Estatística de Extremos em Desporto

Lígia Henriques Rodrigues
Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)
lcphjr@gmail.com

M. Ivette Gomes
Universidade de Lisboa, CEAUL, e DEIO-FCUL
Instituto de Investigação Científica Bento da Rocha Cabral, Portugal
migomes@fc.ul.pt

Dinis Pestana
Universidade de Lisboa, CEAUL, e DEIO-FCUL
Instituto de Investigação Científica Bento da Rocha Cabral, Portugal
dinis.pestana@fc.ul.pt

1 Introdução

É muito usual ouvirmos os comentadores desportivos apoiarem as suas opiniões e argumentos em factos estatísticos. Excelentes programas sobre as mais diversas modalidades, tais como corridas com ou sem barreiras, natação, salto em altura e salto à vara, entre outras, mostram que a Estatística é um instrumento indispensável para quem ambiciona alcançar uma medalha, sendo um instrumento que os instrutores de campeões têm obrigatoriamente de usar. As *estatísticas ordinais* (EO's) extremais, e em particular os máximos (ou mínimos) e os recordes, são pois as estatísticas de maior relevância.

Muitas questões da vida real, incluindo o Desporto, requerem a estimação sobre acontecimentos acerca dos quais os dados são inexistentes ou se existem são escassos, os chamados acontecimentos extremos ou raros. A *Teoria de Valores Extremos* (EVT, do Inglês "*extreme value theory*") é um ramo probabilístico de suporte à Estatística que lida exatamente com tais situações, ajudando a descrever e a quantificar os ditos acontecimentos raros. Em particular, permite a estimação de probabilidades de acontecimentos para os quais não existem dados, ou como usualmente dizemos, permite *extrapolar para além da amostra*.

O *Teorema de Fisher-Tippett-Gnedenko* (o teorema fulcral dos tipos em EVT) é um resultado acerca da distribuição assintótica das EO's extremais. O teorema dos tipos extremais desempenha um papel análogo ao tão famoso *Teorema Limite Central* (TLC) para as médias (somas). Basicamente, estabelece que o máximo amostral linear e convenientemente normalizado converge nas mais variadas situações para uma *variável aleatória* (VA) com uma de 3 distribuições possíveis, a Gumbel, a Fréchet ou a Max-Weibull, que podem ser unificadas (Von Mises, 1936; Jenkinson, 1955) na forma funcional,

$$G_\xi(x) = \begin{cases} \exp\left(-(1+\xi x)^{-1/\xi}\right), 1+\xi x > 0, \text{ se } \xi \neq 0 \\ \exp(-\exp(-x)), \ x \in \Re, \text{ se } \xi = 0, \end{cases} \quad (1)$$

a chamada *função de distribuição* (FD) de *valores extremos* (EV, do Inglês "*extreme value*"). O parâmetro de forma ξ em (1) é o chamado *índice de valores extremos* (EVI, do Inglês "*extreme value índex*"). Independentemente da forma do centro da distribuição, a *cauda assume formas sempre muito características* quando estamos suficientemente longe nessa cauda. O crédito deste resultado é devido essencialmente a Gnedenko (1943), embora versões anteriores tivessem sido estabelecidas por Fréchet (1927) e Fisher and Tippett (1928). Em abordagens paramétricas à Estatística de Extremos trabalhamos usualmente com o modelo G_ξ, em (1), ou com o modelo generalizado de Pareto (GP) associado aos excessos acima de um nível elevado, $\mathrm{GP}_\xi(x) = 1 + \ln G_\xi(x), \ x \geq 0$ (veja-se Gomes *et al.*, 2013).

Apesar da existência de abordagens paramétricas em Estatística de Extremos para aplicações a dados de Desporto, como se pode ver em Robinson and Tawn (1995) e Barão and Tawn (1999), que consideram os melhores tempos anuais em corridas de 3000 metros para mulheres, e em Smith (1988), que trabalha com dados da maratona, entre outros, enquadrar-nos-emos aqui essencialmente numa abordagem de índole semi-paramétrica à Estatística de Extremos. Trabalhamos então com estimadores de parâmetros de acontecimentos raros baseados nas k EO's de topo. O parâmetro fundamental continua a ser o EVI, ξ, em (1), que deve ser estimado de forma "precisa", uma vez que é a base para a estimação de outros parâmetros de acontecimentos extremos, tais como o *limite superior do suporte* do modelo subjacente aos dados,

$$x^F = \sup\{x : F(x) < 1\}, \tag{2}$$

de importância fundamental na área do Desporto. Tendo em vista a aplicação a dados de melhores marcas em modalidades de atletismo, e em contexto semelhante ao usado em Einmahl and Magnus (2008), Gomes and Pestana (2009), Einmahl and Smeets (2011) e Henriques-Rodrigues *et al.* (2011), daremos atenção à estimação do índice de cauda, ξ, bem como à estimação do limite superior do suporte x^F, em (2), se finito, o "recorde mundial" possível face às condições atuais.

