02	8.406	8.483

HJ						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale θ	0.12	0.12	0.001	0.119	0.122
MLE	shape s	151.097	151.09	0.121	150.87	151.349
QME	scale θ	0.125	0.126	0.002	0.122	0.129
QME	shape s	150.623	150.586	0.297	150.094	151.213
Bayesian	shape s	0.121	0.121	0.001	0.12	0.122
Bayesian	scale θ	151.093	151.092	0.121	150.86	151.333

400						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale θ	0.088	0.088	0.001	0.087	0.089
MLE	shape s	7.076	7.076	0.004	7.067	7.085
QME	scale θ	0.065	0.065	0.001	0.063	0.067
QME	shape s	7.18	7.18	0.006	7.168	7.192
Bayesian	shape s	0.088	0.088	0.0	0.087	0.089
Bayesian	scale θ	7.076	7.076	0.004	7.068	7.084

110h						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale θ	0.132	0.132	0.001	0.13	0.134
MLE	shape s	6.206	6.206	0.014	6.178	6.233
QME	scale θ	0.098	0.098	0.003	0.093	0.103
QME	shape s	6.406	6.405	0.023	6.36	6.45
Bayesian	shape s	0.132	0.132	0.002	0.129	0.135
Bayesian	scale θ	6.205	6.205	0.014	6.177	6.234

DT						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale θ	0.328	0.328	0.002	0.324	0.333
MLE	shape s	23.569	23.567	0.082	23.421	23.742
QME	scale θ	0.3	0.3	0.008	0.283	0.315
QME	shape s	24.214	24.21	0.247	23.74	24.714
Bayesian	shape s	0.329	0.329	0.002	0.324	0.334
Bayesian	scale θ	23.57	23.57	0.082	23.41	23.73

PV						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale θ	0.273	0.273	0.002	0.27	0.276
MLE	shape s	283.689	283.677	0.811	282.167	285.291
QME	scale θ	0.23	0.23	0.003	0.223	0.236
QME	shape s	300.814	300.761	1.335	298.466	303.718
Bayesian	shape s	0.277	0.277	0.002	0.273	0.282
Bayesian	scale θ	283.69	283.681	0.831	282.065	285.331

JT						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale θ	0.326	0.326	0.002	0.322	0.33
MLE	shape s	30.54	30.537	0.094	30.366	30.721
QME	scale θ	0.29	0.291	0.006	0.28	0.3
QME	shape s	31.333	31.334	0.216	30.921	31.73
Bayesian	shape s	0.327	0.327	0.002	0.322	0.331
Bayesian	scale θ	30.54	30.54	0.093	30.358	30.723

1500						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale θ	0.107	0.107	0.0	0.106	0.108
MLE	shape s	5.381	5.381	0.004	5.373	5.388
QME	scale θ	0.082	0.082	0.001	0.08	0.084
QME	shape s	5.494	5.494	0.007	5.482	5.507
Bayesian	shape s	0.107	0.107	0.001	0.106	0.108
Bayesian	scale θ	5.381	5.381	0.004	5.373	5.388

Gamma

Table 3: Parameter estimates for the Gamma model across events (MLE, QME, Bayesian)

100						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape k	126.843	126.868	1.212	124.449	129.055
MLE	scale θ	0.063	0.063	0.001	0.062	0.064
QME	shape k	63.404	63.382	0.303	62.855	64.013
QME	scale θ	0.117	0.117	0.001	0.116	0.118
Bayesian	shape k	126.854	126.865	0.943	125.024	128.64
Bayesian	scale θ	0.063	0.063	0.0	0.062	0.064

LJ						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape k	50.588	50.6	0.402	49.76	51.33
MLE	scale θ	10.219	10.217	0.084	10.061	10.391
QME	shape k	79.075	79.024	1.165	76.973	81.419
QME	scale θ	6.929	6.932	0.094	6.74	7.109
Bayesian	shape k	50.575	50.574	0.429	49.743	51.405
Bayesian	scale θ	10.222	10.221	0.087	10.053	10.394

SP						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape k	17.454	17.448	0.255	16.969	17.943
MLE	scale θ	0.498	0.498	0.008	0.483	0.512
QME	shape k	14.304	11.891	5.54	11.091	28.654
QME	scale θ	0.652	0.704	0.126	0.346	0.751
Bayesian	shape k	17.441	17.439	0.236	16.996	17.9
Bayesian	scale θ	0.498	0.499	0.007	0.485	0.512

HJ						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape k	68.517	68.514	0.633	67.21	69.705
MLE	scale θ	2.222	2.222	0.021	2.182	2.264
QME	shape k	61.751	60.312	3.268	57.282	67.683
QME	scale θ	2.475	2.517	0.113	2.277	2.644
Bayesian	shape k	68.506	68.522	0.642	67.275	69.795
Bayesian	scale θ	2.222	2.221	0.021	2.18	2.263

400						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape k	132.745	132.743	1.867	129.198	136.47
MLE	scale θ	0.054	0.054	0.001	0.052	0.055
QME	shape k	65.067	65.065	0.34	64.419	65.703
QME	scale θ	0.101	0.101	0.001	0.1	0.102
Bayesian	shape k	132.719	132.716	1.296	130.2	135.234
Bayesian	scale θ	0.054	0.054	0.001	0.053	0.055

110h						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape k	58.21	58.257	1.146	55.997	60.276
MLE	scale θ	0.108	0.108	0.002	0.104	0.112
QME	shape k	76.828	76.756	2.741	71.586	82.396
QME	scale θ	0.082	0.082	0.003	0.077	0.087
Bayesian	shape k	58.051	58.055	1.404	55.284	60.772
Bayesian	scale θ	0.108	0.108	0.003	0.103	0.113

DT						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape k	9.519	9.521	0.132	9.267	9.768
MLE	scale θ	2.612	2.611	0.038	2.536	2.684
QME	shape k	8.159	8.093	1.013	7.275	9.19
QME	scale θ	3.02	3.025	0.185	2.707	3.312
Bayesian	shape k	9.509	9.51	0.142	9.224	9.785
Bayesian	scale θ	2.615	2.614	0.04	2.54	2.698

PV						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape k	13.387	13.386	0.157	13.097	13.68
MLE	scale θ	22.011	22.009	0.284	21.473	22.554
QME	shape k	14.602	14.57	0.463	13.838	15.596
QME	scale θ	20.669	20.692	0.595	19.497	21.744
Bayesian	shape k	13.37	13.372	0.196	12.987	13.756
Bayesian	scale θ	22.043	22.036	0.33	21.411	22.698

JT						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape k	9.889	9.886	0.132	9.652	10.14
MLE	scale θ	3.252	3.252	0.044	3.166	3.334
QME	shape k	8.904	8.691	1.67	8.099	9.753
QME	scale θ	3.61	3.64	0.233	3.295	3.872
Bayesian	shape k	9.885	9.881	0.121	9.648	10.123
Bayesian	scale θ	3.253	3.253	0.041	3.174	3.335

1500						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape k	89.027	89.058	0.787	87.53	90.536
MLE	scale θ	0.061	0.061	0.001	0.06	0.062
QME	shape k	66.389	68.365	4.203	57.579	69.469
QME	scale θ	0.079	0.076	0.005	0.075	0.089
Bayesian	shape k	89.036	89.029	0.887	87.27	90.855
Bayesian	scale θ	0.061	0.061	0.001	0.06	0.062

GenGamma

Table 4: Parameter estimates for the Generalized Gamma model across events (MLE, Bayesian)

