# Extreme value prediction: an application to sport records

Giovanni Fonseca[1] and Federica Giummolè[2]

[1] Dipartimento di Scienze Economiche e Statistiche,
Università di Udine
via Tomadini 30/A, I-33100 Udine, Italy.
giovanni.fonseca@uniud.it

[2] Dipartimento di Scienze Ambientali, Informatica e Statistica,
Università Ca' Foscari Venezia,
via Torino 155, I-30172 Venice, Italy.
giummole@unive.it

**Abstract**

Extreme value theory studies the extreme deviations from the central portion of a probability distribution. Results in this field have considerable importance in assessing the risk that characterises rare events, such as collapse of the stock market, or earthquakes of exceptional intensity, or floods. In the last years, application of extreme value theory for prediction of sport records have received increased interest by the scientific community. In this work we face the problem of constructing prediction limits for series of extreme values coming from sport data. We propose the use of a calibration procedure applied to the generalised extreme value distribution, in order to obtain a proper predictive distribution for future records. The calibrated procedure is applied to series of real data related to sport records. In particular, we consider sequences of annual maxima for different athletic events. Using the proposed calibrated predictive distribution, we show how to correctly predict the probability of future records and we discuss the existence and interpretation of ultimate records.

Keywords: athletic records, bootstrap, generalised extreme value distribution, prediction.

## 1 Introduction

From the very beginning, a big effort has been put into understanding the limits of human being capabilities: how fast can we run or swim? How far can we jump? In the last decades interest has mainly regarded the application of mathematical or statistical results in order to describe the progression of records in several sport events and in particular for track and field competitions.

Different approaches are used for assessing the probability of a new record or eventually the determination of an ultimate record, that is a measure that will not be overcome ever. [8] and [7] apply the theory of records to best annual performances. [10] propose a model for series of records, based on a random walk structure. In [11] a nonlinear regression model is introduced for fitting the progression of best annual results. Extreme value theory is applied in [9] to model the tail of the distribution for annual best records. [4] also takes advantage of the theory of extremes, enlarging the sample dimension by considering the personal best performance of as many athletes as possible over a period of several years.

In this work we apply the generalised extreme value (GEV) model to best annual results in the period from 2001 to 2018 for different athletic competitions. Depending on the data, the estimated model may comprise an end point that depends on the estimated parameters, or not. We propose a bootstrap procedure that allows the computation of a calibrated predictive distribution for best annual performances. The proposed predictive distribution works well in regular cases, i.e. when the estimated model is unbounded, but can also be useful when the end point of the support depends on the estimated parameters. Being calibrated, it allows to compute the correct probability of improving a world record in regular cases and to assess the quality of the endpoint of the estimated model in non regular cases.

The paper is organised as follows. In Section 2 we briefly describe the bootstrap calibrating procedure and in Section 3 we define the family of GEV distributions. In Section 4 we apply the proposed predictive procedure to athletic records.

## 2 Calibrated distributions for prediction

In this section we briefly review the calibrating approach proposed by [5], that provides a predictive distribution function whose quantiles give prediction limits with well-calibrated coverage probability.

Suppose that $\{Y_i\}_{i\geq 1}$ is a sequence of continuous random variables with probability distribution specified by the unknown $d$-dimensional parameter $\theta \in \Theta \subseteq \mathbf{R}^d$, $d \geq 1$; $Y = (Y_1, \cdots, Y_n)$, $n > 1$, is observable, while $Z = Y_{n+1}$ is a future or not yet available observation. For simplicity, we consider the case of $Y$ and $Z$ being independent random variables and we indicate with $G(z;\theta)$ the distribution function of $Z$.

Given the observed sample $y = (y_1, \ldots, y_n)$, an $\alpha$-prediction limit for $Z$ is a function $c_\alpha(y)$ such that, exactly or approximately,

$$P_{Y,Z}\{Z \leq c_\alpha(Y); \theta\} = \alpha, \tag{1}$$

for every $\theta \in \Theta$ and for any fixed $\alpha \in (0,1)$. The above probability is called coverage probability and it is calculated with respect to the joint distribution of $(Z, Y)$.

