Ranking Elite Swimmers using Extreme Value Theory

Harry Spearing
Supervisors: Jon Tawn*, Tim Paulden[†], David Irons[†], Grace Bennett[†]
*Lancaster University, [†]ATASS sports

June 29, 2019

Motivation

Motivation

- FINA rankings are currently used to compare swimmers across all 34 Olympic events.

Motivation

- FINA rankings are currently used to compare swimmers across all 34 Olympic events.
- Based on points system,

$$p_{i,j} \propto (b_j / t_{i,j})^3$$

$p_{i,j}$ given to swimmer i in event j is where b_j is the *base time* for event j.

Motivation

- FINA rankings are currently used to compare swimmers across all 34 Olympic events.
- Based on points system,

$$p_{i,j} \propto (b_j/t_{i,j})^3$$

$p_{i,j}$ given to swimmer i in event j is where b_j is the *base time* for event j.

- Very sensitive to changes in world record.

Motivation

- FINA rankings are currently used to compare swimmers across all 34 Olympic events.
- Based on points system,

$$p_{i,j} \propto (b_j/t_{i,j})^3$$

$p_{i,j}$ given to swimmer i in event j is where b_j is the *base time* for event j.

- Very sensitive to changes in world record.
- Bias between events.

A good ranking system should be **robust** to changes in the data, and quantify **uncertainty** in the ranks.
Other features of interest:

A good ranking system should be **robust** to changes in the data, and quantify **uncertainty** in the ranks.
Other features of interest:

- What is the ultimate possible swim-time for a given event?

A good ranking system should be **robust** to changes in the data, and quantify **uncertainty** in the ranks.
Other features of interest:

- What is the ultimate possible swim-time for a given event?
- What will be the swim-time of the next world record?

A good ranking system should be **robust** to changes in the data, and quantify **uncertainty** in the ranks.
Other features of interest:

- What is the ultimate possible swim-time for a given event?
- What will be the swim-time of the next world record?
- When will current world records next be broken?

A good ranking system should be **robust** to changes in the data, and quantify **uncertainty** in the ranks.
Other features of interest:

- What is the ultimate possible swim-time for a given event?
- What will be the swim-time of the next world record?
- When will current world records next be broken?
- Which event will next have a new record set?

A good ranking system should be **robust** to changes in the data, and quantify **uncertainty** in the ranks.
Other features of interest:

- What is the ultimate possible swim-time for a given event?
- What will be the swim-time of the next world record?
- When will current world records next be broken?
- Which event will next have a new record set?
- Can we adjust for technological advances?

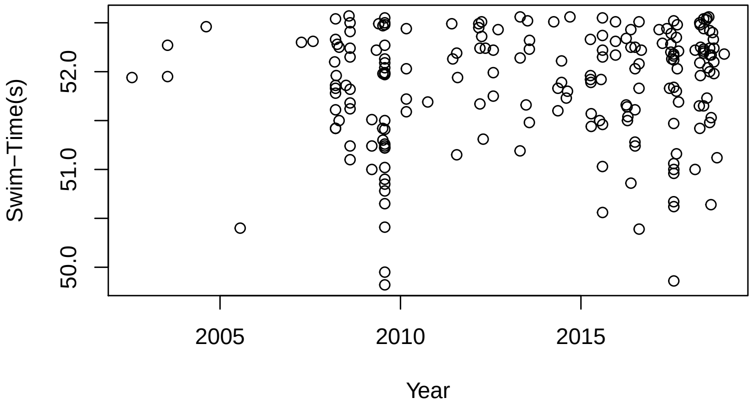


Figure: Data for the men's 100m butterfly. The fastest 200 times over the period 2001 to late 2018.

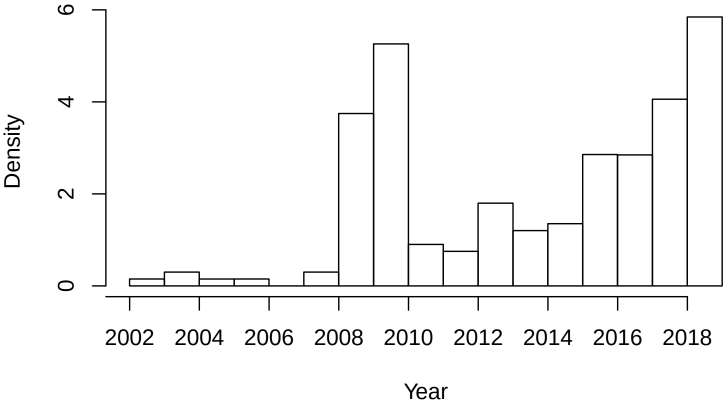


Figure: Rate of observations

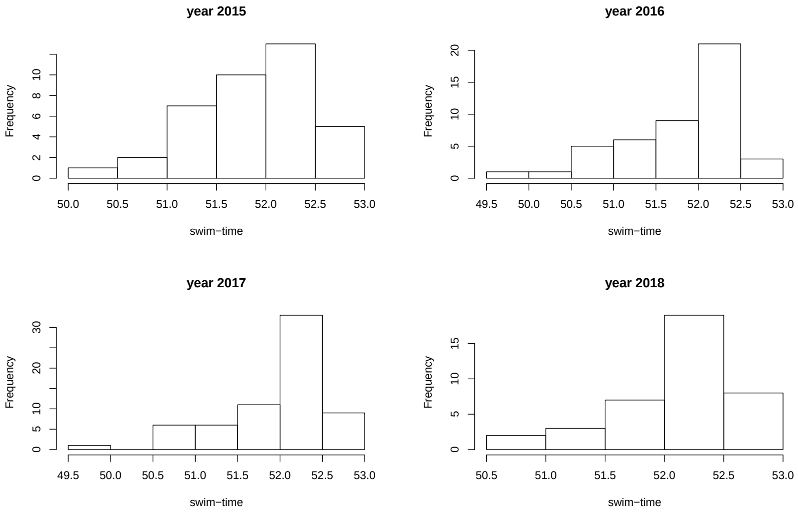


Figure: Distribution of swim-times.