100						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale a	6.823	6.815	0.161	6.524	7.122
MLE	shape d	22.549	22.591	0.947	20.776	24.396
MLE	shape p	6.785	6.746	0.349	6.206	7.485
Bayesian	scale a	7.506	7.507	0.046	7.413	7.596
Bayesian	shape d	18.533	18.525	0.294	17.969	19.122
Bayesian	shape p	8.65	8.649	0.169	8.318	8.986

LJ						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale a	312.555	312.906	13.219	286.212	334.825
MLE	shape d	15.77	15.758	0.51	14.871	16.765
MLE	shape p	3.103	3.106	0.111	2.881	3.299
Bayesian	scale a	79.859	79.786	5.597	69.111	91.146
Bayesian	shape d	30.851	30.843	0.68	29.531	32.224
Bayesian	shape p	1.584	1.584	0.036	1.515	1.655

SP						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale a	1.777	1.774	0.164	1.485	2.067
MLE	shape d	12.088	12.079	0.436	11.329	12.961
MLE	shape p	1.364	1.362	0.039	1.289	1.437
Bayesian	scale a	0.376	0.37	0.078	0.241	0.545
Bayesian	shape d	18.177	18.151	0.708	16.837	19.584
Bayesian	shape p	0.94	0.939	0.038	0.868	1.014

HJ						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale a	96.46	96.445	3.202	90.442	102.724
MLE	shape d	18.887	18.882	0.608	17.746	20.022
MLE	shape p	3.53	3.525	0.109	3.334	3.742
Bayesian	scale a	41.157	41.198	2.126	37.014	45.41
Bayesian	shape d	31.811	31.791	0.699	30.443	33.232
Bayesian	shape p	2.075	2.075	0.048	1.983	2.17

400						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale a	6.075	6.069	0.139	5.809	6.345
MLE	shape d	23.163	23.186	0.962	21.251	24.988
MLE	shape p	6.924	6.898	0.353	6.298	7.631
Bayesian	scale a	6.6	6.6	0.05	6.501	6.699
Bayesian	shape d	19.553	19.552	0.377	18.807	20.29
Bayesian	shape p	8.538	8.534	0.199	8.16	8.943

110h						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale a	4.056	4.052	0.121	3.816	4.288
MLE	shape d	17.438	17.446	0.52	16.364	18.514
MLE	shape p	3.521	3.516	0.105	3.321	3.737
Bayesian	scale a	3.754	3.756	0.225	3.299	4.194
Bayesian	shape d	18.543	18.513	0.861	16.93	20.342
Bayesian	shape p	3.277	3.274	0.179	2.936	3.645

DT						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale a	5.757	5.536	0.777	4.639	7.314
MLE	shape d	7.56	7.668	0.367	6.825	8.102
MLE	shape p	1.232	1.223	0.049	1.152	1.332
Bayesian	scale a	0.779	0.756	0.245	0.363	1.322
Bayesian	shape d	11.94	11.909	0.614	10.824	13.28
Bayesian	shape p	0.782	0.782	0.042	0.699	0.864

PV						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale a	78.61	78.575	4.87	71.204	83.409
MLE	shape d	8.856	8.861	0.197	8.498	9.226
MLE	shape p	1.39	1.39	0.031	1.344	1.416
Bayesian	scale a	8.321	8.072	2.123	4.854	13.078
Bayesian	shape d	15.809	15.822	0.663	14.518	17.123
Bayesian	shape p	0.825	0.823	0.035	0.759	0.896

JT						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale a	11.04	10.689	0.862	10.324	13.455
MLE	shape d	6.781	6.818	0.172	6.34	7.018
MLE	shape p	1.431	1.413	0.047	1.39	1.568
Bayesian	scale a	4.443	4.411	0.88	2.84	6.274
Bayesian	shape d	9.21	9.187	0.46	8.373	10.155
Bayesian	shape p	1.077	1.077	0.056	0.968	1.188

1500						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	scale a	3.589	3.582	0.085	3.431	3.762
MLE	shape d	23.234	23.268	0.551	22.11	24.307
MLE	shape p	4.077	4.069	0.107	3.88	4.308
Bayesian	scale a	3.379	3.38	0.106	3.164	3.582
Bayesian	shape d	24.57	24.563	0.701	23.236	26.009
Bayesian	shape p	3.829	3.826	0.121	3.59	4.07

Burr

Table 5: Parameter estimates for the Burr model across events (MLE, QME, Bayesian)

100						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	15.494	15.484	0.091	15.347	15.676
MLE	shape k	3.992	3.997	0.116	3.747	4.191
MLE	scale s	8.963	8.965	0.023	8.911	9.002
QME	shape c	20.279	20.409	0.612	18.59	21.347
QME	shape k	1.923	1.87	0.168	1.683	2.412
QME	scale s	8.428	8.413	0.05	8.352	8.571
Bayesian	shape c	15.467	15.468	0.107	15.255	15.671
Bayesian	shape k	4.188	4.183	0.188	3.837	4.57
Bayesian	scale s	8.996	8.996	0.035	8.929	9.065

LJ						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	12.771	12.748	0.13	12.56	13.031
MLE	shape k	0.916	0.919	0.016	0.883	0.941
MLE	scale s	504.404	504.447	1.284	501.782	506.635
QME	shape c	7.007	6.984	0.207	6.636	7.445
QME	shape k	4.334	4.288	0.551	3.414	5.5
QME	scale s	650.862	650.396	16.727	621.407	684.06
Bayesian	shape c	9.319	9.319	0.047	9.228	9.411
Bayesian	shape k	2.428	2.428	0.03	2.369	2.487
Bayesian	scale s	582.517	582.535	1.005	580.52	584.449

SP						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	7.154	7.152	0.065	7.029	7.291
MLE	shape k	1.062	1.062	0.013	1.031	1.086
MLE	scale s	8.542	8.544	0.032	8.481	8.603
QME	shape c	5.795	5.55	0.776	4.864	8.008
QME	shape k	1.811	1.88	0.379	0.986	2.506
QME	scale s	9.665	9.751	0.511	8.559	10.574
Bayesian	shape c	7.157	7.158	0.118	6.928	7.389
Bayesian	shape k	1.062	1.06	0.045	0.979	1.153
Bayesian	scale s	8.545	8.544	0.08	8.392	8.704

HJ						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	14.766	14.762	0.122	14.559	15.049
MLE	shape k	0.941	0.942	0.012	0.913	0.962
MLE	scale s	149.756	149.774	0.278	149.213	150.248
QME	shape c	8.29	8.297	0.301	7.671	8.826
QME	shape k	3.797	3.659	0.59	3.026	5.213
QME	scale s	181.275	180.342	4.695	174.45	192.084
Bayesian	shape c	14.067	14.067	0.166	13.749	14.395
Bayesian	shape k	1.047	1.046	0.03	0.987	1.106
Bayesian	scale s	151.684	151.683	0.494	150.719	152.641

400						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	16.112	16.114	0.146	15.827	16.389
MLE	shape k	3.504	3.538	0.153	3.229	3.759
MLE	scale s	7.856	7.861	0.03	7.8	7.906
QME	shape c	14.995	14.987	0.448	14.121	15.966
QME	shape k	4.807	4.752	0.704	3.56	6.295
QME	scale s	8.079	8.075	0.109	7.871	8.299
Bayesian	shape c	16.0	16.001	0.147	15.712	16.29
Bayesian	shape k	3.763	3.755	0.207	3.389	4.19
Bayesian	scale s	7.901	7.9	0.037	7.832	7.974