Consider the maximum likelihood estimator $\hat{\theta} = \hat{\theta}(Y)$ for $\theta$, or an asymptotically equivalent alternative, and the estimative prediction limit $z_\alpha(\hat{\theta})$, which is obtained as the $\alpha$-quantile of the estimative distribution function $G(\cdot;\hat{\theta})$. The associated coverage probability is

$$P_{Y,Z}\{Z \leq z_\alpha(\hat{\theta}(Y)); \theta\} = E_Y[G\{z_\alpha(\hat{\theta}(Y)); \theta\}; \theta] = C(\alpha, \theta) \tag{2}$$

and, although its explicit expression is rarely available, it is well-known that it does not match the target value $\alpha$ even if, asymptotically, $C(\alpha,\theta) = \alpha + O(n^{-1})$, as $n \to +\infty$, see e.g. [1]. As proved in [5], function

$$G_c(z; \hat{\theta}, \theta) = C\{G(z;\hat{\theta}), \theta\}, \tag{3}$$

which is obtained by substituting $\alpha$ with $G(z;\hat{\theta})$ in $C(\alpha,\theta)$, is a proper predictive distribution function, provided that $C(\cdot,\theta)$ is a sufficiently smooth function. Furthermore, it gives, as quantiles, prediction limits $z^c_\alpha(\hat{\theta},\theta)$ with coverage probability equal to the target nominal value $\alpha$, for all $\alpha \in (0,1)$.

The calibrated predictive distribution (3) is not useful in practice, since it depends on the unknown parameter $\theta$. However, a suitable parametric bootstrap estimator for $G_c(z;\hat{\theta},\theta)$ may be readily defined. Let $y^b$, $b = 1, \ldots, B$, be parametric bootstrap samples generated from the estimative distribution of the data and let $\hat{\theta}^b$, $b = 1, \ldots, B$, be the corresponding maximum likelihood estimates. Since $C(\alpha,\theta) = E_Y[G\{z_\alpha(\hat{\theta}(Y)); \theta\}; \theta]$, we define the bootstrap-calibrated predictive distribution as

$$G_c^b(z;\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B} G\{z_\alpha(\hat{\theta}^b);\hat{\theta}\}|_{\alpha=G(z;\hat{\theta})}. \tag{4}$$

The corresponding $\alpha$-quantile defines, for each $\alpha \in (0,1)$, a prediction limit having coverage probability equal to the target $\alpha$, with an error term which depends on the efficiency of the bootstrap simulation procedure.

## 3 Generalised extreme value distribution

The previous result can be applied, with some care, to the context of extreme value prediction. Indeed, assume that $\{X_t\}_{t\geq 1}$ is a discrete-time stochastic process with probability distribution specified by an unknown parameter. Furthermore, let $Y_i = \max_{k\in T_i} X_k$ be the maximum of the process over time interval $T_i$, $i \geq 1$. It is a well known result in extreme value theory that, under suitable conditions and if the number of observations in each period is big enough, the $Y_i$'s are approximately independent and with the same generalised extreme value (GEV) distribution; see for instance [2].

Now, assume that $Y = (Y_1, \cdots, Y_n)$, $n > 1$, is observable, while $Z = Y_{n+1}$ is a future or not yet available observation of the maximum of the process over the next time interval. Then $Y_1, \cdots, Y_n$ and $Z = Y_{n+1}$ can be considered as independent random variables with the same GEV distribution function

$$G(z;\mu,\sigma,\xi) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \tag{5}$$

with $z$ such that $1+\xi(z-\mu)/\sigma > 0$ and $\sigma > 0$.

The GEV distribution has three parameters: a location parameter $\mu$, a scale parameter $\sigma$ and a shape parameter $\xi$. In particular, the values of $\xi$ determine the type of GEV distribution:

- $\xi \to 0$ corresponds to the Gumbel distribution (type $I$);
- $\xi > 0$ corresponds to the Fréchet distribution (type $II$);
- $\xi < 0$ corresponds to the (negative) Weibull distribution (type $III$).