2 Alguns Resultados em EVT

Consideremos um qualquer acontecimento desportivo, denotemos as melhores marcas pessoais de n atletas por $X_1,\ldots,X_n$ e por $X_{1:n} \le \ldots \le X_{n:n}$ as EO's ascendentes associadas. Admitamos ainda que, caso necessário, os dados são transformados de modo a podermos falar de máximos (e não de mínimos). Iremos pois trabalhar com EO's superiores. Por simplicidade, admitamos também que $(X_1,\ldots,X_n)$ podem ser consideradas como observações independentes e identicamente distribuídas (IID) de um modelo F, desconhecido.

Um dos resultados fundamentais em EVT tem a ver com a identificação das possíveis leis limites de $X_{n:n} := \max(X_1,\ldots X_n)$. Tem-se obviamente $X_{n:n} \xrightarrow[n\to\infty]{p} x^F$, com x^F dado em (2). Para se obter um possível comportamento limite não-degenerado, é necessário normalizar $X_{n:n}$. De forma análoga ao TLC, sabemos que se o máximo $X_{n:n}$, linearmente normalizado, convergir para uma VA não degenerada, existem sucessões de reais $\{a_n\}_{n\ge1}$ (a_n>0) e $\{b_n\}_{n\ge1}$ tais que

$$\lim_{n\to\infty} P\left(a_n^{-1}\left(X_{n:n} - b_n\right) \le x\right) = \lim_{n\to\infty} F^n\left(a_n x + b_n\right) = G_\xi(x) \tag{3}$$

para algum $\xi \in \Re$ (Gnedenko, 1943), com $G_\xi(x)$ definida em (1). Dizemos então que F está no *max-domínio de atração* de G_ξ e usamos a notação óbvia $F \in D_M(G_\xi)$. O índice de valores extremos ξ mede pois essencialmente o peso da cauda direita $\overline{F} = 1 - F$:

- se $\xi < 0$, temos uma cauda direita leve, i.e. F tem limite superior de suporte finito;
- se $\xi > 0$, temos uma cauda direita pesada de tipo polinomial negativo, também chamada de tipo Pareto, i.e. F tem limite superior de suporte infinito;
- se $\xi = 0$, a cauda direita é de tipo exponencial e o limite superior do suporte de F pode ser finito ou infinito.

Observação 2.1. *Note-se que qualquer resultado obtido para máximos pode ser reformulado para mínimos, uma vez que* $\min(X_1,\ldots,X_n) = -\max(-X_1,\ldots,-X_n)$.

Observação 2.2. *Note-se ainda que dizer que* $F \in D_M(G_\xi)$ *é equivalente a dizer que para qualquer* $x \in \Re$ *tal que* $0 < G_\xi(x) < 1$, *tem-se* $\lim_{n\to\infty} n \ln F(a_n x + b_n) = \ln G_\xi(x) = -(1+\xi x)^{-1/\xi}$, *tendo-se, equivalentemente,*

$$\lim_{n\to\infty} n\left(1 - F(a_n x + b_n)\right) = -\ln G_\xi(x) = (1+\xi x)^{-1/\xi}. \tag{4}$$

Definamos

$$U(t) := F^{\leftarrow}(1-1/t) \ \ (t>1), \quad F^{\leftarrow}(x) := \inf\{y : F(y) \geq x\}. \tag{5}$$

É fácil de demonstrar (veja-se o Teorema 1.1.2 em de Haan and Ferreira, 2006) que, com $a_t \equiv a(t) := a_{\lfloor t \rfloor}$ e $b_t \equiv b(t) := b_{\lfloor t \rfloor}$, com $\lfloor t \rfloor$ = parte inteira de t, (a_n, b_n) definidos em (3), e denotando G_ξ^{-1} a função inversa da FD G_ξ em (1),

$$\lim_{n\to\infty} \left(U(tx) - b_t\right)/a_t = D(x) = G_\xi^{-1}(\exp(-1/x)) = (x^\xi - 1)/\xi \tag{6}$$

para todo o $x>0$, podendo-se escolher $b_t=U(t)$, com $U(.)$ definido em (5).