110h						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	10.067	10.071	0.187	9.719	10.436
MLE	shape k	2.993	2.95	0.344	2.404	3.686
MLE	scale s	7.185	7.176	0.118	6.972	7.417
QME	shape c	8.109	8.044	0.379	7.561	8.936
QME	shape k	49.314	20.294	59.172	6.706	203.374
QME	scale s	10.018	9.557	1.531	8.052	12.99
Bayesian	shape c	8.508	8.507	0.111	8.292	8.728
Bayesian	shape k	48.615	48.639	1.031	46.56	50.595
Bayesian	scale s	10.429	10.428	0.064	10.303	10.557

DT						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	4.842	4.844	0.037	4.769	4.913
MLE	shape k	1.291	1.291	0.004	1.282	1.299
MLE	scale s	25.536	25.534	0.11	25.319	25.735
QME	shape c	3.735	3.723	0.279	3.246	4.361
QME	shape k	3.196	3.015	0.815	2.032	5.12
QME	scale s	35.063	34.455	3.39	29.782	42.628
Bayesian	shape c	4.507	4.506	0.073	4.369	4.655
Bayesian	shape k	1.677	1.675	0.094	1.499	1.868
Bayesian	scale s	27.821	27.813	0.522	26.816	28.865

PV						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	5.75	5.746	0.039	5.675	5.826
MLE	shape k	1.224	1.224	0.004	1.216	1.232
MLE	scale s	298.104	298.13	1.072	296.057	300.208
QME	shape c	3.142	3.139	0.042	3.066	3.229
QME	shape k	258.394	262.259	69.658	107.535	408.448
QME	scale s	1835.012	1858.234	171.778	1413.667	2129.373
Bayesian	shape c	5.785	5.784	0.057	5.669	5.896
Bayesian	shape k	1.197	1.197	0.02	1.159	1.237
Bayesian	scale s	296.448	296.447	1.001	294.482	298.403

JT						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	4.874	4.995	0.225	4.465	5.079
MLE	shape k	1.508	1.316	0.306	1.308	2.066
MLE	scale s	34.719	33.452	2.093	33.205	38.497
QME	shape c	3.899	3.884	0.2	3.54	4.319
QME	shape k	3.012	2.977	0.466	2.26	4.081
QME	scale s	43.967	43.844	2.538	39.719	49.525
Bayesian	shape c	4.485	4.484	0.053	4.384	4.59
Bayesian	shape k	2.046	2.043	0.088	1.88	2.226
Bayesian	scale s	38.346	38.334	0.537	37.322	39.424

1500						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape c	12.899	12.892	0.11	12.69	13.11
MLE	shape k	2.636	2.646	0.109	2.438	2.847
MLE	scale s	5.964	5.966	0.028	5.913	6.02
QME	shape c	10.021	10.013	0.283	9.51	10.545
QME	shape k	14.341	12.172	7.386	7.443	31.55
QME	scale s	7.259	7.189	0.361	6.77	8.077
Bayesian	shape c	12.692	12.691	0.135	12.43	12.959
Bayesian	shape k	2.936	2.926	0.169	2.635	3.288
Bayesian	scale s	6.035	6.034	0.038	5.964	6.112

Inverse Burr

Table 6: Parameter estimates for the Inverse Burr model across events (MLE, QME, Bayesian)

100						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape α	0.418	0.418	0.007	0.407	0.431
MLE	shape τ	31.129	31.151	0.304	30.516	31.686
MLE	scale θ	8.453	8.454	0.008	8.437	8.466
QME	shape α	0.292	0.294	0.02	0.247	0.329
QME	shape τ	35.802	35.615	0.944	34.288	38.14
QME	scale θ	8.593	8.59	0.027	8.545	8.655
Bayesian	shape α	0.414	0.414	0.007	0.401	0.428
Bayesian	shape τ	31.418	31.413	0.315	30.819	32.043
Bayesian	scale θ	8.456	8.456	0.007	8.441	8.47

LJ						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape α	1.575	1.548	0.035	1.537	1.626
MLE	shape τ	11.001	10.999	0.042	10.925	11.087
MLE	scale θ	479.957	480.648	1.583	477.384	482.15
QME	shape α	0.203	0.203	0.009	0.187	0.222
QME	shape τ	18.571	18.564	0.398	17.781	19.3
QME	scale θ	606.232	606.472	2.619	600.967	611.131
Bayesian	shape α	0.76	0.76	0.011	0.74	0.781
Bayesian	shape τ	13.26	13.26	0.095	13.074	13.445
Bayesian	scale θ	532.063	532.066	0.846	530.381	533.721

SP						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape α	1.342	1.342	0.005	1.331	1.352
MLE	shape τ	6.649	6.647	0.047	6.561	6.738
MLE	scale θ	7.911	7.911	0.022	7.867	7.954
QME	shape α	0.374	0.353	0.08	0.285	0.606
QME	shape τ	9.268	9.344	0.454	8.204	10.02
QME	scale θ	10.147	10.208	0.327	9.306	10.618
Bayesian	shape α	1.187	1.186	0.052	1.091	1.292
Bayesian	shape τ	6.912	6.911	0.109	6.705	7.13
Bayesian	scale θ	8.134	8.135	0.08	7.977	8.289

HJ						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape α	1.492	1.491	0.006	1.481	1.504
MLE	shape τ	12.861	12.862	0.061	12.743	12.984
MLE	scale θ	144.051	144.05	0.167	143.727	144.384
QME	shape α	0.22	0.222	0.012	0.196	0.24
QME	shape τ	20.649	20.543	0.559	19.839	21.782
QME	scale θ	173.273	173.153	0.836	171.855	174.932
Bayesian	shape α	1.437	1.435	0.043	1.357	1.526
Bayesian	shape τ	12.937	12.936	0.12	12.699	13.172
Bayesian	scale θ	144.917	144.932	0.507	143.912	145.887

400						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape α	0.461	0.46	0.01	0.445	0.481
MLE	shape τ	30.19	30.204	0.364	29.475	30.894
MLE	scale θ	7.461	7.462	0.01	7.441	7.478
QME	shape α	0.193	0.192	0.009	0.176	0.214
QME	shape τ	41.413	41.485	1.01	39.135	43.321
QME	scale θ	7.766	7.765	0.016	7.733	7.795
Bayesian	shape α	0.441	0.441	0.009	0.423	0.459
Bayesian	shape τ	31.161	31.155	0.386	30.424	31.939
Bayesian	scale θ	7.476	7.476	0.009	7.46	7.494

110h						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape α	0.877	0.77	0.278	0.592	1.401
MLE	shape τ	13.92	14.194	1.418	11.329	16.186
MLE	scale θ	6.35	6.421	0.221	5.932	6.624
QME	shape α	0.142	0.141	0.013	0.119	0.167
QME	shape τ	30.131	30.139	1.665	27.308	33.545
QME	scale θ	7.358	7.362	0.044	7.266	7.43
Bayesian	shape α	0.441	0.44	0.03	0.387	0.506
Bayesian	shape τ	19.22	19.216	0.734	17.811	20.675
Bayesian	scale θ	6.783	6.786	0.042	6.697	6.861