Theory

Theory

- Want to model both **rate** and **distribution** of these extreme events.

Theory

- Want to model both **rate** and **distribution** of these extreme events.
- By definition, there are very few observations.

Theory

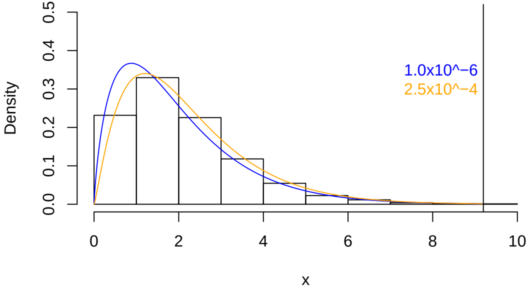
- Want to model both **rate** and **distribution** of these extreme events.
- By definition, there are very few observations.
- Data of the swimmer population not available.

Theory

- Want to model both **rate** and **distribution** of these extreme events.
- By definition, there are very few observations.
- Data of the swimmer population not available.
- Even if it was, this would not be useful!

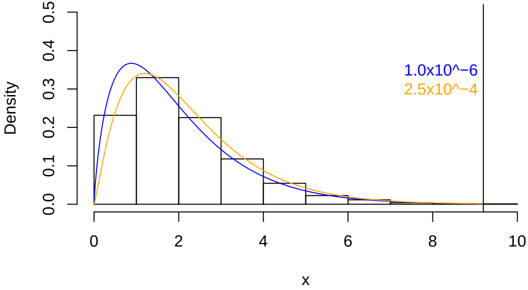
Theory

- Want to model both **rate** and **distribution** of these extreme events.
- By definition, there are very few observations.
- Data of the swimmer population not available.
- Even if it was, this would not be useful!



Theory

- Want to model both **rate** and **distribution** of these extreme events.
- By definition, there are very few observations.
- Data of the swimmer population not available.
- Even if it was, this would not be useful!



- A Separate framework for analysing extreme events is required.

Extreme Value Theory

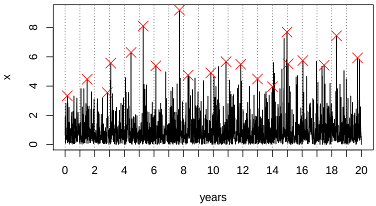


Figure: Block maxima

Extreme Value Theory

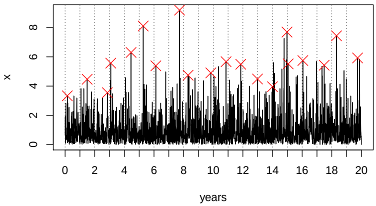


Figure: Block maxima

Generalised extreme value (GEV) distribution function:

$$G(x) = \exp\left(-[1 + \xi(x - \mu)/\sigma]_+^{-1/\xi}\right),$$

$\mu,\ \xi \in \mathbb{R},\ \sigma \in \mathbb{R}^+.$

Extreme Value Theory

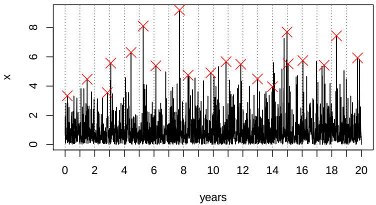


Figure: Block maxima

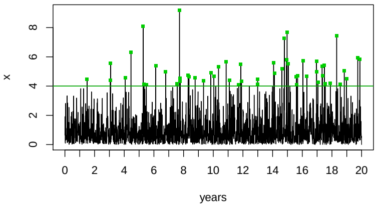


Figure: Peaks over threshold, u

Generalised extreme value (GEV) distribution function:

$$G(x) = \exp\left(-[1+\xi(x-\mu)/\sigma]_+^{-1/\xi}\right),$$

$\mu,\ \xi \in \mathbb{R},\ \sigma \in \mathbb{R}^+.$

Extreme Value Theory

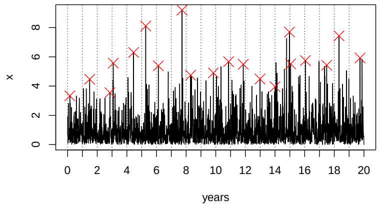


Figure: Block maxima

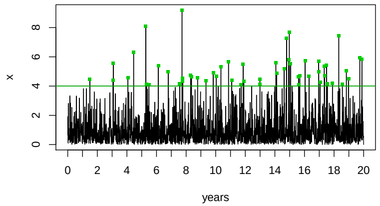


Figure: Peaks over threshold, u

Generalised extreme value (GEV) distribution function:

$$G(x) = \exp\left(-[1+\xi(x-\mu)/\sigma]_+^{-1/\xi}\right),$$

$\mu,\ \xi \in \mathbb{R},\ \sigma \in \mathbb{R}^+.$

Generalised pareto distribution (GPd) function:

$$H_u(x) = 1 - \left[1 + \xi\left(\frac{x-u}{\tilde{\sigma}_u}\right)\right]_+^{-\frac{1}{\xi}},$$