Observação 2.3. *Quando* $\xi=0$*, as funções* $-\ln G_\xi(x) = (1+\xi x)^{-1/\xi}$ *e* $G_\xi^{-1}(\exp(-1/x)) = (x^\xi - 1)/\xi$*, em* (4) *e em* (6)*, devem ser interpretadas como* $\exp(-x)$ *e* $\ln x$*, respetivamente.*

3 Estimação semi-paramétrica de alguns parâmetros de interesse

Como estimar o índice de valores extremos ξ, a escala a, a localização b e o limite superior do suporte x^F?

3.1 Estimadores do índice de valores extremos

Para $j \geq 1$, denotemos

$$L_{k,n}^{(j)} := \frac{1}{k}\sum_{i=1}^{k}\left\{1 - \frac{X_{n-k:n}}{X_{n-i+1:n}}\right\}^j, \quad M_{k,n}^{(j)} := \frac{1}{k}\sum_{i=1}^{n}\left\{\ln\frac{X_{n-i+1:n}}{X_{n-k:n}}\right\}^j. \tag{7}$$

Neste estudo, consideraremos os estimadores seguintes, válidos para qualquer $\xi \in \Re$.

1. O estimador de *momentos* (M), introduzido em Dekkers *et al*. (1989), com a forma funcional

$$\hat{\xi}_{k,n}^M \equiv M_{k,n} := M_{k,n}^{(1)} + \frac{1}{2}\left\{1 - \left(M_{k,n}^{(2)} \Big/ \left[M_{k,n}^{(1)}\right]^2 - 1\right)^{-1}\right\}. \tag{8}$$

2. O estimador *generalizado de Hill* (GH), introduzido em Beirlant *et al*. (1996) e estudado posteriormente em Beirlant *et al*. (2005). Tem-se

$$\hat{\xi}_{k,n}^{GH} \equiv GH_{k,n} := \hat{\xi}_{k,n}^H + \frac{1}{k}\sum_{i=1}^{k}\left\{\ln\frac{\hat{\xi}_{i,n}^H}{\hat{\xi}_{k,n}^H}\right\}, \quad \xi_{k,n}^H := \frac{1}{k}\sum_{i=1}^{k}\left\{\ln\frac{X_{n-i+1:n}}{X_{n-k:n}}\right\}. \tag{9}$$

3. O estimador de *momentos mistos* (MM) (Fraga Alves *et al*., 2009), com a forma funcional

$$\hat{\xi}_{k,n}^{MM} \equiv MM_{k,n} := \frac{\hat{\varphi}_{k,n} - 1}{1 + 2\min(\hat{\varphi}_{k,n} - 1, 0)}, \quad \hat{\varphi}_{k,n} := \frac{M_{k,n}^{(1)} - L_{k,n}^{(1)}}{\left(L_{k,n}^{(1)}\right)^2}. \tag{10}$$

Devido à especificidade dos dados, consideraremos também um estimador simples, válido só para $\xi < -1/2$,

4. o estimador introduzido em Falk (1995), denotado F, e dado por

$$\hat{\xi}_{k,n}^F \equiv F_{k,n} := \frac{1}{k}\sum_{i=1}^{k-1}\ln\frac{X_{n:n} - X_{n-1:n}}{X_{n:n} - X_{n-k:n}}. \tag{11}$$

Trata-se de estimadores que são consistentes desde que $k=k_n$ seja uma sucessão intermédia, i.e. uma sucessão de inteiros tais que

$$k = k_n \to \infty \text{ e } k_n = o(n), \quad \text{quando } n \to \infty. \tag{12}$$

Numa sub-classe vasta do domínio de atração, e sob condições adequadas em k, conseguimos ainda garantir a normalidade assintótica de qualquer destes estimadores, bem como dos estimadores de parâmetros de acontecimentos extremos a introduzir na Secção 3.2.

