Improved prediction of 100-meter sprint records

Giovanni Fonseca[1], Federica Giummolè[2],
Michele Lambardi di San Miniato[1], Valentina Mameli[1]
[1] *University of Udine*
[2] *Ca' Foscari University of Venice*

Abstract

In the last years, prediction of sport records has received increased attention by the scientific community. In particular, it is of great interest the evaluation of the goodness of a record. The application of extreme value theory in this context is quite natural. In this work, we use the Gumbel model to analyze the annual speed records in men's and women's 100-meter sprint races from 2001 to 2024. We propose the use of a new calibration procedure in order to correctly estimate the probability of future records and the expected time needed to break the current world record.
Keywords: predictive distribution, estimative distribution, bootstrap, athletic records

1 Introduction

The study of human capabilities has been a topic of interest since ancient times, particularly in sports. Mathematical models and statistical analyses have been used to describe the peak performances in various sports events. In athletics, the most debated aspects are the probability of breaking the record and, relatedly, the expected time to break the record. Another aspect is the magnitude, if not the very existence, of an ultimate record, that is an upper limit to human performances.

In sports context, technologies, rules and leading athletes can rapidly change over time, so it seems reasonable to assume that only the most recent records provide relevant information on the future. As a consequence, statistical analysis involves small sample sizes. The classical estimative approach to prediction consists in replacing any unknown model parameters with suitable efficient estimates. Unfortunately, small sample sizes produce unstable estimates, making the estimative approach unreliable.

In this work, in order to overcome this problem, we apply the bootstrap-based calibration method proposed in [1] to obtain an improved predictive distribution for future records. The resulting predictive distribution performs better than the estimative one since it provides more valid prediction of probabilities. Here, we present an analysis of the annual records in men's and women's 100-meter sprint races from 2001 to 2024 based on the Gumbel distribution.

2 Model and prediction methods

Let $(Y_i)_{i\geq 1}$ denote a sequence of i.i.d. continuous random variables, whose density is indexed by a d-dimensional parameter $\theta \in \Theta \subseteq \mathbf{R}^d$. Let $Y^{(n)} = (Y_1, \cdots, Y_n)$ be a n-sample and let $Z = Y_{n+1}$ be the prediction target. The distribution function of Z is denoted by $F(z;\theta)$, while its inverse is called the quantile function and is denoted by $Q(\alpha;\theta)$.

Let the Y_i's be the best annual performances in a sport discipline. Under suitable conditions, according to extreme value theory, best annual records follow a Generalized Extreme Value (GEV) distribution. In this work, supported by a preliminary data analysis, we consider a special case of the GEV model, i.e. the Gumbel model. The Gumbel distribution function is defined as

$$F(z;\theta) = \exp\left\{-\exp\left(-\frac{z-\mu}{\sigma}\right)\right\}, \quad z \in R, \tag{1}$$

with $\theta = (\mu, \sigma)$, $\mu \in R$, $\sigma > 0$. The Gumbel model has an unbounded support, thus excluding the existence of an ultimate record.

2.1 Prediction methods

The ideal predictive distribution is $F(z;\theta)$, which is unavailable since θ is unknown. As previously said, the estimative approach proposes to replace θ with an efficient estimate $\hat{\theta}_n$ by plug-in, so,

$$\hat{F}_n(z) = F(z;\hat{\theta}_n), \quad \hat{Q}_n(\alpha) = Q(\alpha;\hat{\theta}_n)\,. \tag{2}$$

Unfortunately, both $\hat{F}_n$ and $\hat{Q}_n$ can perform poorly, especially for small samples, because they do not take into account the variability introduced by substituting an unknown constant θ with an estimate.

Using the bootstrap approach, [2] develop a predictive distribution with better coverage properties. This improved predictive distribution, denoted by $\hat{F}_b$, is defined as

$$\hat{F}_b(z) = \frac{1}{B}\sum_{b=1}^{B} \hat{F}_n(Q(\hat{F}_n(z);\hat{\theta}_n^b)), \tag{3}$$

where $(\hat{\theta}_n^b)_{b=1}^B$ are parametric bootstrap estimates of θ, generated under $\theta = \hat{\theta}_n$. $\hat{F}_b(z)$ improves on the estimative solution in terms of coverage probabilities of corresponding quantiles. It thus can be useful when the focus is on quantiles of the true distribution function. Similarly, [1] propose an improved quantile function in the following form:

$$\hat{Q}_b(\alpha) = \frac{1}{B}\sum_{b=1}^{B} \hat{Q}_n(F(\hat{Q}_n(\alpha);\hat{\theta}_n^b))\,. \tag{4}$$

$\hat{Q}_b$ in (4) and its inverse are meant to improve on the estimative approach as regards the prediction of probabilities. Thus, they can be employed for predicting probabilities of overcoming a certain fixed limit. In particular, the inverse of $\hat{Q}_b$ could be fruitfully used to predict the probability $\mathsf{PBR} = 1 - F(\mathsf{WR};\theta)$ of breaking the current world record (WR) within the next year.

3 Application to athletics data

Here, we apply the estimative distribution and the two proposed bootstrap calibrated distributions, to world athletic records. In particular, we consider the annual speed records in the men's and women's 100-meter sprint races from 2001 to 2024. Figure 1 shows the histogram of the data (grey) and the estimative density under the Gumbel model (red). In this application, we consider the maximum likelihood estimator. The considered predictive distribution functions are the estimative one (E), the improved distribution function in equation (3) (FB), and the inverse of the improved quantile function in equation (4) (QB). The bootstrap procedures are based on 1000 replications. Table 1 reports the expected time (in years) to break the WR, ETBR = 1/PBR. Figure 2 shows the upper tails of the three predictive distributions, namely, the estimative (black), and the bootstrap-calibrated FB (red) and QB (blue).