$\xi \in \mathbb{R},\ \tilde{\sigma} \in \mathbb{R}^+.$

Extreme Value Theory

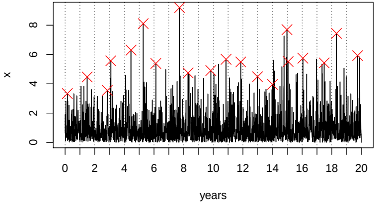


Figure: Block maxima

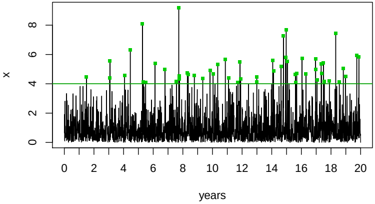
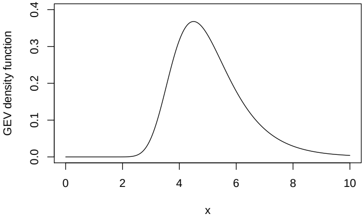
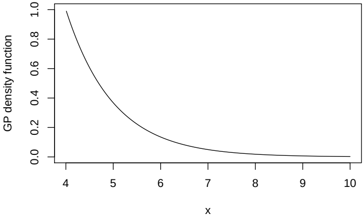


Figure: Peaks over threshold, u





Extreme Value Theory

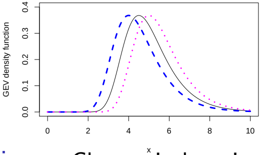


Figure: Change in location, μ

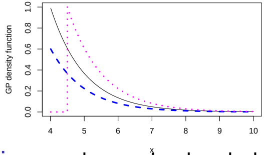


Figure: change in threshold, u

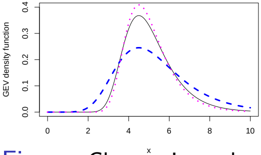


Figure: Change in scale, σ

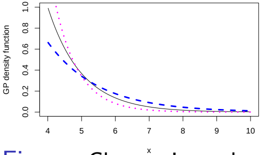


Figure: Change in scale, σ

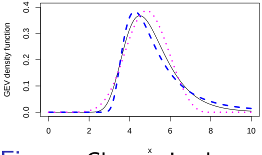


Figure: Change in shape, ξ

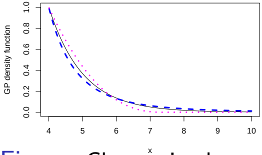


Figure: Change in shape, ξ

Point process

Point process

- Point process framework combines annual maxima (GEV), and peaks over threshold (GPd), to allow for non-homogeneous rates of occurrence.

Point process

- Point process framework combines annual maxima (GEV), and peaks over threshold (GPd), to allow for non-homogeneous rates of occurrence.
- Connects GEV parameters (μ, σ, ξ) to GPd parameters $(\tilde{\sigma}, \xi)$ via:

$$\tilde{\sigma} = \sigma + \xi(u - \mu).$$

Point process

- Point process framework combines annual maxima (GEV), and peaks over threshold (GPd), to allow for non-homogeneous rates of occurrence.
- Connects GEV parameters (μ, σ, ξ) to GPd parameters $(\tilde{\sigma}, \xi)$ via:

$$\tilde{\sigma} = \sigma + \xi(u - \mu).$$

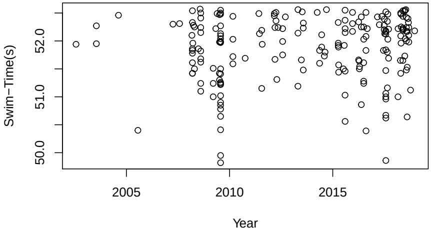
- Model both **rate** and **distribution** of extreme data, for any event.

Point process

- Point process framework combines annual maxima (GEV), and peaks over threshold (GPd), to allow for non-homogeneous rates of occurrence.
- Connects GEV parameters (μ, σ, ξ) to GPd parameters $(\tilde{\sigma}, \xi)$ via:

$$\tilde{\sigma} = \sigma + \xi(u - \mu).$$

- Model both **rate** and **distribution** of extreme data, for any event.
- Observations occur via a Poisson random variable with rate $R(\mu, \sigma, \xi | x, t)$ for a swim-time x at time t.



Model: independent events

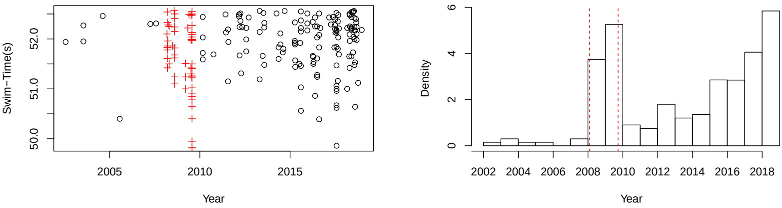


Figure: Data for the men's 100m butterfly.