3.2 Estimação de outros parâmetros de interesse

3.2.1 Estimadores dos parâmetros de localização e escala

Já mencionámos que podemos escolher $b_t=U(t)$, com $U(.)$ definida em (5). Por outro lado, a transformação uniformizante permite-nos garantir que $\forall F$ desconhecido e subjacente à VA X, $X \overset{d}{=} U(Y)$ com Y VA Pareto unitária, i.e., VA com FD $F_Y(y) = 1 - 1/y,\ y \geq 1$. Consequentemente,

$$X_{n-k:n} \overset{d}{=} U\left(Y_{n-k:n}\right), \quad \text{e como} \quad Y_{n-k:n} \overset{p}{\sim} n/k, \quad \text{quando } n \to \infty,$$

é sensato considerar

$$\hat{b} = \hat{b}_{k,n} = \hat{U}(n/k) = X_{n-k:n}.$$

E qualquer que seja o estimador de ξ, denotado $\hat{\xi}^{\bullet}_{k,n}$, pode-se considerar

$$\hat{a}^{\bullet} = \hat{a}^{\bullet}_{k,n} = X_{n-k:n} M^{(1)}_{k,n} \left(1 - \min\left(0, \hat{\xi}^{\bullet}_{k,n}\right)\right),$$

com $M^{(1)}_{k,n}$, definido em (7).

3.2.2 Estimadores do limite superior do suporte, caso $\xi<0$

É possível provar (de Haan, 1984) que $F \in D_M\left(G_\xi\right)$ se e só se, para todo o $x>0$,

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = \begin{cases} \dfrac{x^\xi - 1}{\xi} & \text{se} \quad \xi \neq 0 \\ \ln x & \text{se} \quad \xi = 0, \end{cases}$$

com $U(.)$ definida em (5). Para valores elevados de t, podemos pois escrever

$$U(tx) \approx U(t) + a(t)(x^\xi - 1)/\xi.$$

Mas $x^F=U(\infty)$ e para $\xi<0$, $(x^\xi - 1)/\xi \to -1/\xi$, quando $x \to \infty$. Fazendo $t=n/k$, podemos pois garantir que, sempre que $\hat{\xi}^{\bullet} < 0$,

$$x^F \approx U(n/k) - a(n/k)/\xi \ \Rightarrow\ \hat{x}^F_{\bullet} := \hat{b} - \hat{a}^{\bullet}/\hat{\xi}^{\bullet}.$$

Para mais detalhes sobre o assunto, vejam-se os Capítulos 1 e 4 em de Haan and Ferreira (2006).

4 Análise de um conjunto de dados em atletismo

Como ilustração, procedemos a uma análise dos dados associados a corridas de 200 metros, para homens. Os dados foram recolhidos em http://www.iaaf.org/statistics/toplists/index.htmx e http://hem.bredband.net/athletics/athletics_all-time_best.htm. As observações foram colecionadas até ao fim de 2007, e para cada atleta foi unicamente considerada a sua melhor marca. A dimensão da amostra é n=352. Tal como anteriormente referido, estamos interessados na cauda direita do modelo F subjacente aos dados. Consequentemente, convertemos *tempos de corrida* em *velocidades*, i.e., 20 segundos nos 200 metros (ou seja, nos 0.2 quilómetros) são transformados na velocidade de 3600 × 0.2/20=36 km/h. Com esta transformação é óbvio que quanto maior for a velocidade tanto melhor, e estamos pois interessados em valores máximos, usando unicamente o melhor resultado de cada um dos n atletas. Análises semelhantes de outros conjuntos de dados em Desporto são efectuadas em Gomes e Pestana (2009) e em Henriques-Rodrigues *et al*. (2011).

4.1 Análise paramétrica dos dados

Como os dados observados (após a mudança de escala atrás referida) já são valores máximos, possivelmente de um número pequeno de marcas dependentes associadas com cada um dos *n* atletas, mas a lei limite em (1) é robusta relativamente ao relaxamento quer da hipótese de independência quer de identidade distribucional, tentámos primeiro ajustar, através da máxima verosimilhança (ML, do Inglês "*maximum likelihood*"), um modelo de valores extremos $F(x;\lambda,\delta,\xi)=G_\xi((x-\lambda)/\delta)$, com $G_\xi(x)$ dada em (1). Usámos para isso o "package" EVIR, do "software" **R** e os resultados são apresentados na Tabela 1. Tal como esperado, a estimativa de ξ é negativa. Um pouco mais dececionante é a estimação do limite superior do suporte, dada por $\hat{x}^F = \max\left(x_{n:n}, \hat{\lambda} - \hat{\delta}/\hat{\xi}\right)$, que é igual ao valor máximo dos dados.

