Extreme Value Theory and Sports: the Maximal Oxygen Uptake

S. Vicente, M. I. Fraga Alves and M. I. Gomes

CEAUL and DEIO, Faculty of Sciences, University of Lisbon

Abstract

The parameters of the Generalized Extreme Value and of the Generalized Pareto distributions constitute a key for estimating quantities such as the right endpoint of the underlying distribution function F and the probability of exceeding the current sample maximum. This work exemplifies an application of such estimation to a well-known variable of the world of Sports and Physiology: the Maximal Oxygen Uptake or $\dot{V}O_2max$. Following a classical parametric approach with the Peaks Over Threshold method and a semi-parametric approach in a modest sample size context, a relative proximity between the two approaches is notorious.

1 Introduction

1.1 Extreme Value Theory

The statistical analysis of extreme large (or small) values, typical of areas such as Hydrology, Finance or Environment, finds all the necessary tools in Extreme Value Theory (EVT). In particular, let $(X_1, X_2, \ldots, X_n)$ be a random sample of the random variable (r.v.) X with distribution function (d.f.) F and let $M_n = \max(X_1, X_2, \ldots, X_n)$ denote the respective *sample maximum*. The non-degenerate limiting d.f. of a sequence of suitably normalized sample maxima has been established by the works of R. Fisher and L. Tippett, later completed by B. Gnedenko. See Fisher and Tippett [1] and Gnedenko [2]. They settled on the **Generalized Extreme Value distribution** (GEVd) as the unified version of all the possible non-degenerate weak limits of a suitably normalized sequence of sample maxima. The GEVd is defined as follows:

$$G(x|\gamma) = G_\gamma(x) = \begin{cases} \exp\left(-(1+\gamma x)^{-\frac{1}{\gamma}}\right), & 1+\gamma x > 0 \quad \text{if } \gamma \neq 0, \\ \exp(-\exp(-x)), & x \in \mathbb{R} \quad \text{if } \gamma = 0, \end{cases} \tag{1}$$

where the shape parameter γ is known as the ***extreme value index*** (EVI). For $\gamma < 0$, $\gamma = 0$ and $\gamma > 0$, the GEVd reduces to Weibull, Gumbel and Fréchet d.f.s, respectively. Therefore, under necessary and sufficient conditions, there exist normalizing sequences $a_n > 0$ and $b_n \in \mathbb{R}$ such that, for every continuity point x of G_γ,

$$\lim_{n\to\infty} P(M_n \leq a_n x + b_n) = \lim_{n\to\infty} F^n(a_n x + b_n) = G_\gamma(x) \tag{2}$$

and we say that F belongs to the max-domain of attraction of G_γ, which will be denoted by $F \in \mathcal{D}(G_\gamma)$. A more general version of the GEVd defined in (1) can be obtained incorporating a location parameter, $\lambda \in \mathbb{R}$, and a scale parameter, $\delta > 0$:

$$G_\gamma(x|\lambda, \delta) = G_\gamma\left(\frac{x-\lambda}{\delta}\right), \quad \lambda \in \mathbb{R}, \delta > 0. \tag{3}$$

Following the contributions of B. Gnedenko and looking for an appropriate model for the right tail of the underlying d.f. F, L. de Haan, A. Balkema and J. Pickands introduced the **Generalized Pareto distribution** (GPd) as the limiting d.f. for the conditional d.f. of suitably normalized excesses, $Y_i = X_i - u | X_i > u$, $i = 1, 2, \ldots, n_u$, over a sufficiently high threshold u; that is, $F \in \mathcal{D}(G_\gamma)$ if and only if $F_{X|X>u}(x) \simeq H_\gamma(x|u, \sigma_u)$, where

$$H_\gamma(x|u,\sigma_u) = \begin{cases} 1-(1+\gamma\frac{x-u}{\sigma_u})^{-1/\gamma}, & 1+\gamma\frac{x-u}{\sigma_u} > 0, x \geq u, \quad \text{if } \gamma \neq 0, \\ 1-\exp(-\frac{x-u}{\sigma_u}), & x \geq u, \quad \text{if } \gamma = 0, \end{cases} \tag{4}$$

stands for the GPd, with shape, location and scale parameters γ, u and $\sigma_u > 0$, respectively. For $\gamma < 0$, $\gamma = 0$ and $\gamma > 0$, the GPd reduces to Beta, Exponential and Pareto d.f.s, respectively. See Balkema and de Haan [3] and Pickands [4].