It is important noticing that when $\xi > 0$ or $\xi \to 0$ the support of the distribution is not limited from above. Only in the case when $\xi < 0$ the support has an upper bound equal to $\mu - \sigma/\xi$.

Inverting (5) we can achieve an explicit expression for the quantiles of the distribution:

$$z_\alpha = z_\alpha(\mu,\sigma,\xi) = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1-\{-\log(\alpha)\}^{-\xi}\right] & \text{if } \xi \neq 0 \\ \mu - \sigma\log\{-\log(\alpha)\} & \text{if } \xi = 0 \end{cases} \tag{6}$$

with $G(z_\alpha;\mu,\sigma,\xi) = \alpha$. The value $z_\alpha$ is also called return level and it indicates the value that is expected to be exceeded on average once every $1/(1-\alpha)$ time intervals.

# 4 An application to athletic records

In this section we apply the calibrated procedure to athletic records for two different purposes. First, we estimate the probability of observing a new record in the next year and we predict the expected time for a future record. This also allows to evaluate the goodness of a current world record. Secondly, we discuss the existence of the ultimate record and we give the correct interpretation to its estimate.

We have collected data from the web site of the International Association of Athletics Federations (IAAF) [6]. Starting from 2001, we have registered the annual records for males and females, for the following events: 100 m, 200 m, 400 m, 10,000 m, long jump and javelin. We have transformed times into mean speeds so that, for each event, the higher the best.

## 4.1 Parameter estimation

The first step consists of estimating the unknown parameters of each distribution. In particular, the estimates obtained for the shape parameters are very important because they determine the particular distribution to be used inside the GEV family.

Here we consider three different methods of estimation: maximum likelihood, L-moments and generalised maximum likelihood. In spite of its optimal asymptotic properties, the method of maximum likelihood does not perform very well for small sample sizes. Instead L-moments and generalised maximum likelihood estimates ensure a better fit, especially for the shape parameter. In particular, estimation based on the generalised maximum likelihood retains large sample properties of the maximum likelihood but improves on its small sample performance (see, for instance, [3]).

Table 1 and Table 2 show estimates for the shape parameters $\xi$ obtained using the three different estimating methods, for men and women, respectively. The first row of each table reports maximum likelihood estimates (mle), the second row contains estimates obtained by the method of L-moments (Lmom) and the third row is for generalized maximum likelihood (gmle). All the estimated values for the shape parameters are negative, with an exception for women long jump, for which the three estimates are positive. Thus, the corresponding estimative distributions are reverse Weibull distributions for all events with negative shape parameter and a Fréchet distribution for women long jump. In the sequel we will use generalised maximum likelihood estimates.

| estimate | 100 m | 200 m | 400 m | 10,000 m | long jump | javelin |
|---|---|---|---|---|---|---|
| mle | -0.1618 | -0.0826 | -0.1781 | -0.2157 | -0.3984 | -0.3231 |
| Lmom | -0.1281 | -0.0553 | -0.2246 | -0.0819 | -0.3104 | -0.3755 |
| gmle | -0.1281 | -0.0553 | -0.2246 | -0.0819 | -0.3104 | -0.3755 |

Table 1: Men: estimates of the shape parameters for different events

| estimate | 100 m | 200 m | 400 m | 10,000 m | long jump | javelin |
|---|---|---|---|---|---|---|
| mle | -0.3343 | -0.4416 | -0.1658 | -0.1269 | 0.1116 | -0.3033 |
| Lmom | -0.3069 | -0.3006 | -0.1330 | -0.1803 | 0.1126 | -0.1864 |
| gmle | -0.3069 | -0.3006 | -0.1330 | -0.1803 | 0.3311 | -0.1864 |

Table 2: Women: estimates of the shape parameters for different events

## 4.2 Prediction

In this section we compare the estimative distribution function obtained from the generalised maximum likelihood estimator with the bootstrap calibrated one, for each of the considered events.