DT						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape α	1.258	1.257	0.005	1.248	1.269
MLE	shape τ	4.848	4.849	0.035	4.78	4.912
MLE	scale θ	22.019	22.017	0.099	21.832	22.214
QME	shape α	0.221	0.219	0.021	0.183	0.264
QME	shape τ	8.653	8.619	0.37	8.027	9.474
QME	scale θ	33.85	33.881	0.642	32.621	35.01
Bayesian	shape α	0.964	0.962	0.045	0.878	1.056
Bayesian	shape τ	5.304	5.303	0.097	5.112	5.498
Bayesian	scale θ	23.941	23.946	0.337	23.272	24.597

PV						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape α	2.975	2.976	0.156	2.85	3.101
MLE	shape τ	4.881	4.878	0.064	4.817	4.94
MLE	scale θ	208.242	208.016	4.136	204.845	211.588
QME	shape α	0.118	0.117	0.005	0.108	0.127
QME	shape τ	13.37	13.346	0.385	12.671	14.181
QME	scale θ	426.87	426.896	2.932	421.415	432.167
Bayesian	shape α	0.393	0.393	0.008	0.378	0.408
Bayesian	shape τ	8.678	8.677	0.127	8.433	8.93
Bayesian	scale θ	361.443	361.445	0.965	359.575	363.305

JT						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape α	0.677	0.677	0.027	0.624	0.739
MLE	shape τ	6.338	6.326	0.143	6.073	6.622
MLE	scale θ	34.21	34.214	0.379	33.313	34.97
QME	shape α	0.259	0.259	0.02	0.224	0.301
QME	shape τ	8.629	8.608	0.254	8.196	9.147
QME	scale θ	42.017	42.027	0.637	40.783	43.201
Bayesian	shape α	0.641	0.641	0.021	0.6	0.683
Bayesian	shape τ	6.511	6.511	0.112	6.299	6.73
Bayesian	scale θ	34.718	34.72	0.291	34.15	35.294

1500						
Method	**Param**	**Mean**	**Median**	**Std**	**CI Low.**	**CI Up.**
MLE	shape α	0.668	0.667	0.014	0.642	0.698
MLE	shape τ	19.069	19.065	0.185	18.693	19.433
MLE	scale θ	5.598	5.598	0.012	5.574	5.622
QME	shape α	0.149	0.149	0.007	0.138	0.162
QME	shape τ	35.318	35.271	0.947	33.591	37.062
QME	scale θ	6.155	6.154	0.014	6.127	6.182
Bayesian	shape α	0.615	0.615	0.017	0.584	0.649
Bayesian	shape τ	20.046	20.043	0.278	19.496	20.588
Bayesian	scale θ	5.634	5.634	0.013	5.609	5.658

Full Model Fit

100m											
Event	**Model**	**Method**	**KS**	**p-KS**	**CvM**	**p-CvM**	**AD**	**p-AD**	**AD right**	**p-AD right**	**LogLik**
100	Burr	Bayesian	0.009	0.0	0.635	0.018	4.68	0.002	-16711.608	0.34	-34549.927
100	Burr	QME	0.059	0.0	34.997	0.0	317.737	0.0	-18102.996	1.0	-35463.279
100	Burr	MLE	0.009	0.0	0.733	0.01	5.201	0.0	-16711.619	0.342	-34551.184
100	Gamma	MLE	0.057	0.0	37.567	0.0	224.17	0.0	-15506.299	0.0	-35962.579
100	Gamma	QME	0.337	0.0	1754.871	0.0	8314.913	0.0	989.721	0.0	-45126.195
100	Gamma	Bayesian	0.056	0.0	36.741	0.0	221.423	0.0	-15539.798	0.0	-35962.591
100	Gengamma	Bayesian	0.015	0.0	2.66	0.0	16.208	0.0	-16626.448	0.104	-34650.989
100	Gengamma	MLE	0.023	0.0	5.399	0.0	32.999	0.0	-16402.552	0.0	-34732.753
100	Invburr	Bayesian	0.014	0.0	1.414	0.0	10.654	0.0	-16903.052	0.928	-34640.477
100	Invburr	QME	0.034	0.0	8.505	0.0	82.303	0.0	-16126.183	0.0	-34952.523
100	Invburr	MLE	0.013	0.0	1.232	0.0	9.725	0.0	-16905.328	0.894	-34640.389
100	Lognorm	Bayesian	0.063	0.0	46.799	0.0	280.361	0.0	-15355.311	0.0	-36366.808
100	Lognorm	QME	0.08	0.0	45.963	0.0	804.766	0.0	-18005.575	1.0	-39953.19
100	Lognorm	MLE	0.063	0.0	46.577	0.0	279.551	0.0	-15362.115	0.0	-36366.807
100	Weibull	Bayesian	0.031	0.0	7.621	0.0	51.192	0.0	-16938.816	0.972	-34935.664
100	Weibull	QME	0.143	0.0	274.299	0.0	1412.138	0.0	-11413.581	0.0	-37109.878
100	Weibull	MLE	0.031	0.0	7.615	0.0	51.081	0.0	-16943.137	0.974	-34935.662

Table 7: Fitting statistics of the full distribution : 100m

Long Jump											
Event	**Model**	**Method**	**KS**	**p-KS**	**CvM**	**p-CvM**	**AD**	**p-AD**	**AD right**	**p-AD right**	**LogLik**
LJ	Burr	MLE	0.045	0.0	17.725	0.0	137.488	0.0	-13918.817	0.284	-159728.547
LJ	Burr	QME	0.132	0.0	62.837	0.0	535.402	0.0	-12367.53	0.0	-161624.101
LJ	Burr	Bayesian	0.069	0.0	28.51	0.0	213.388	0.0	-14676.018	1.0	-160356.417
LJ	Gamma	MLE	0.044	0.0	15.392	0.0	116.726	0.0	-14545.65	1.0	-159285.608
LJ	Gamma	QME	0.237	0.0	712.149	0.0	4783.938	0.0	-20575.596	1.0	-164569.718
LJ	Gamma	Bayesian	0.044	0.0	15.124	0.0	115.693	0.0	-14530.532	1.0	-159285.593
LJ	Gengamma	Bayesian	0.051	0.0	15.835	0.0	124.331	0.0	-14586.602	1.0	-159456.111
LJ	Gengamma	MLE	0.062	0.0	23.137	0.0	179.485	0.0	-14683.565	1.0	-159988.235
LJ	Invburr	Bayesian	0.061	0.0	31.787	0.0	206.101	0.0	-14949.804	1.0	-160136.984
LJ	Invburr	QME	0.209	0.0	194.994	0.0	1273.384	0.0	-11081.09	0.0	-164831.195
LJ	Invburr	MLE	0.042	0.0	17.527	0.0	125.918	0.0	-13643.122	0.0	-159420.327
LJ	Lognorm	MLE	0.04	0.0	11.834	0.0	92.886	0.0	-14385.943	1.0	-159058.992
LJ	Lognorm	Bayesian	0.04	0.0	10.561	0.0	86.135	0.0	-14378.675	1.0	-159061.407
LJ	Lognorm	QME	0.043	0.0	14.835	0.0	103.502	0.0	-14716.725	1.0	-159072.69
LJ	Weibull	Bayesian	0.106	0.0	52.978	0.0	420.608	0.0	-14442.135	1.0	-162017.924
LJ	Weibull	QME	0.202	0.0	229.891	0.0	1421.132	0.0	-10273.85	0.0	-164050.855
LJ	Weibull	MLE	0.103	0.0	61.091	0.0	444.095	0.0	-14755.922	1.0	-162012.518