Modalidade	*n*	$(x_{1:n}, x_{n:n})$ em Km/h	$\hat{\xi}$ (IC a 95%)	$\hat{\lambda}$	$\hat{\delta}$	$\hat{x}^F$
200M	352	(33.72, 36.14)	-0.22 (-0.340,-0.101)	34.08	0.28	36.14

Tabela 1: Estimativas ML de ξ, (λ,δ) e x^F para o modelo $G_\xi((x-\lambda)/\delta)$, com $G_\xi(x)$ dada em (1)

A um nível de significância α=0.05, o modelo (unificado) de *valores extremos* foi rejeitado pelo teste de Kolmogorov-Smirnov, tal como pode ser inferido graficamente da Figura 1, em que representamos à *esquerda* a FD empírica e a FD de valores extremos estimada, para o conjunto de dados em análise.

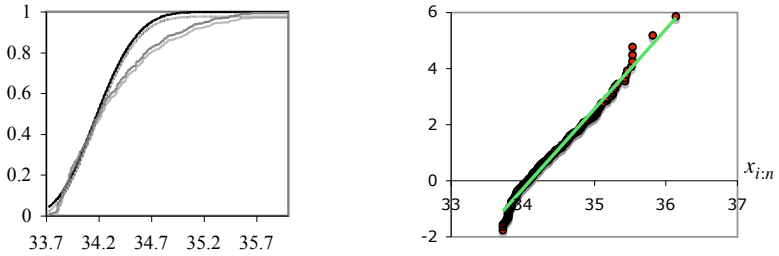

Figura 1: FD empírica (a cinzento) e FD de valores extremos (a preto) (*esquerda*), e Gumbel QQ-plot (*direita*) do conjunto de dados relativos a corridas de 200 metros, para homens

Antes de procedermos a uma análise em contexto semi-paramétrico, ilustramos, também na Figura 1, à *direita*, o gráfico em papel de probabilidade Gumbel associado aos dados em análise, com a marcação dos pontos $\left(x_{i:n},\ p_i^{\Lambda} = -\ln\left(-\ln\left(i/(n+1)\right)\right)\right)$, $1 \le i \le n$, e o ajustamento de uma recta de mínimos quadrados. E este QQ-plot evidencia também um comportamento de cauda direita leve, i.e. um índice de cauda ξ <0.

Face à rejeição do modelo EV, e de outros modelos paramétricos alternativos, sentimos pois a necessidade de uma análise de dados em contexto semi-paramétrico, a desenvolver em seguida.

4.2 Análise semi-paramétrica dos dados

4.2.1 Teste ao sinal do índice de valores extremos

Tal como mencionámos anteriormente, e sempre que nos colocamos em contexto semi-paramétrico, admitimos unicamente que $F \in D_M\left(G_\xi\right)$, sendo ξ o parâmetro fundamental de valores extremos. E em muitas áreas em que os extremos são relevantes, o caso mais frequente e simples é considerar ξ =0. Além disso, se claramente pensamos que ξ <0 ou que ξ >0, temos procedimentos específicos para a

estimação ξ, frequentemente mais fiáveis do que os procedimentos válidos para $\xi \in \Re$. Antes de procedermos a uma análise mais aprofundada da cauda direita de F é pois sensato testar

$$H_0 : F \in D_M\left(G_\xi\right)_{\xi=0} \ \left(\text{ou } F \in D_M\left(G_\xi\right)_{\xi\geq 0}\right) \quad \text{versus} \quad H_1 : F \in D_M\left(G_\xi\right)_{\xi<0}. \tag{13}$$

Iremos aqui considerar duas estatísticas de teste baseadas nos excessos acima de um limiar elevado, $X_{n-k:n}$, com k intermédio, i.e. tal que se tem a validade de (12). A primeira estatística de teste foi introduzida em Greenwood (1946) e a segunda em Hasofer and Wang (1992). Estas estatísticas foram estudadas em contexto semi-paramétrico em Neves and Fraga Alves (2007) e são dadas por

$$G_{k,n} := \frac{\frac{1}{k}\sum_{i=1}^{k}\left(X_{n-i+1:n} - X_{n-k:n}\right)^2}{\left(\frac{1}{k}\sum_{i=1}^{k}\left(X_{n-i+1:n} - X_{n-k:n}\right)\right)^2}, \qquad W_{k,n} := \frac{1}{k\left(G_{k,n}-1\right)}.$$