We will see that, using the bootstrap calibrated predictive distribution, we can properly calculate probabilities related to the variable $Z$ which represents the best performance in the year to come. In particular we can predict the probability of having a new world record in the next year as $\alpha_{WR} = P(Z > WR)$, where $WR$ represents the present world record. This probability can also be used to evaluate the goodness of the world record. Moreover, from $\alpha_{WR}$ we can calculate the expected number of years for the next record, $T_{WR} = 1/\alpha_{WR}$.

In all the cases when the estimate of the shape parameter $\xi$ is negative, the estimative distribution is a (negative) Weibull distribution. It has a bounded upper tail at $UL = \hat{\mu} - \hat{\sigma}/\hat{\xi}$. In the analysis of sport data $UL$ is the estimate of what is called the ultimate record, which is a value that cannot be exceeded by any performance. Instead, using the calibrated predictive distribution we can show that the probability of exceeding $UL$, $\alpha_{UL} = P(Z > UL)$, is different from 0. Unfortunately, in non regular cases when the support of the distribution depends on unknown parameters, formula (4) is only useful for calculating the bootstrap calibrated predictive distribution inside the estimated domain. As a consequence, when the present world record exceeds the estimated upper bound, we cannot calculate $\alpha_{WR}$. This drawback is not present for women long jump, since in this case the estimated shape parameter is positive, giving rise to a Fréchet estimative distribution whose upper tail is unbounded.

Table 3 and Table 4 summarise the main results obtained for each considered event. In particular they report for men and women, respectively: the estimate of the ultimate record $UL$, the probability $\alpha_{UL}$ of exceeding that estimate, the present world record $WR$, the probability $\alpha_{WR}$ of exceeding it and the expected time $T_{WR}$ for improving it.

| | 100 m | 200 m | 400 m | 10,000 m | long jump | javelin |
|---|---|---|---|---|---|---|
| $UL$ | 10.797 | 12.052 | 9.431 | 6.804 | 8.871 | 95.364 |
| $\alpha_{UL}$ | 0.009 | 0.008 | 0.011 | 0.008 | 0.011 | 0.013 |
| $WR$ | 10.438 | 10.422 | 9.296 | 6.339 | 8.95* | 98.48* |
| $\alpha_{WR}$ | 0.031 | 0.057 | 0.029 | 0.054 | - | - |
| $T_{WR}$ | 31.79 | 17.51 | 33.91 | 18.62 | - | - |

Table 3: Men's summary results. * means that the corresponding world record is not included in the data.

| | 100 m | 200 m | 400 m | 10,000 m | long jump | javelin |
|---|---|---|---|---|---|---|
| $UL$ | 9.477 | 9.368 | 8.449 | 5.897 | - | 78.318 |
| $\alpha_{UL}$ | 0.012 | 0.011 | 0.009 | 0.010 | - | 0.010 |
| $WR$ | 9.533* | 9.372* | 8.403* | 5.690 | 7.52* | 72.28 |
| $\alpha_{WR}$ | - | - | 0.009 | 0.028 | 0.056 | 0.084 |
| $T_{WR}$ | - | - | 105.55 | 36.05 | 17.93 | 11.83 |

Table 4: Women's summary results. * means that the corresponding world record is not included in the data.

Three different possible situations are illustrated and commented using data from men's 400 m, women's 100 m and women's long jump.

#### 4.2.1 Men's 400 m

Figure 1 shows the estimative (red dashed) and bootstrap calibrated (black solid) distribution functions for men's 400 m data. The bootstrap procedure is based on 5000 replications. The present world record (blue dash-dotted) and the estimated ultimate record (red dotted) are also represented. Here the original time data (sec) have been transformed into mean speeds (m/sec) since the GEV model fits to maxima data.