Table 8: Fitting statistics of the full distribution : Long Jump

Shot Put											
Event	**Model**	**Method**	**KS**	**p-KS**	**CvM**	**p-CvM**	**AD**	**p-AD**	**AD right**	**p-AD right**	**LogLik**
SP	Burr	MLE	0.019	0.002	1.177	0.0	8.01	0.0	-5068.968	0.464	-21645.975
SP	Burr	Bayesian	0.02	0.002	1.192	0.0	8.046	0.0	-5078.933	0.528	-21645.963
SP	Burr	QME	0.044	0.0	3.043	0.0	34.156	0.0	-4850.229	0.002	-21782.21
SP	Gamma	MLE	0.02	0.0	1.131	0.002	7.907	0.0	-5191.464	0.99	-21643.152
SP	Gamma	QME	0.095	0.0	33.325	0.0	202.458	0.0	-4121.064	0.0	-22067.474
SP	Gamma	Bayesian	0.021	0.0	1.337	0.002	8.659	0.0	-5212.131	0.994	-21643.105
SP	Gengamma	Bayesian	0.016	0.002	0.877	0.004	7.183	0.0	-5155.808	0.934	-21636.741
SP	Gengamma	MLE	0.028	0.0	3.121	0.0	20.664	0.0	-5258.827	1.0	-21708.594
SP	Invburr	QME	0.102	0.0	20.72	0.0	166.057	0.0	-4503.329	0.0	-22368.762
SP	Invburr	Bayesian	0.019	0.002	1.192	0.0	8.0	0.0	-5061.524	0.43	-21638.341
SP	Invburr	MLE	0.02	0.0	1.251	0.002	8.844	0.0	-5046.639	0.284	-21641.407
SP	Lognorm	Bayesian	0.008	0.562	0.052	0.858	0.488	0.79	-5096.128	0.64	-21562.527
SP	Lognorm	QME	0.042	0.0	7.339	0.0	43.333	0.0	-4610.516	0.0	-21642.41
SP	Lognorm	MLE	0.008	0.558	0.05	0.876	0.468	0.796	-5095.1	0.632	-21562.524
SP	Weibull	MLE	0.072	0.0	19.664	0.0	139.553	0.0	-5250.035	1.0	-22559.387
SP	Weibull	Bayesian	0.072	0.0	19.705	0.0	139.845	0.0	-5249.39	1.0	-22559.473
SP	Weibull	QME	0.214	0.0	155.786	0.0	824.699	0.0	-2967.914	0.0	-23710.69

Table 9: Fitting statistics of the full distribution : Shot Put

High Jump											
Event	**Model**	**Method**	**KS**	**p-KS**	**CvM**	**p-CvM**	**AD**	**p-AD**	**AD right**	**p-AD right**	**LogLik**
HJ	Burr	QME	0.121	0.0	62.437	0.0	467.06	0.0	-9497.071	0.0	-98048.83
HJ	Burr	Bayesian	0.065	0.0	10.168	0.0	71.377	0.0	-11290.683	0.97	-96564.243
HJ	Burr	MLE	0.063	0.0	10.18	0.0	71.557	0.0	-11120.647	0.476	-96539.888
HJ	Gamma	QME	0.046	0.0	6.128	0.0	59.115	0.0	-11127.48	0.548	-96357.279
HJ	Gamma	MLE	0.06	0.0	9.539	0.0	61.723	0.0	-11516.45	1.0	-96265.272
HJ	Gamma	Bayesian	0.059	0.0	9.168	0.0	60.407	0.0	-11488.855	1.0	-96265.209
HJ	Gengamma	Bayesian	0.066	0.0	12.418	0.0	81.662	0.0	-11603.612	1.0	-96461.669
HJ	Gengamma	MLE	0.078	0.0	20.397	0.0	129.281	0.0	-11779.835	1.0	-96795.733
HJ	Invburr	Bayesian	0.061	0.0	8.513	0.0	59.19	0.0	-11193.894	0.782	-96389.487
HJ	Invburr	QME	0.187	0.0	155.413	0.0	995.308	0.0	-8683.123	0.0	-100333.373
HJ	Invburr	MLE	0.056	0.0	8.892	0.0	60.358	0.0	-10996.615	0.064	-96378.281
HJ	Lognorm	Bayesian	0.056	0.0	6.624	0.0	44.255	0.0	-11376.918	0.994	-96135.522
HJ	Lognorm	QME	0.039	0.0	4.234	0.0	39.597	0.0	-11029.872	0.124	-96185.009
HJ	Lognorm	MLE	0.056	0.0	6.802	0.0	45.34	0.0	-11376.388	0.994	-96135.264
HJ	Weibull	MLE	0.09	0.0	51.607	0.0	355.433	0.0	-11693.72	1.0	-98451.969
HJ	Weibull	QME	0.203	0.0	229.597	0.0	1329.617	0.0	-7682.98	0.0	-100348.619
HJ	Weibull	Bayesian	0.09	0.0	50.862	0.0	353.091	0.0	-11668.197	1.0	-98452.022

Table 10: Fitting statistics of the full distribution : High Jump

400m											
Event	**Model**	**Method**	**KS**	**p-KS**	**CvM**	**p-CvM**	**AD**	**p-AD**	**AD right**	**p-AD right**	**LogLik**
400	Burr	Bayesian	0.008	0.082	0.337	0.106	2.629	0.032	-10468.345	0.454	-18636.863
400	Burr	QME	0.025	0.0	3.122	0.0	20.028	0.0	-10158.614	0.0	-18668.677
400	Burr	MLE	0.008	0.106	0.277	0.154	2.315	0.054	-10461.438	0.426	-18637.213
400	Gamma	QME	0.349	0.0	1169.925	0.0	5519.282	0.0	1048.044	0.0	-25612.658
400	Gamma	Bayesian	0.05	0.0	19.136	0.0	121.347	0.0	-9751.528	0.0	-19538.499
400	Gamma	MLE	0.05	0.0	19.169	0.0	121.44	0.0	-9749.624	0.0	-19538.501
400	Gengamma	Bayesian	0.016	0.0	1.486	0.0	10.225	0.0	-10432.82	0.316	-18698.495
400	Gengamma	MLE	0.021	0.0	2.554	0.0	16.461	0.0	-10222.167	0.0	-18743.024
400	Invburr	MLE	0.013	0.0	0.796	0.006	5.566	0.002	-10522.25	0.682	-18700.345
400	Invburr	QME	0.1	0.0	54.959	0.0	370.065	0.0	-9177.132	0.0	-19914.486
400	Invburr	Bayesian	0.016	0.0	1.274	0.002	7.933	0.0	-10548.296	0.83	-18703.254
400	Lognorm	Bayesian	0.056	0.0	24.519	0.0	155.195	0.0	-9657.137	0.0	-19800.97
400	Lognorm	QME	0.084	0.0	41.756	0.0	609.961	0.0	-11711.716	1.0	-22725.446
400	Lognorm	MLE	0.056	0.0	24.427	0.0	154.78	0.0	-9659.225	0.0	-19800.967
400	Weibull	Bayesian	0.031	0.0	7.586	0.0	48.407	0.0	-10638.637	0.978	-18892.195
400	Weibull	MLE	0.031	0.0	7.667	0.0	48.527	0.0	-10652.771	0.99	-18892.189
400	Weibull	QME	0.162	0.0	216.404	0.0	1096.199	0.0	-6633.104	0.0	-20505.453