Model: independent events

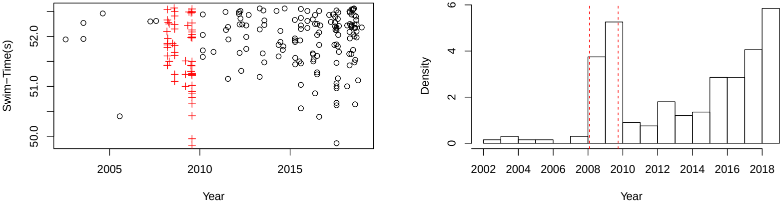
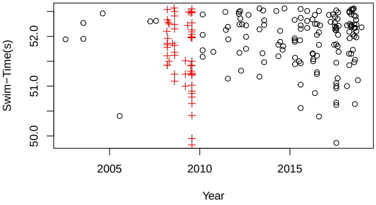


Figure: Data for the men's 100m butterfly.

- A trend in the rates of occurrences.

Model: independent events



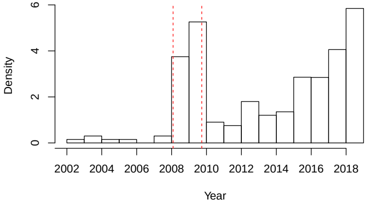


Figure: Data for the men's 100m butterfly.

- A trend in the rates of occurrences.
- Observations above threshold have a time independent distribution.

Model: independent events

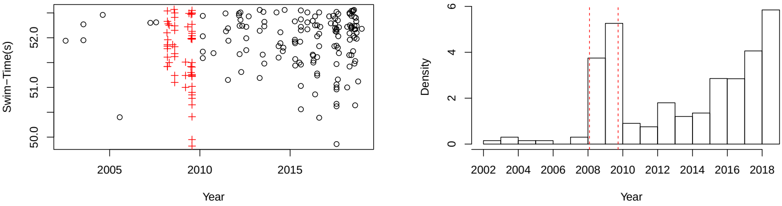


Figure: Data for the men's 100m butterfly.

- A trend in the rates of occurrences.
- Observations above threshold have a time independent distribution.
- Swim-suit effect.

Model: independent events

This is achieved via two additional parameters:

$$(\xi^{(e)}, \mu_0^{(e)}, \sigma_0^{(e)}, \beta^{(e)}, \gamma^{(e)}),$$

Model: independent events

This is achieved via two additional parameters:

$$(\xi^{(e)}, \mu_0^{(e)}, \sigma_0^{(e)}, \beta^{(e)}, \gamma^{(e)}),$$

such that:

$$\begin{aligned}
\xi^{(e)}(t) &= \xi^{(e)}, \\
\mu^{(e)}(t) &= \mu_0^{(e)} + \beta^{(e)} t + \gamma^{(e)} \mathbb{1}_{\{t \in S_t\}}, \\
\sigma^{(e)}(t) &= \sigma_0^{(e)} + \xi^{(e)} \beta^{(e)} t + \xi^{(e)} \gamma^{(e)} \mathbb{1}_{\{t \in S_t\}},
\end{aligned}$$

...independent event model

- Have model with with five parameters: μ_0, σ_0, ξ, β, γ.

...independent event model

- Have model with with five parameters: μ_0, σ_0, ξ, β, γ.
- For all 34 Olympic swimming events, this means 170 parameters in total.

...independent event model

- Have model with with five parameters: μ_0, σ_0, ξ, β, γ.
- For all 34 Olympic swimming events, this means 170 parameters in total.
- Can improve model **robustness** and **predictive power** by pooling information across events.

...independent event model

- Have model with with five parameters: μ_0, σ_0, ξ, β, γ.
- For all 34 Olympic swimming events, this means 170 parameters in total.
- Can improve model **robustness** and **predictive power** by pooling information across events.
- Try introducing **distance** as a covariate.

...independent event model

- Have model with with five parameters: μ_0, σ_0, ξ, β, γ.
- For all 34 Olympic swimming events, this means 170 parameters in total.
- Can improve model **robustness** and **predictive power** by pooling information across events.
- Try introducing **distance** as a covariate.
- A log-log relationship with distance works well for athletics, but..

Across event model

- The different strokes means this isn't much of an improvement.

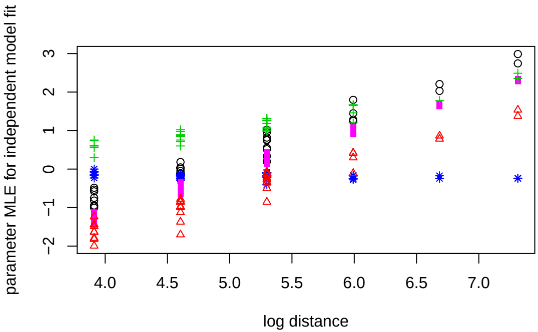


Figure: Transformed parameter estimates against log threshold swim-time $\log(u)$: $\log(\tilde{\sigma})$, ξ, $\log(\beta)$, $\log(-\mu_0)$, $\sqrt{\gamma}$.