Sob a validade da hipótese nula H_0, em (13), e condições extras de regularidade impostas na cauda direita de F e no crescimento de $k=k_n$, ambas as estatísticas têm um comportamento assintótico normal. Mais especificamente,

$$G^*_{k,n} := \sqrt{k/4}\left(G_{k,n}-2\right)\Big|_{F\in D_M(G_0)} \xrightarrow[n\to\infty]{d} N(0,1) \tag{14}$$

e

$$W^*_{k,n} := \sqrt{k/4}\left(kW_{k,n}-1\right)\Big|_{F\in D_M(G_0)} \xrightarrow[n\to\infty]{d} N(0,1). \tag{15}$$

Face à importante contribuição do máximo para a soma dos k excessos, $X_{n-i+1:n} - X_{n-k:n}$, $1 \leq i \leq k$, Neves *et al.* (2006) introduziram a estatística,

$$R_{k,n} := \frac{X_{n:n} - X_{n-k:n}}{\frac{1}{k}\sum_{i=1}^{k}\left(X_{n-i+1:n} - X_{n-k:n}\right)},$$

também incluída nesta análise de dados. O comportamento assintótico de $R_{k,n}$ é Gumbel, i.e. $\Lambda=G_0$ é a FD da VA limite, com G_ξ definida em (1). Mais especificamente,

$$R^*_{k,n} := R_{k,n} - \ln k\Big|_{F\in D_M(G_0)} \xrightarrow[n\to\infty]{d} Z \sim G_0. \tag{16}$$

Como função de k, $G^*_{k,n}$ e $R^*_{k,n}$ tendem a ter uma inclinação com o sinal de ξ. A estatística $W^*_{k,n}$ funciona em sentido reverso.

Na Figura 2, e para os dados dos 200 metros, apresentamos as trajetórias amostrais das três estatísticas de teste, $G^*_{k,n}$, $W^*_{k,n}$ e $R^*_{k,n}$ em (14), (15) e (16), respectivamente. Nessa mesma figura também marcamos os quantis $\left(\chi^{\cdot}_{0.025}, \chi^{\cdot}_{0.975}\right)$ da normal padrão Φ, os valores (-1.96,+1.96), e da Gumbel padrão, $\Lambda \equiv G_0$, os valores (-1.31, +3.68).

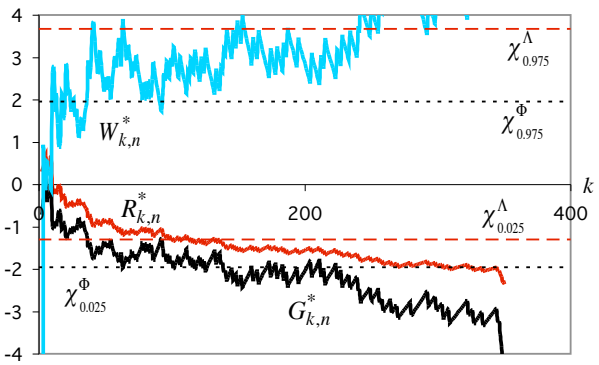

Figura 2: Estatísticas de teste para as corridas de 200 metros, para homens

Este gráfico mostra claramente uma tendência decrescente de $R^*_{k,n}$ e de $G^*_{k,n}$ (respectivamente abaixo de $\chi^{\Lambda}_{0.025}$ e de $\chi^{\Phi}_{0.025}$ para um grande número de valores de k), bem como uma tendência crescente de $W^*_{k,n}$ (acima de $\chi^{\Phi}_{0.975}$ desde valores moderados até valores elevados de k). Essa tendência está essencialmente relacionada com o viés, mas o viés está por sua vez fortemente relacionado com o sinal do EVI. Apesar disto note-se que para o ficheiro 200M (dos 200 metros, para homens), a trajetória amostral de $G^*_{k,n}$ e mais acentuadamente a de $R^*_{k,n}$ estão dentro do intervalo de confiança a 95% para uma zona vasta de valores de k. Tal situação era também esperada, uma vez que se sabe (veja-se, por exemplo Neves and Fraga Alves, 2008) que $R^*_{k,n}$ é um teste conservativo e o verdadeiro valor de ξ não está certamente muito longe de zero. Temos de qualquer modo uma indicação clara de um índice de valores extremos negativo, tal como esperado.