Table 11: Fitting statistics of the full distribution : 400m

110 m Hurdles											
Event	**Model**	**Method**	**KS**	**p-KS**	**CvM**	**p-CvM**	**AD**	**p-AD**	**AD right**	**p-AD right**	**LogLik**
110h	Burr	MLE	0.055	0.0	1.698	0.0	12.044	0.0	-1717.304	0.792	-4163.886
110h	Burr	QME	0.053	0.0	1.15	0.0	14.909	0.0	-1681.197	0.38	-4201.645
110h	Burr	Bayesian	0.05	0.0	1.829	0.0	15.064	0.0	-1752.637	0.974	-4193.036
110h	Gamma	MLE	0.043	0.0	1.963	0.0	11.398	0.0	-1670.043	0.246	-4110.445
110h	Gamma	QME	0.093	0.0	7.163	0.0	54.811	0.0	-1777.395	0.998	-4189.874
110h	Gamma	Bayesian	0.041	0.0	1.991	0.0	11.348	0.0	-1660.789	0.176	-4110.393
110h	Gengamma	Bayesian	0.047	0.0	1.409	0.0	8.845	0.0	-1691.605	0.514	-4109.789
110h	Gengamma	MLE	0.049	0.0	1.433	0.0	9.091	0.0	-1700.904	0.628	-4111.345
110h	Invburr	Bayesian	0.071	0.0	3.405	0.0	22.015	0.0	-1755.959	0.976	-4231.615
110h	Invburr	QME	0.141	0.0	11.842	0.0	89.847	0.0	-1466.629	0.0	-4537.371
110h	Invburr	MLE	0.058	0.0	2.533	0.0	17.137	0.0	-1697.737	0.57	-4212.53
110h	Lognorm	QME	0.179	0.0	28.225	0.0	279.485	0.0	-2074.929	1.0	-4660.844
110h	Lognorm	MLE	0.047	0.0	2.435	0.0	13.711	0.0	-1642.366	0.064	-4120.689
110h	Lognorm	Bayesian	0.047	0.0	2.407	0.0	13.516	0.0	-1640.868	0.06	-4120.696
110h	Weibull	Bayesian	0.049	0.0	1.903	0.0	15.732	0.0	-1756.902	0.98	-4198.354
110h	Weibull	MLE	0.049	0.0	1.939	0.0	15.852	0.0	-1758.918	0.982	-4198.354
110h	Weibull	QME	0.067	0.0	1.863	0.0	22.306	0.0	-1610.316	0.01	-4229.561

Table 12: Fitting statistics of the full distribution : 110m Hurdles

Discus Throw											
Event	**Model**	**Method**	**KS**	**p-KS**	**CvM**	**p-CvM**	**AD**	**p-AD**	**AD right**	**p-AD right**	**LogLik**
DT	Burr	MLE	0.03	0.0	2.664	0.0	18.503	0.0	-4407.826	0.552	-30645.663
DT	Burr	QME	0.041	0.0	1.884	0.0	26.721	0.0	-4327.232	0.1	-30731.234
DT	Burr	Bayesian	0.031	0.0	2.511	0.0	18.167	0.0	-4450.924	0.836	-30645.146
DT	Gamma	QME	0.033	0.0	2.2	0.0	22.786	0.0	-4163.822	0.0	-30596.304
DT	Gamma	Bayesian	0.025	0.0	1.382	0.0	9.41	0.0	-4486.183	0.972	-30531.087
DT	Gamma	MLE	0.025	0.0	1.385	0.0	9.439	0.0	-4484.284	0.964	-30531.094
DT	Gengamma	Bayesian	0.022	0.0	1.051	0.0	7.186	0.0	-4473.821	0.932	-30511.128
DT	Gengamma	MLE	0.019	0.002	0.956	0.0	11.172	0.0	-4309.704	0.026	-30565.925
DT	Invburr	Bayesian	0.031	0.0	2.985	0.0	21.081	0.0	-4422.335	0.688	-30671.111
DT	Invburr	QME	0.13	0.0	29.896	0.0	223.271	0.0	-3763.696	0.0	-31572.339
DT	Invburr	MLE	0.032	0.0	3.232	0.0	23.071	0.0	-4356.901	0.242	-30671.727
DT	Lognorm	QME	0.062	0.0	9.9	0.0	71.999	0.0	-4760.032	1.0	-30597.404
DT	Lognorm	MLE	0.021	0.0	1.262	0.0	7.767	0.0	-4360.596	0.23	-30485.727
DT	Lognorm	Bayesian	0.02	0.0	1.197	0.0	7.441	0.0	-4362.802	0.238	-30485.745
DT	Weibull	MLE	0.052	0.0	8.611	0.0	66.155	0.0	-4588.294	1.0	-31021.667
DT	Weibull	QME	0.149	0.0	60.886	0.0	368.726	0.0	-3200.218	0.0	-31594.592
DT	Weibull	Bayesian	0.052	0.0	8.588	0.0	66.176	0.0	-4586.124	1.0	-31021.668

Table 13: Fitting statistics of the full distribution : Discus Throw

Pole Vault											
Event	**Model**	**Method**	**KS**	**p-KS**	**CvM**	**p-CvM**	**AD**	**p-AD**	**AD right**	**p-AD right**	**LogLik**
PV	Burr	Bayesian	0.082	0.0	10.24	0.0	72.856	0.0	-4568.697	0.952	-52029.966
PV	Burr	MLE	0.081	0.0	10.284	0.0	73.426	0.0	-4563.438	0.934	-52032.837
PV	Burr	QME	0.158	0.0	26.437	0.0	212.618	0.0	-3960.177	0.0	-52697.644
PV	Gamma	MLE	0.085	0.0	11.373	0.0	74.217	0.0	-4684.168	1.0	-51846.895
PV	Gamma	Bayesian	0.085	0.0	11.388	0.0	74.177	0.0	-4686.535	1.0	-51846.894
PV	Gamma	QME	0.124	0.0	30.164	0.0	161.526	0.0	-5124.043	1.0	-51898.239
PV	Gengamma	Bayesian	0.08	0.0	9.817	0.0	65.756	0.0	-4659.395	1.0	-51820.482
PV	Gengamma	MLE	0.074	0.0	8.616	0.0	67.734	0.0	-4576.999	0.97	-51934.332
PV	Invburr	QME	0.259	0.0	96.167	0.0	572.31	0.0	-3313.515	0.0	-54148.919
PV	Invburr	Bayesian	0.128	0.0	37.55	0.0	202.59	0.0	-5105.682	1.0	-52601.175
PV	Invburr	MLE	0.069	0.0	8.187	0.0	56.588	0.0	-4262.566	0.0	-51793.827
PV	Lognorm	MLE	0.078	0.0	8.408	0.0	56.549	0.0	-4598.92	0.978	-51715.308
PV	Lognorm	Bayesian	0.074	0.0	7.329	0.0	50.716	0.0	-4598.766	0.978	-51717.856
PV	Lognorm	QME	0.184	0.0	83.367	0.0	543.43	0.0	-5630.232	1.0	-52294.203
PV	Weibull	MLE	0.105	0.0	21.768	0.0	145.284	0.0	-4755.012	1.0	-52478.912
PV	Weibull	QME	0.159	0.0	26.442	0.0	212.426	0.0	-3989.41	0.0	-52708.337
PV	Weibull	Bayesian	0.109	0.0	18.478	0.0	136.475	0.0	-4633.441	0.998	-52481.748