In both distributions defined in (3) and (4), the EVI is directly related to the right tail's weight of the underlying d.f. F: $\gamma < 0$ refers to light right-tailed d.f.s with finite right endpoint, $x^F = \sup\{x : F(x) < 1\}$, $\gamma = 0$ encompasses exponential right-tailed d.f.s with finite or infinite right endpoint and $\gamma > 0$ evidences heavy right-tailed d.f.s with infinite right endpoint.

1.2 The Maximal Oxygen Uptake or $\dot{V}O_2max$

EVT reveals to be a very useful tool in Sports, with extreme events such as minimum time, maximum speed, maximum length and maximum height, which characterize most categories. The success of the athletes with such variables is directly related to their cardiorespiratory capacity, which can be monitored by the **Maximal Oxygen Uptake** or $\dot{V}O_2max$. The $\dot{V}O_2max$ represents the maximum quantity of oxygen that can be assimilated and used by the human body during an intense effort, measured in milliliters per kilogram of bodyweight and per minute (ml/kg/min). The maintenance of a high $\dot{V}O_2max$ level is then a key for a high cardiorespiratory capacity. In this work, a sample of 74 male athletes picked up among cross-country skiers, runners and road racing cyclists was analyzed. For each athlete, the highest $\dot{V}O_2max$ was selected, resulting in a sample of size 74. The sample includes the current world's record of 96 ml/kg/min, hold by the Norwegian cross-country skiers, Bjørn Dæhlie and Espen Harald Bjerke.

2 Estimation of parameters

The parameters in (3) and (4) are crucial for the estimation of quantities of interest in EVT: the location and scale constants b_n and a_n, respectively, defined in (2), the right endpoint x^F for the case $\gamma \leq 0$ and the probability of exceeding the current world record of 96 ml/kg/min. The estimations were obtained by two approaches: a parametric approach and a semi-parametric approach.

2.1 Parametric analysis: the Peaks Over Threshold (POT) method

Following a classical parametric approach, the main assumption is that we can use the limiting distributions defined in (3) and (4) as an exact model that can be fitted to the data in hand, using point estimation methods, such that the Maximum Likelihood (ML) method and the Probability Weighted Moments (PWM) method. Since the limiting distributions are piecewise-defined distributions, the choice of a model to be fitted to the data can be improved if we make *a priori* assumptions about the sign of γ. Therefore, the following hypothesis test was used in order to choose the "best" model for the data:

$$H_0 : \gamma = 0 \quad vs \quad H_1 : \gamma \neq 0. \tag{5}$$

The Peaks Over Threshold (POT) method considers the data in hand as a collection obtained from the right tail of the underlying d.f. F, where, above a fixed high threshold u, the GPd defined in (4) can be fitted to the excesses $Y_i = X_i - u | X_i > u$, $i = 1, 2, \ldots, n_u$. After finding a concordance between the preliminary statistical analysis of the data by means of the sample Mean Excess plot and the literature in the Physiology area, we chose a threshold of $u = 80$ (i.e. $\hat{b}_n = 80$) and obtained $m = 49$ excesses above 80ml/kg/min. Moreover, the preliminary statistical analysis with a Quantile-Quantile plot (QQ-plot) and convenient test statistics for the test defined in (5) pointed clearly to the GPd with $\gamma < 0$ as a suitable model for the excesses. See Vicente [5] for details. The following results were obtained using the ML (PWM) method: $\hat{\gamma} = -0.727$ (-0.509); $\hat{\sigma}_u \equiv \hat{a}_n = 12.045$ (10.025); $\hat{P}(X > 96) = 0.006$ (0.025) and $\hat{x}^F = 96.573$ (99.708).