Across event model

Therefore, use the threshold time as a covariate:

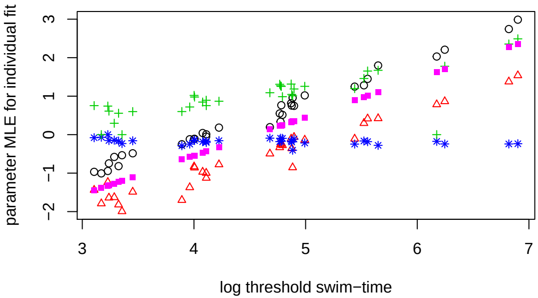


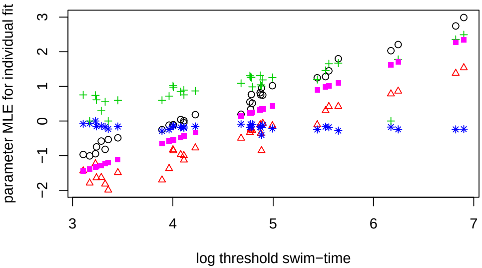
Figure: Transformed parameter estimates against log threshold swim-time, $\log(u)$: $\log(\tilde{\sigma})$, ξ, $\log(\beta)$, $\log(-\mu_0)$, $\sqrt{\gamma}$.

Across event model

These pooled estimates would result in model:

$$\begin{aligned} \xi^{(e)} &= \xi, \\ \log\left(\mu^{(e)}\right) &= \alpha_1 + \theta_1 \tilde{u}_e, \\ \log\left(\beta^{(e)}\right) &= \alpha_2 + \theta_2 \tilde{u}_e, \\ \sqrt{\gamma^{(e)}} &= \alpha_3 + \theta_3 \tilde{u}_e, \\ \log\left(\tilde{\sigma}_u^{(e)}\right) &= \alpha_4 + \theta_4 \tilde{u}_e, \end{aligned}$$

$\tilde{u}_e = \log(u_e)$.

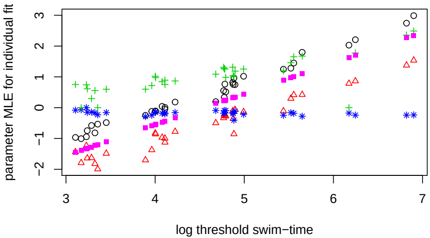


Across event model

These pooled estimates would result in model:

$$\xi^{(e)} = \xi,$$
$$\log\left(\mu^{(e)}\right) = \alpha_1 + \theta_1 \tilde{u}_e,$$
$$\log\left(\beta^{(e)}\right) = \alpha_2 + \theta_2 \tilde{u}_e,$$
$$\sqrt{\gamma^{(e)}} = \alpha_3 + \theta_3 \tilde{u}_e,$$
$$\boldsymbol{\log\left(\tilde{\sigma}_u^{(e)}\right) = \alpha_4 + \theta_4 \tilde{u}_e},$$

$\tilde{u}_e = \log(u_e)$.



Splines

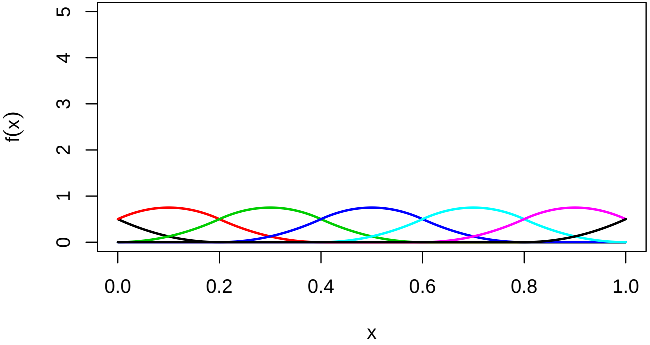


Figure: different degree B-spline.

Splines

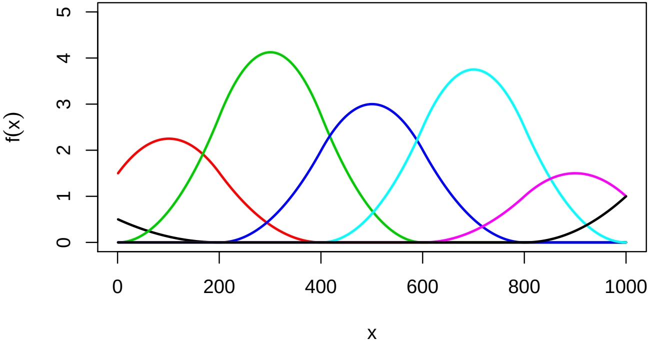


Figure: different degree B-spline.

Splines

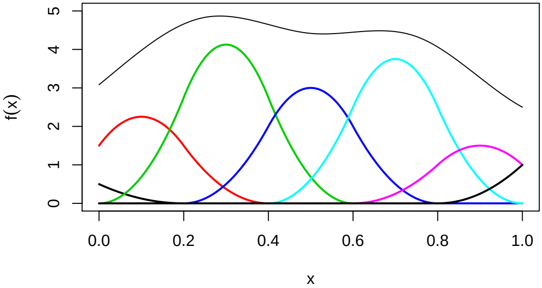


Figure: different degree B-spline.

$$\log\left(\tilde{\sigma}(\tilde{u})\right) = \sum_{k=1}^{q} a_k B_k(\tilde{u})$$

where a_k is the k^{th} element of the spline coefficient vector $\boldsymbol{a}$.