Table 14: Fitting statistics of the full distribution : Pole Vault

Javelin Throw											
Event	**Model**	**Method**	**KS**	**p-KS**	**CvM**	**p-CvM**	**AD**	**p-AD**	**AD right**	**p-AD right**	**LogLik**
JT	Burr	QME	0.027	0.0	2.104	0.0	20.815	0.0	-5677.952	0.0	-43703.878
JT	Burr	Bayesian	0.019	0.002	0.539	0.024	4.028	0.014	-5914.558	0.738	-43626.917
JT	Gamma	Bayesian	0.016	0.004	0.533	0.034	2.725	0.04	-5852.563	0.34	-43585.48
JT	Gamma	MLE	0.016	0.004	0.525	0.03	2.71	0.042	-5855.797	0.364	-43585.479
JT	Gamma	QME	0.044	0.0	7.17	0.0	37.717	0.0	-5426.764	0.0	-43647.386
JT	Gengamma	MLE	0.022	0.0	1.364	0.0	8.721	0.0	-5705.467	0.004	-43617.595
JT	Gengamma	Bayesian	0.016	0.004	0.481	0.042	2.476	0.046	-5852.651	0.372	-43586.71
JT	Invburr	Bayesian	0.027	0.0	1.367	0.0	9.366	0.0	-5940.737	0.846	-43682.02
JT	Invburr	QME	0.091	0.0	22.099	0.0	173.692	0.0	-5273.757	0.0	-44371.867
JT	Invburr	MLE	0.024	0.0	1.282	0.002	8.846	0.0	-5898.691	0.67	-43681.372
JT	Lognorm	MLE	0.038	0.0	3.922	0.0	21.617	0.0	-5665.137	0.0	-43679.399
JT	Lognorm	QME	0.048	0.0	6.27	0.0	73.412	0.0	-6186.544	1.0	-43891.265
JT	Lognorm	Bayesian	0.038	0.0	3.905	0.0	21.593	0.0	-5667.697	0.0	-43679.423
JT	Weibull	MLE	0.045	0.0	8.499	0.0	63.932	0.0	-6061.356	0.998	-44062.683
JT	Weibull	QME	0.145	0.0	90.949	0.0	512.782	0.0	-4187.698	0.0	-44865.918
JT	Weibull	Bayesian	0.044	0.0	8.399	0.0	63.628	0.0	-6054.72	0.998	-44062.69

Table 15: Fitting statistics of the full distribution : Javelin Throw

1500m											
Event	**Model**	**Method**	**KS**	**p-KS**	**CvM**	**p-CvM**	**AD**	**p-AD**	**AD right**	**p-AD right**	**LogLik**
1500	Burr	QME	0.046	0.0	9.481	0.0	101.114	0.0	-10512.242	0.0	-19528.181
1500	Burr	Bayesian	0.022	0.0	3.219	0.0	25.671	0.0	-11223.923	0.952	-19187.281
1500	Burr	MLE	0.02	0.0	2.954	0.0	24.249	0.0	-11170.481	0.846	-19185.669
1500	Gamma	MLE	0.025	0.0	4.38	0.0	28.838	0.0	-10796.1	0.0	-19047.489
1500	Gamma	QME	0.149	0.0	260.278	0.0	1288.501	0.0	-6119.66	0.0	-20462.183
1500	Gamma	Bayesian	0.025	0.0	4.406	0.0	28.895	0.0	-10792.849	0.0	-19047.488
1500	Gengamma	MLE	0.012	0.002	1.115	0.0	8.877	0.0	-11077.763	0.484	-18948.543
1500	Gengamma	Bayesian	0.013	0.0	1.086	0.0	8.658	0.0	-11099.984	0.608	-18945.783
1500	Invburr	MLE	0.023	0.0	4.878	0.0	39.076	0.0	-11175.714	0.884	-19413.094
1500	Invburr	QME	0.133	0.0	96.474	0.0	665.148	0.0	-9266.868	0.0	-21791.37
1500	Invburr	Bayesian	0.028	0.0	6.268	0.0	45.282	0.0	-11221.932	0.942	-19416.939
1500	Lognorm	MLE	0.031	0.0	6.89	0.0	44.669	0.0	-10689.03	0.0	-19146.517
1500	Lognorm	QME	0.117	0.0	105.278	0.0	1043.138	0.0	-13090.77	1.0	-21649.77
1500	Lognorm	Bayesian	0.031	0.0	6.927	0.0	44.727	0.0	-10685.053	0.0	-19146.518
1500	Weibull	QME	0.067	0.0	26.714	0.0	219.888	0.0	-10032.661	0.0	-19884.621
1500	Weibull	Bayesian	0.034	0.0	10.413	0.0	77.165	0.0	-11498.752	1.0	-19517.408
1500	Weibull	MLE	0.034	0.0	10.414	0.0	77.151	0.0	-11499.027	1.0	-19517.408

Table 16: Fitting statistics of the full distribution : 1500m

Tail's Model Fit

100m				
Event	**Model**	**Method**	$\boldsymbol{\delta}_{\mathbf{score}}$	**p-value**
100	Burr	QME	0.002	0.015
100	Burr	Bayesian	0.004	0.069
100	Burr	MLE	0.004	0.126
100	Gamma	MLE	0.051	0.001
100	Gamma	QME	0.15	0.016
100	Gamma	Bayesian	0.052	0.001
100	Gengamma	Bayesian	0.009	0.118
100	Gengamma	MLE	0.006	0.083
100	Invburr	MLE	0.006	0.024
100	Invburr	QME	0.001	0.128
100	Invburr	Bayesian	0.005	0.028
100	Lognorm	MLE	0.083	0.0
100	Lognorm	QME	0.003	0.068
100	Lognorm	Bayesian	0.083	0.0
100	Weibull	MLE	0.02	0.021
100	Weibull	QME	0.011	0.064
100	Weibull	Bayesian	0.02	0.02

Long Jump				
Event	**Model**	**Method**	$\boldsymbol{\delta}_{\mathbf{score}}$	**p-value**
LJ	Burr	MLE	1125.657	0.091
LJ	Burr	QME	1.759	0.466
LJ	Burr	Bayesian	98.687	0.065
LJ	Gamma	QME	376.362	0.002
LJ	Gamma	Bayesian	119.883	0.0
LJ	Gamma	MLE	119.118	0.0
LJ	Gengamma	MLE	358.923	0.004
LJ	Gengamma	Bayesian	136.131	0.024
LJ	Invburr	Bayesian	564.78	0.064
LJ	Invburr	QME	14.988	0.172
LJ	Invburr	MLE	1895.806	0.003
LJ	Lognorm	MLE	29.168	0.08
LJ	Lognorm	QME	6.688	0.252
LJ	Lognorm	Bayesian	16.214	0.157
LJ	Weibull	MLE	647.924	0.02
LJ	Weibull	Bayesian	682.478	0.0
LJ	Weibull	QME	40.746	0.052

Table 17: Tail's Goodness of Fit : 100m and Long Jump

Shot Put				
Event	**Model**	**Method**	$\boldsymbol{\delta}_{\mathbf{score}}$	**p-value**
SP	Burr	QME	0.01	0.704
SP	Burr	Bayesian	0.461	0.142
SP	Burr	MLE	0.452	0.147
SP	Gamma	MLE	0.474	0.097
SP	Gamma	QME	0.051	0.17
SP	Gamma	Bayesian	0.461	0.095
SP	Gengamma	Bayesian	0.389	0.136
SP	Gengamma	MLE	0.417	0.061
SP	Invburr	MLE	1.989	0.073
SP	Invburr	QME	0.012	0.552
SP	Invburr	Bayesian	1.348	0.055
SP	Lognorm	MLE	0.077	0.406
SP	Lognorm	QME	0.021	0.267
SP	Lognorm	Bayesian	0.075	0.399
SP	Weibull	Bayesian	1.119	0.045
SP	Weibull	QME	0.171	0.153
SP	Weibull	MLE	1.123	0.048