4.2.2 Estimativas semi-paramétricas do EVI e do limite superior do suporte

Dificuldades na estimação do valor ótimo de k, no sentido de erro quadrático médio mínimo, para ξ <0, levaram-nos a considerar um método heurístico, apresentado em seguida, estudado em Henriques-Rodrigues *et al.* (2011). Seja $\hat{\xi}^{(i)}_{k,n}$, $i \in T = \{1,2,3,4\}$, o conjunto de estimadores alternativos do EVI aqui considerados, i.e. os estimadores em (8), (9), (10) e (11). Sugerimos então a escolha

$$k^*_{\min} := \arg\min_k \sum_{(i,j)\in T, i\neq j} \left(\hat{\xi}^{(i)}_{k,n} - \hat{\xi}^{(j)}_{k,n}\right)^2 \qquad (17)$$

e a consideração dos estimadores adaptativos

$$T^*_{\min} = T_{k^*_{\min},n}, \text{ para } T\text{=M, GH, MM e F,} \qquad (18)$$

com $M_{k,n}$, $GH_{k,n}$, $MM_{k,n}$, $F_{k,n}$ e $k^*_{\min}$ dados em (8), (9), (10), (11) e (17), respectivamente.

Na Tabela 2 apresentamos uma estimativa de ξ e associado intervalo de confiança a 95%. Esta estimativa de ξ foi obtida através do estimador MM em (10), calculado no valor $k^*_{\min}$ em (17) também apresentado na última coluna da tabela. A escolha de $MM^*_{\min}$ prende-se com um estudo de simulação desenvolvido para o modelo de *valores extremos* G_ξ em (1), que nos forneceu, para o viés absoluto e o erro quadrático médio (EQM) dos estimadores $T^*_{\min}$ em (18), padrões já referidos em Gomes e Pestana (2009), do tipo do fornecido na Figura 3.

Modalidade	n	$(x_{1:n}, x_{n:n})$	MM^* (IC a 95%)	$k^*_{\min}$
200M	352	$(33.72, 36.14)^*$	-0.25 (-0.379,-0.129)	269

Tabela 2: Estimativas do índice de valores extremos: *km/h

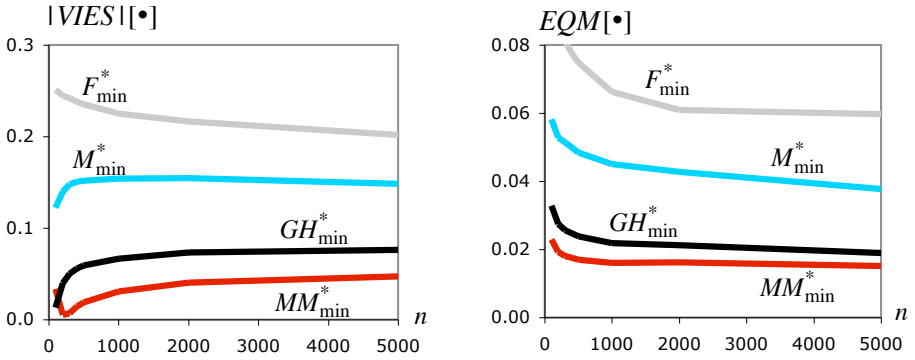

Figura 3: Valor absoluto do viés (*esquerda*) e erros médios quadráticos (*direita*) dos estimadores $T^*_{\min}$ em (18), para um modelo de valores extremos com ξ=-0.1 e para dimensões de amostra n=100, 200, 500, 1000,

2000 e 5000.

Finalmente, na Tabela 3, apresentamos a estimativa do limite superior do suporte obtida com base na estimativa de ξ associada ao estimador MM em (10), para as corridas de 200 metros, para homens.

Modalidade	$x_{n:n}$	k^{*F}	$\hat{x}^F_{MM}$
200M	36.14^{*} | (00:19.92)$^{\bullet}$	269	36.47^{*} | (00:19.74)$^{\bullet}$

Tabela 3: Estimativa do limite superior do suporte, *km/h, e conversão para o limite inferior, $^{\bullet}$minutos

O resultado apresentado na Tabela 3 não é surpreendente e mostra-nos que, nas condições atuais, existe um limite superior para a velocidade, que implica que haja um limite inferior para o tempo de corrida da modalidade analisada, tal como já tinha sido detectado em artigos anteriores sobre o assunto.