2.2 Semi-parametric analysis

Following a semi-parametric approach, there is no fit of a specific parametric model to the data. No assumption is then made about the global form of the underlying d.f. F. The only assumption is that the d.f. F belongs to the max-domain of attraction of some extreme value distribution, i.e., $F \in \mathcal{D}(G_\gamma)$. The inference is then based on the $k+1$ top order statistics defined from a random sample point, $X_{n-k:n}$, where the determination of k plays a central role. Using convenient semi-parametric test statistics, the Weibull max-domain of attraction was elected and the value of k was determined by the following heuristic process: $k^{opt} = \arg\min_k \sum_{i \neq j} \left(\hat{\gamma}^{(i)}_{n,k} - \hat{\gamma}^{(j)}_{n,k} \right)^2$, where $\hat{\gamma}^{(i)}$ stands for any semi-parametric estimator of the EVI valid for $\gamma < 0$. The plot of this function depicted in Figure 1 shows an optimal choice of $k = 19$, but in a highly volatile region. A better choice of $k = 43$ was selected in a more stable region. See Vicente [5] for details. This latter choice of k produced the estimates $\hat{b}_n = 81.125$ and $\hat{a}_n = 9.956$ for the attraction coefficients defined in (2) and the estimates of the remaining parameters can be found in Table 1. A clear proximity with the POT approach is evident, namely for the PWM method, reputed to perform better in a modest sample size context. This proximity evidences how the semi-parametric approach can be viewed as a parametric approach of the right tail of the underlying d.f. F.

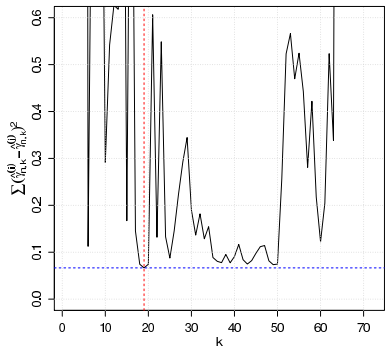

Figure 1: Plot of $\sum_{i \neq j} \left(\hat{\gamma}_{n,k}^{(i)} - \hat{\gamma}_{n,k}^{(j)} \right)^2$ vs. k

Table 1: *Estimation results, including confidence intervals for a choice of* $\alpha = 5\%$.

Semi-parametric estimators	$\hat{\gamma}$ for $k = 43$	$\hat{x}^F$ for $k = 43$	$\hat{P}(X > 96)$ for $k = 43$
Moment	-0.581 (-1.018,-0.144)	98.261 (96,103.712)	0.018 -
Generalized Hill	-0.632 (-0.906,-0.358)	96.884 (96,99.776)	0.006 -
Mixed Moment	-0.449 (-0.829,-0.069)	103.304 (96,111.238)	0.049 -
PORT-Moment	-0.588 (-1.025,-0.151)	98.059 (96,103.382)	0.016 -

References

[1] Fisher, R. A., Tippett, L. H. C., 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc. Camb. Phil. Soc.*, **24**, 180–190.

[2] Gnedenko, B., 1943. Sur La Distribution Limite Du Terme Maximum D'Une Série Aléatoire. *Ann. Math.*, **44** (3), 423–453.

[3] Balkema, A. A., de Haan, L., 1974. Residual life time at great age. *Ann. Probab.*, **2** (5), 792–804.

[4] Pickands III, J., 1975. Statistical inference using extreme order statistics. *Ann. Stat.*, **3** (1), 119–131.

[5] Vicente, S., 2012. *Extreme Value Theory: an Application to Sports*. Master's Thesis. Department of Statistics and Operations Research, Faculty of Sciences, University of Lisbon. `http://docentes.deio.fc.ul.pt/fragaalves/MSc_thesis_Serge.pdf`

Acknowledgements

The authors would like to thank the support of all the contributors for this research partially supported by National Funds through FCT - Fundação para a Ciência e a Tecnologia, projects PEst-OE/MAT/UI0006/2011 and PTDC/FEDER / EXTREMA.