For the transformed data, the estimate of the shape parameter is negative. This implies that the estimative distribution function is a reverse Weibull distribution with upper bound $UL = \hat{\mu} - \hat{\sigma}/\hat{\xi} = 9.431$ m/sec. This corresponds to a time of 42.41 sec. The value $UL$ is usually interpreted as an estimate of the ultimate record which is the best possible performance in the event. It is important noticing that this is just an estimate and, of course, it is subject to variability. To account for this variability, we can use the calibrated predictive distribution and correctly predict the probability of exceeding $UL$. As one can see in the plot, in fact, this probability is the difference between the value 1 of the estimative distribution at $UL$ and the value of the bootstrap calibrated distribution at $UL$. Thus, $\alpha_{UL} = P(Z > UL) = 0.011$. Similarly, we can calculate the probability of improving the present world record of 43.03 sec, $WR = 9.296$ m/sec, as $\alpha_{WR} = P(Z > WR) = 0.029$, meaning that we expect to improve the present world record about 3 times every 100 years. This can also be taken as a measure of goodness of a world record. Both probabilities $\alpha_{UL}$ and $\alpha_{WR}$ are wrongly underestimated by the estimative distribution function because in the estimation procedure the true parameters are substituted by their estimates without taking into account for the additional uncertainty introduced. In particular, the estimative distribution underestimates to 0.016 the probability of improving the current world record.

#### 4.2.2 Women's 100 m

Figure 2 shows the estimative (red dashed) and bootstrap calibrated (black solid) distribution functions for women's 100 m data. The bootstrap procedure is based on 5000 replications. The present world record (blue dash-dotted) and the estimated ultimate record (red dotted) are also represented. As in the previous example, the original time data (sec) have been transformed into mean speeds (m/sec) since the GEV model fits to maxima data.

For the transformed data, the estimate of the shape parameter is negative, thus the estimative distribution function is a reverse Weibull distribution with upper bound $UL = \hat{\mu} - \hat{\sigma}/\hat{\xi} = 9.477$ m/sec. This corresponds to a time of 10.55 sec. We can use the calibrated predictive distribution to correctly predict the probability of exceeding $UL$. As one can see in the plot, in fact, this probability is the difference between the value 1 of the estimative distribution at $UL$ and the value of the bootstrap calibrated distribution at $UL$. Thus, $\alpha_{UL} = P(Z > UL) = 0.012$. In this example, the present world record $WR = 9.533$ m/sec (10.49 sec) exceeds the upper limit $UL$, as can be seen in figure 2. This may occur when the data used for estimation do not include the world record. Indeed, the present world record dates back to 1988, while we have considered data from 2001 to 2018. A methodological problem arises in this situation, since we are not able to calculate the values of the bootstrap calibrated predictive distribution (4) in points that exceed the upper bound of the estimative distribution. The upper tail of the calibrated predictive distribution can be estimated using non linear regression, but this issue requires further research. At the moment, we can only conclude by saying that the probability of improving the present world record is $\alpha_{WR} = P(Z > WR) < P(Z > UL) = 0.012$. Actually, the present world record seems to be an exceptional result that can be hardly improved at the moment.