Agradecimentos

Investigação parcialmente financiada por Fundos Nacionais através da **FCT** --- Fundação para a Ciência e a Tecnologia, bolsa SFRH/BPD/77319/2011 e projeto PEst-OE/MAT/UI0006/2014.

Referências

[1] Barão, M.I., and Tawn, J. (1999). Extremal analysis of short series with outliers: sea-levels and athletic records. *Applied Statistics* **48**, 469–487.

[2] Beirlant, J., Dierckx, G., and Guillou, A. (2005). Estimation of the extreme-value index and generalized quantile plots. *Bernoulli* **11**:6, 949–970.

[3] Beirlant, J., Vynckier, P., and Teugels, J. (1996). Excess functions and estimation of the extreme-value index. *Bernoulli* **2**, 293–318.

[4] Dekkers, A.L.M., Einmahl, J.H.J. and Haan, L. de (1989). A moment estimator for the index of an extreme-value distribution. *Ann. Statist.* **17**, 1833–1855.

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

[6] Einmahl, J., and Smeets, S.G.W.R. (2011). Ultimate 100-m world records through extreme-value theory. *Statistica Neerlandica* **65**:1, 32–42.

[7] Falk, M. (1995). Some best parameter estimates for distributions with finite endpoint. *Statistics* **27**(1-2), 115–125.

[8] Fisher, R.A., and Tippett, L.H.C. (1928). Limiting forms of the frequency of the largest or smallest member of a sample. *Proc. Cambridge Phil. Soc.* **24**, 180–190.

[9] Fraga Alves, M.I., Gomes M.I., de Haan, L., and Neves, C. (2009). Mixed moment estimator and location invariant alternatives. *Extremes* **12**, 149–185.

[10] Fréchet, M. (1927). Sur le loi de probabilité de l'écart maximum. *Ann. Société Polonaise de Mathématique* **6**, 93–116.

[11] Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math.* **44**, 423–453.

[12] Gomes, M.I., and Pestana, D. (2009). Caudas leves em desporto: estimação de parâmetros úteis. In Oliveira, I., M.M. *et al.* (eds.), *Estatística: Arte de Explicar o Acaso*, 307–318, Edições S.P.E.

[13] Gomes, M.I., Fraga Alves, M.I., e Neves, C. (2013). *Análise de Valores Extremos: uma Introdução*. Edições S.P.E. and I.N.E.

[14] Greenwood, M. (1946). The statistical study of infectious diseases. *J. Roy. Statist. Soc.* **A109**, 85–109.

[15] Haan, L. de (1984). Slow variation and characterization of domains of attraction. In Tiago de

Oliveira, ed., *Statistical Extremes and Applications*, D. Reidel, Dordrecht, 31–48.
[16] Haan, L. de, and Ferreira, A. (2006). *Extreme Value Theory: an Introduction*. Springer Science+Business Media, LLC, New York.
[17] Hasofer, A., and Wang, J.Z. (1992). A test for extreme value domain of attraction. *J. Amer. Statist. Assoc.* **87**, 171–177.
[18] Henriques-Rodrigues, L., Gomes, M.I., and Pestana, D. (2011). Statistics of extremes in athletics. *Revstat* **9**:2, 127–153.
[19] Jenkinson, A.F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quart. J. Royal Meteorol. Society* **81**, 158–171.
[20] Mises, R. von (1936). La distribution de la plus grande de n valeurs, *Revue Math. Union Interbalcanique* **1**, 141-160. Reprinted in *Selected Papers of Richard von Mises*, Amer. Math. Soc. **2** (1964), 271–294.
[21] Neves, C., and Fraga Alves, M.I. (2007). Semi-parametric approach to Hasofer-Wang and Greenwood statistics in extremes. *Test* **16**, 297–313.
[22] Neves, C., and Fraga Alves, M.I. (2008). Testing extreme value conditions — an overview and recent approaches. *Revstat* **6**:1, 83–100.
[23] Neves, C., Picek, J., and Fraga Alves, M.I. (2006). The contribution of the maximum to the sum of excesses for testing max-domains of attraction. *J. Statist. Planning and Inference* **136** (4), 1281–1301.
[24] Robinson, M.E., and Tawn, J. (1995). Statistics for exceptional athletic records. *Applied Statistics* **44**, 499–511.
[25] Smith, R.L. (1988). Forecasting records by maximum likelihood. *J. American Statistical Association* **83**, 331–338.