Final model

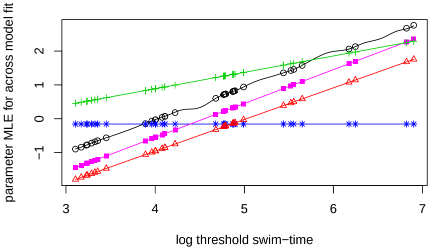


Figure: parameters for the final model: $\log(\tilde{\sigma})$, ξ, $\log(\beta)$, $\log(-\mu_0)$, $\sqrt{\gamma}$.

$$
\begin{aligned}
\xi^{(e)} &= \xi, \\
\log\left(-\mu^{(e)}\right) &= \alpha_1 + \theta_1 \tilde{u}_e, \\
\log\left(\beta^{(e)}\right) &= \alpha_2 + \theta_2 \tilde{u}_e, \\
\sqrt{\gamma^{(e)}} &= \alpha_3 + \theta_3 \tilde{u}_e, \\
\log\left(\tilde{\sigma}_u^{(e)}\right) &= \sum_{k=1}^{q} a_k B_k(\tilde{u}_e),
\end{aligned}
$$

Model fits: rate of observations

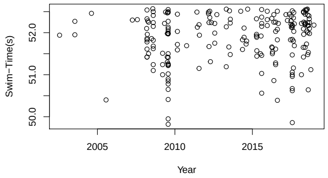


Figure: Data for the men's 100m butterfly.

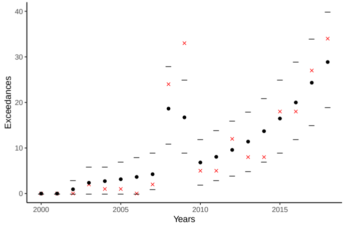


Figure: Fitted rate of occurances.

Model fits: distribution of observations.

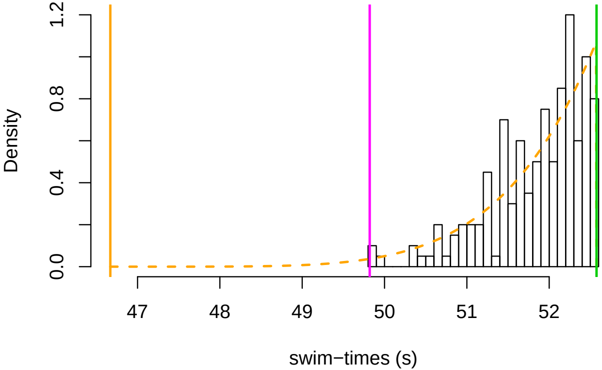


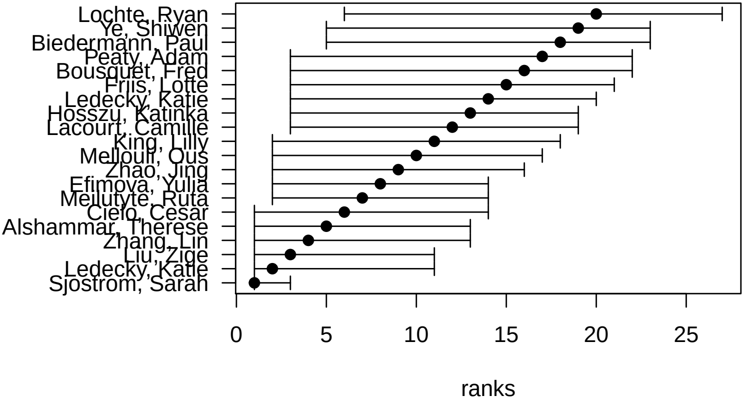
Figure: Fitted distribution above threshold, men's 100m butterfly: Threshold, World record, Ultimate possible swim-time.

Results: Rankings

Can define ranking based on $\Pr\{X_e > x\}$, which gives:

Results: Rankings

Can define ranking based on $\Pr\{X_e > x\}$, which gives:



Results: Expected times

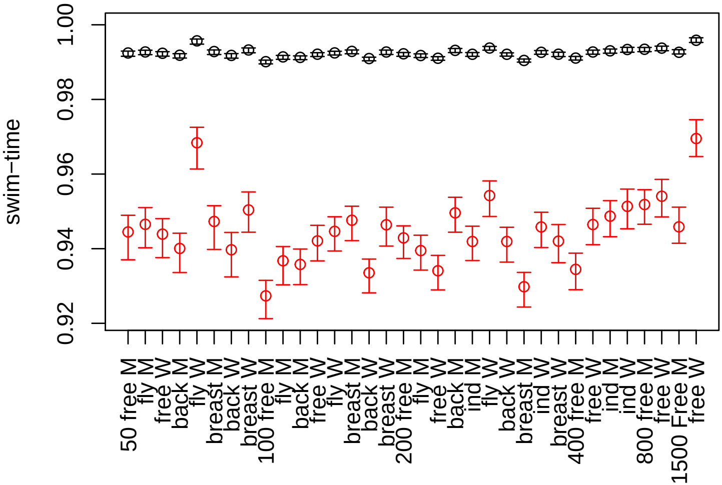
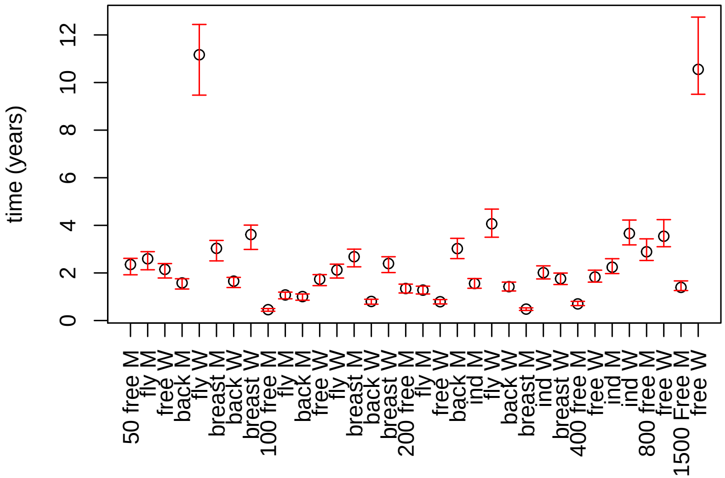
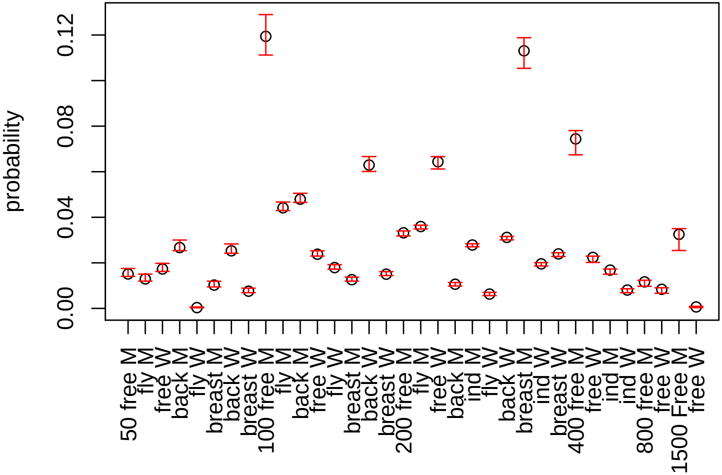


Figure: Expected next world record swim-time. Ultimate possible swim-time

Results: Time of next record



Results: Probability of next record in a given event



Thanks

STOR-i

Lancaster University

...derivations

If $r_e := \max(\boldsymbol{x}_e)$, then

$$
\begin{aligned}
\Pr\{T^{(e)} < t\} &= 1 - \Pr\{T^{(e)} > t\} \qquad (1)\\
&= 1 - \sum_{m=0}^{\infty} \Pr\{\max(X^{(e)}_{1:N_t}) < r_e | N_t = m\} \Pr\{N_t = m\}\\
&= 1 - \sum_{m=0}^{\infty} \left[H^{(e)}_{u_e}(r_e)\right]^m \left[\Lambda^{(e)}\left(\mathcal{A}_{(1,t),u_e}\right)\right]^m \exp\left[-\Lambda^{(e)}\left(\mathcal{A}_{(1,t),u_e}\right)\right] / m!\\
&= 1 - \exp\left\{-\Lambda^{(e)}\left(\mathcal{A}_{(1,t),u_e}\right) \bar{H}^{(e)}_{u_e}(r_e)\right\}\\
&= F_{T^{(e)}}(t),
\end{aligned}
$$

...derivations

Let

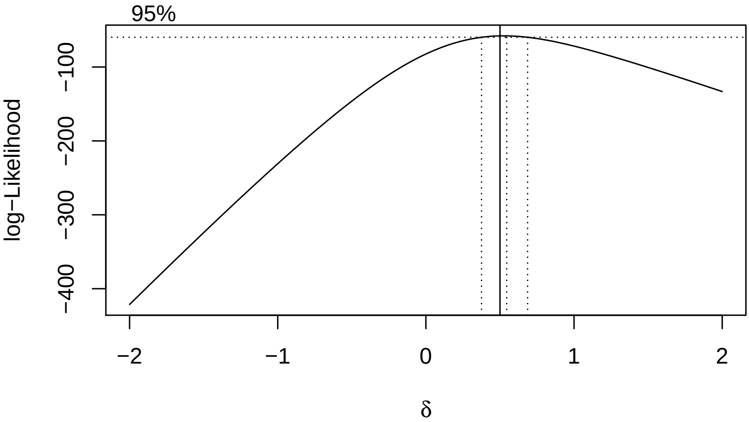
$$T^{(-e)} := \min_k \{ T^{(k)}, k \in E \setminus \{e\} \}.$$

Then the probability that the next world record is set in event e is

$$\begin{aligned}
&\Pr\{T^{(-e)} > T^{(e)}\} \qquad (2)\\
&= \int_1^\infty \Pr\{T^{(-e)} > T^{(e)} | T^{(e)} = t\} \Pr\{T^{(e)} = t\}\, \mathrm{d}t \\
&= \int_1^\infty \prod_{k\in E\setminus\{e\}} \left\{ \exp\left[-\Lambda^{(k)} \left(\mathcal{A}_{(1,t),u_k}\right) \bar{H}^{(k)}_{u_k}(r_k) \right] \right\} \\
&\quad \left[1 + \xi \left(\frac{u_e - \mu^{(e)}(t)}{\sigma^{(e)}(t)} \right) \right]_+^{-\frac{1}{\xi}} \bar{H}^{(e)}_{u_e}(r_e) \exp\left[-\Lambda^{(e)} \left(\mathcal{A}_{(1,t),u_e}\right) \bar{H}^{(e)}_{u_e}(r_e) \right] \mathrm{d}t \\
&= \int_1^\infty \left\{ \exp\left[-\sum_{k\in E} \Lambda^{(k)} \left(\mathcal{A}_{(1,t),u_k}\right) \bar{H}^{(k)}_{u_k}(r_e) \right] \right\} \left[1 + \xi \left(\frac{u_e - \mu^{(e)}(t)}{\sigma^{(e)}(t)} \right) \right]_+^{-\frac{1}{\xi}} \bar{H}^{(e)}_{u_e}(r_e)
\end{aligned}$$