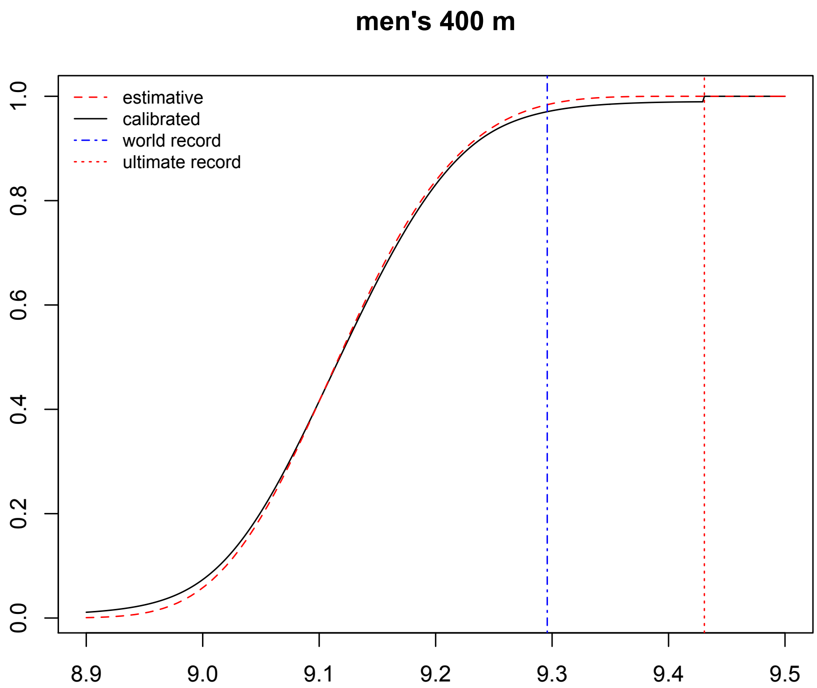


Figure 1: Men's 400 m. Plot of estimative (red dashed) and bootstrap calibrated (black solid) distribution functions for men's 400 m data. Bootstrap procedure is based on 5000 replications. World record (blue dash-dotted) and estimated ultimate record (red dotted) are also represented.

#### 4.2.3 Women's long jump

Figure 3 shows the estimative (red dashed) and bootstrap calibrated (black solid) distribution functions for women's long jump data. The bootstrap procedure is based on 5000 replications. The present world record (blue dash-dotted) is also represented.

This is the only event for which the estimate of the shape parameter of the GEV distribution is positive, thus the estimative distribution function is a Fréchet distribution with no upper bound. The present world record, $WR = 7.52$ m, dates back to 1988 and is not included in the data. Anyway, this is not a problem, being the upper bound $UL = +\infty$. Using the bootstrap calibrated distribution, we can predict the probability of improving the present world record: $\alpha_{WR} = P(Z > WR) = 0.056$. Notice that the estimative distribution wrongly underestimates this probability to 0.040. The expected time for improving the current world record is about 18 years.

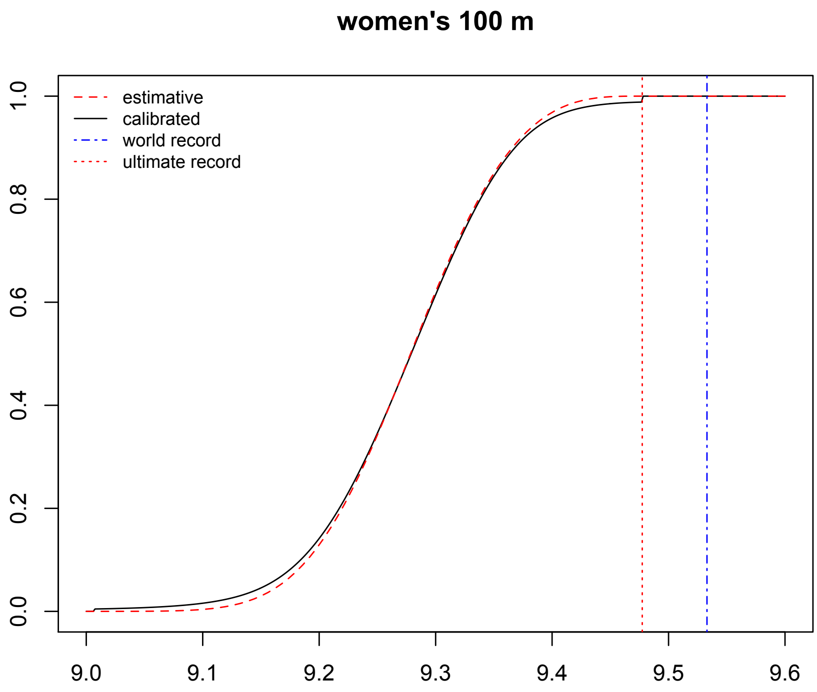


Figure 2: Women's 100 m. Plot of estimative (red dashed) and bootstrap calibrated (black solid) distribution functions for women's 100 m data. Bootstrap procedure is based on 5000 replications. World record (blue dash-dotted) and estimated ultimate record (red dotted) are also represented.