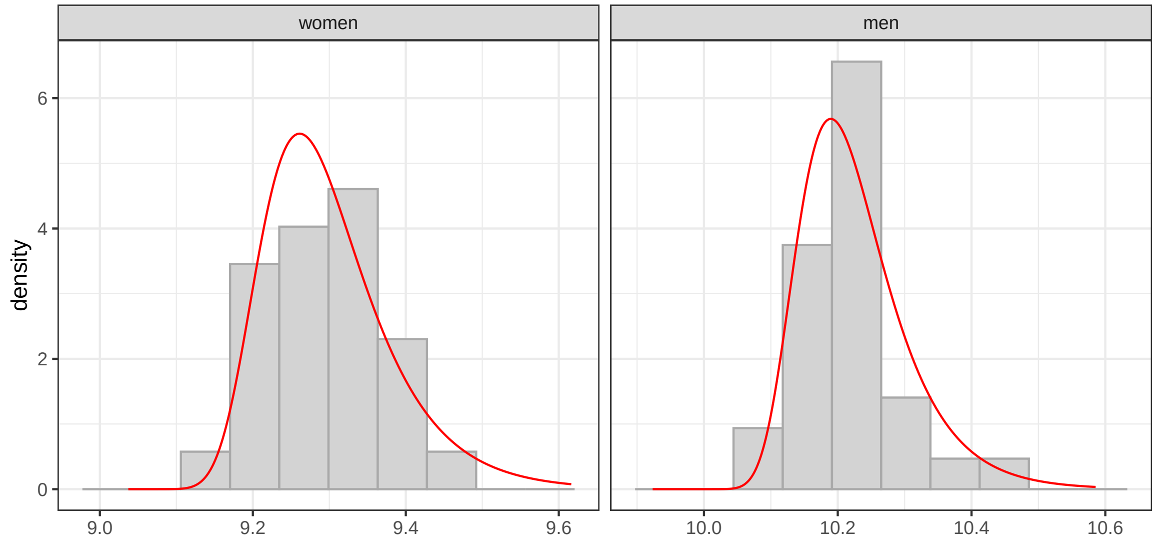

Figure 1: Annual record speeds at the men's and women's 100-meter sprint races, in m/s, 2001-2024. Histogram of the data (grey), along with the estimative Gumbel density (red).

Table 1: WR and estimates of ETBR in years with estimative and improved approaches.

	WR, m/s	set by	year	E	FB	QB
women	9.533	U. Bolt	1988	56.696	41.094	45.797
men	10.438	F. Griffith-Joyner	2009	47.127	36.092	39.647

Both men's and women's WR look statistically hard to break, with PBR ranging between 0.017 and 0.028. However, the PBR obtained with the estimative approach is the lowest and very different from those obtained via improved methods, so the latter should be more trusted. The FB is more reliable than the estimative method in terms of quantiles. However, QB should be preferred for the PBR since it is targeted at probabilities.

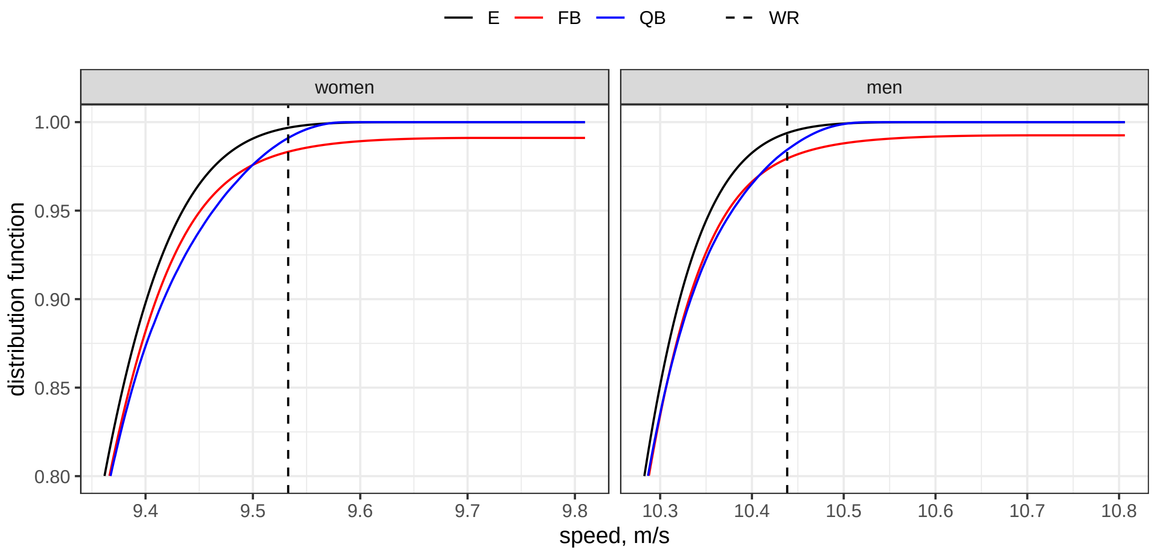

Figure 2: Estimative E (black) and improved FB (red) and QB (blue) distribution functions. WR are reported as dashed lines.

Acknowledgments

This research is funded by PRIN 2022 - Project prot. n. 2022R74PLE-UGOV code PRIN 2022 MAMELI DIES - CUP G53D23001870006 funded by the European Union NextGenerationEU M4C2 inv 1.1.

References

[1] Fonseca G., Giummolè F., and Vidoni P. Optimal prediction for quantiles and probabilities. *Statistical Papers*, 66:24, 2025.

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