High Jump				
Event	**Model**	**Method**	$\boldsymbol{\delta}_{\mathbf{score}}$	**p-value**
HJ	Burr	MLE	48.024	0.037
HJ	Burr	QME	0.23	0.48
HJ	Burr	Bayesian	33.676	0.094
HJ	Gamma	Bayesian	8.175	0.0
HJ	Gamma	QME	0.827	0.138
HJ	Gamma	MLE	7.947	0.0
HJ	Gengamma	MLE	23.982	0.0
HJ	Gengamma	Bayesian	12.044	0.0
HJ	Invburr	Bayesian	101.689	0.049
HJ	Invburr	QME	0.846	0.308
HJ	Invburr	MLE	100.117	0.072
HJ	Lognorm	Bayesian	1.893	0.02
HJ	Lognorm	QME	0.248	0.428
HJ	Lognorm	MLE	2.291	0.013
HJ	Weibull	Bayesian	45.066	0.048
HJ	Weibull	MLE	44.83	0.059
HJ	Weibull	QME	3.912	0.128

Table 18: Tail's Goodness of Fit : Shot Put and High Jump

400m				
Event	**Model**	**Method**	$\boldsymbol{\delta}_{\mathbf{score}}$	**p-value**
400	Burr	Bayesian	0.0	0.372
400	Burr	QME	0.0	0.506
400	Burr	MLE	0.0	0.441
400	Gamma	MLE	0.043	0.001
400	Gamma	Bayesian	0.043	0.001
400	Gamma	QME	0.125	0.009
400	Gengamma	MLE	0.0	0.295
400	Gengamma	Bayesian	0.001	0.247
400	Invburr	MLE	0.014	0.144
400	Invburr	Bayesian	0.01	0.054
400	Invburr	QME	0.001	0.08
400	Lognorm	QME	0.0	0.238
400	Lognorm	Bayesian	0.069	0.001
400	Lognorm	MLE	0.069	0.001
400	Weibull	QME	0.006	0.077
400	Weibull	Bayesian	0.007	0.082
400	Weibull	MLE	0.007	0.08

110m Hurdles				
Event	**Model**	**Method**	$\boldsymbol{\delta}_{\mathbf{score}}$	**p-value**
110h	Burr	MLE	0.018	0.221
110h	Burr	QME	0.0	0.66
110h	Burr	Bayesian	0.007	0.128
110h	Gamma	MLE	0.053	0.088
110h	Gamma	QME	0.007	0.228
110h	Gamma	Bayesian	0.052	0.085
110h	Gengamma	Bayesian	0.007	0.215
110h	Gengamma	MLE	0.005	0.256
110h	Invburr	Bayesian	0.048	0.103
110h	Invburr	QME	0.001	0.376
110h	Invburr	MLE	0.337	0.064
110h	Lognorm	Bayesian	0.101	0.08
110h	Lognorm	QME	0.004	0.268
110h	Lognorm	MLE	0.099	0.078
110h	Weibull	Bayesian	0.007	0.097
110h	Weibull	QME	0.0	0.625
110h	Weibull	MLE	0.007	0.095

Table 19: Tail's Goodness of Fit : 400m and 110m Hurdles

Discus Throw				
Event	**Model**	**Method**	$\boldsymbol{\delta}_{\mathbf{score}}$	**p-value**
DT	Burr	QME	0.032	0.578
DT	Burr	Bayesian	4.047	0.091
DT	Burr	MLE	18.86	0.1
DT	Gamma	Bayesian	1.504	0.176
DT	Gamma	QME	0.215	0.339
DT	Gamma	MLE	1.544	0.161
DT	Gengamma	Bayesian	0.387	0.345
DT	Gengamma	MLE	3.444	0.084
DT	Invburr	QME	0.043	0.578
DT	Invburr	MLE	104.358	0.053
DT	Invburr	Bayesian	56.851	0.065
DT	Lognorm	MLE	2.488	0.087
DT	Lognorm	QME	0.073	0.526
DT	Lognorm	Bayesian	2.698	0.083
DT	Weibull	MLE	10.902	0.084
DT	Weibull	QME	1.121	0.115
DT	Weibull	Bayesian	10.845	0.086

Pole vault				
Event	**Model**	**Method**	$\boldsymbol{\delta}_{\mathbf{score}}$	**p-value**
PV	Burr	Bayesian	3702.543	0.13
PV	Burr	MLE	3390.816	0.131
PV	Burr	QME	11.973	0.405
PV	Gamma	MLE	93.635	0.181
PV	Gamma	QME	78.338	0.218
PV	Gamma	Bayesian	93.007	0.186
PV	Gengamma	MLE	66.75	0.213
PV	Gengamma	Bayesian	124.829	0.09
PV	Invburr	MLE	20676.927	0.009
PV	Invburr	Bayesian	1422.65	0.058
PV	Invburr	QME	41.165	0.273
PV	Lognorm	Bayesian	669.467	0.071
PV	Lognorm	QME	131.899	0.203
PV	Lognorm	MLE	491.063	0.12
PV	Weibull	MLE	371.605	0.001
PV	Weibull	Bayesian	420.504	0.024
PV	Weibull	QME	9.948	0.589

Table 20: Tail's Goodness of Fit : Discus Throw and Pole Vault

Javelin Throw				
Event	**Model**	**Method**	$\boldsymbol{\delta}_{\mathbf{score}}$	**p-value**
JT	Burr	QME	0.08	0.559
JT	Burr	MLE	26.806	0.102
JT	Burr	Bayesian	0.469	0.589
JT	Gamma	Bayesian	1.334	0.211
JT	Gamma	MLE	1.333	0.214
JT	Gamma	QME	0.336	0.307
JT	Gengamma	MLE	3.88	0.154
JT	Gengamma	Bayesian	1.81	0.204
JT	Invburr	MLE	24.613	0.103
JT	Invburr	Bayesian	18.059	0.188
JT	Invburr	QME	0.295	0.333
JT	Lognorm	Bayesian	8.398	0.042
JT	Lognorm	QME	0.034	0.746
JT	Lognorm	MLE	7.975	0.05
JT	Weibull	MLE	18.151	0.005
JT	Weibull	Bayesian	18.184	0.004
JT	Weibull	QME	1.971	0.124

1500 m				
Event	**Model**	**Method**	$\boldsymbol{\delta}_{\mathbf{score}}$	**p-value**
1500	Burr	QME	0.0	0.477
1500	Burr	MLE	0.006	0.117
1500	Burr	Bayesian	0.004	0.108
1500	Gamma	QME	0.03	0.029
1500	Gamma	Bayesian	0.021	0.015
1500	Gamma	MLE	0.021	0.016
1500	Gengamma	Bayesian	0.002	0.146
1500	Gengamma	MLE	0.001	0.168
1500	Invburr	MLE	0.082	0.099
1500	Invburr	QME	0.0	0.163
1500	Invburr	Bayesian	0.058	0.004
1500	Lognorm	Bayesian	0.039	0.049
1500	Lognorm	QME	0.001	0.129
1500	Lognorm	MLE	0.038	0.049
1500	Weibull	Bayesian	0.006	0.052
1500	Weibull	QME	0.0	0.307
1500	Weibull	MLE	0.006	0.052

Table 21: Tail's Goodness of Fit : Javelin Throw and 1500m

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des Sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/sc