## Acknowledgments

This research is partially supported by the Italian Ministry for University and Research under the PRIN2015 grant No. 2015EASZFS 003.

## References

[1] O.E. Barndorff-Nielsen and D.R. Cox. Prediction and asymptotics. *Bernoulli*, 2:319–340, 1996.

[2] S. Coles. *An introduction to statistical modeling of extreme values.* Springer-Verlag, London, 2001.

[3] S.J. Coles and M.J. Dixon. Likelihood-based inference for extreme value models. *Extremes*, 2(1):5–23, 1999.

[4] J.H.J. Einmahl and J.R. Magnus. Records in athletics through extreme-value theory. *Journal of the American Statistical Association*, 103:1382–1391, 2008.

[5] G. Fonseca, F. Giummolè, and P. Vidoni. Calibrating predictive distributions. *Journal of Statistical Computation and Simulation*, 84:373–383, 2014.

[6] IAAF. International Association of Athletics Federations. *https://www.iaaf.org/home.*

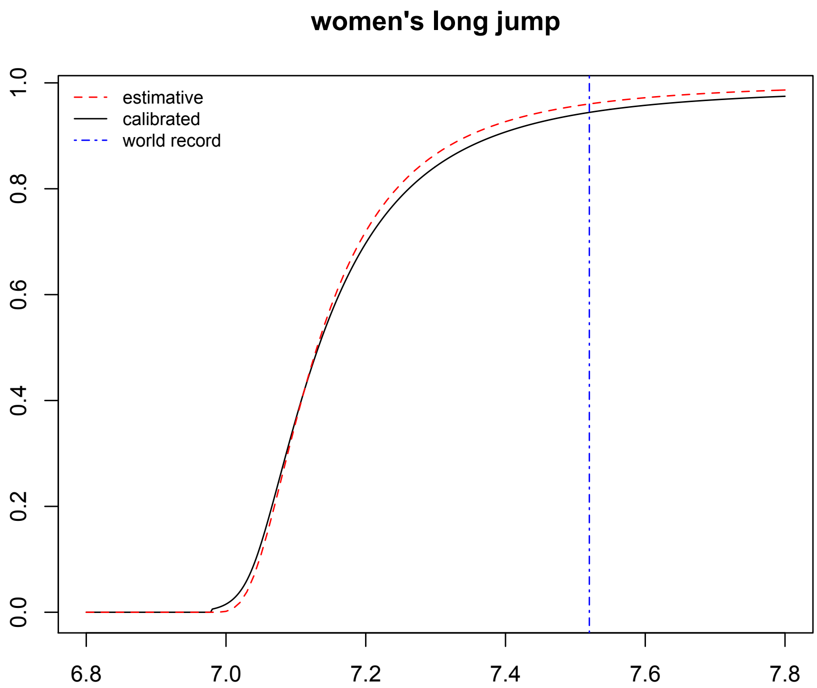


Figure 3: Women's long jump. Plot of estimative (red dashed) and bootstrap calibrated (black solid) distribution functions for women's long jump data. Bootstrap procedure is based on 5000 replications. World record is also represented (blue dash-dotted). Since the estimative density is not bounded from above, the ultimate record is $+\infty$.

[7] F. Noubary and R. Noubary. Survival analysis of the men's 100 meter dash record. *Applications and Applied Mathematics*, 11(1):115–126, 2016.

[8] R.D. Noubary. A procedure for prediction of sports records. *Journal of Quantitative Analysis in Sports*, 1(1):Article 4, 2005.

[9] R.D. Noubary. Tail modeling, track and field records, and Bolt's effect. *Journal of Quantitative Analysis in Sports*, 6(3), 2010.

[10] J.T. Terpstra and N.D. Schauer. A simple random walk model for predicting track and field world record. *Journal of Quantitative Analysis in Sports*, 3(3):Article 4, 2007.

[11] P. Volf. A stochastic model of progression of athletic records. *Journal of Management Mathematics*, 22:157–169, 2011.