where the second equality follows because

$$\Pr\{T^{(-e)} > T^{(e)} | T^{(e)} = t\} = \prod_{k\in E\setminus\{e\}} \left\{ \exp\left[-(\Lambda^{(k)} \left(\mathcal{A}_{(1,t),u_k}\right) \bar{H}^{(k)}_{u_k}(r_k) \right] \right\}$$


95%
log−Likelihood
−100
−200
−300
−400
−2
−1
0
1
2
δ

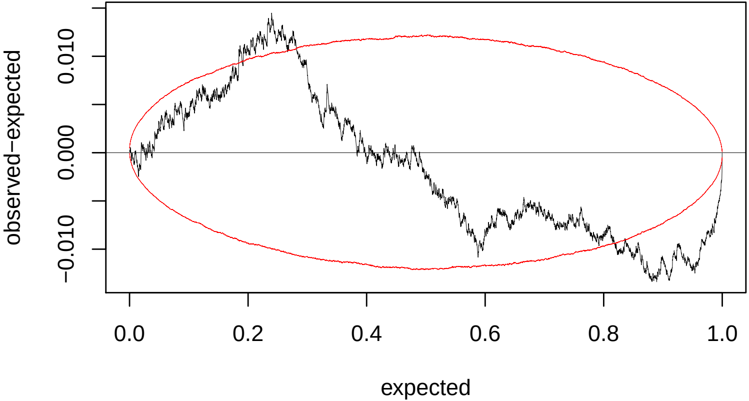


Figure: Pooled pp plot over all events, with 95 % tolerance interval.

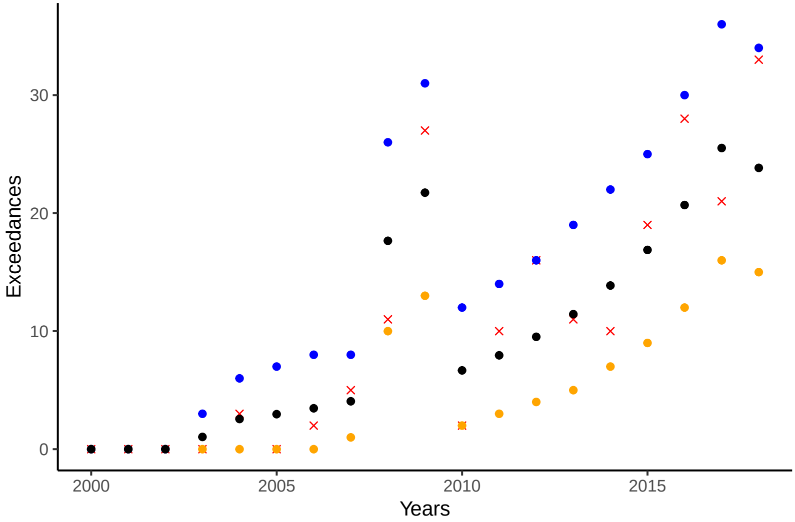


Figure: The expected number of observations for Women's 100m freestyle and 95% confidence intervals, against the number of observations in the data (red crosses).

$$L(\theta; \boldsymbol{x}, \boldsymbol{t}) = \prod_{e \in E} \left\{ \exp\left[-\Lambda^{(e)}\left(\mathcal{A}_{1,u_e}\right)\right] \prod_{i=1}^{200} \int_{x_i^{(e)} - s_i/2}^{x_i^{(e)} + s_i/2} \lambda^{(e)}(t_i, x)\, \mathrm{d}x \right\} \exp\left[-(\phi_1 \boldsymbol{p}_r + \phi_2 \boldsymbol{p}_m)\right]$$

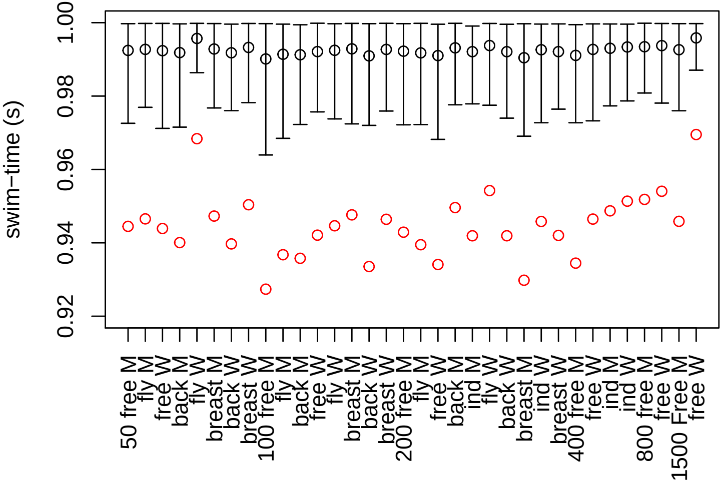


Figure: Expected next world record swim-time. Ultimate possible swim-time