Is Ruth Chepngetich's Marathon World Record Valid? A Statistical Analysis of Elite Marathon Performances

Abigail Mabe and Richard Smith

Department of Statistics and Operations Research

University of North Carolina, Chapel Hill, USA

STOR 834, Spring 2025

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

I. Motivation and Background

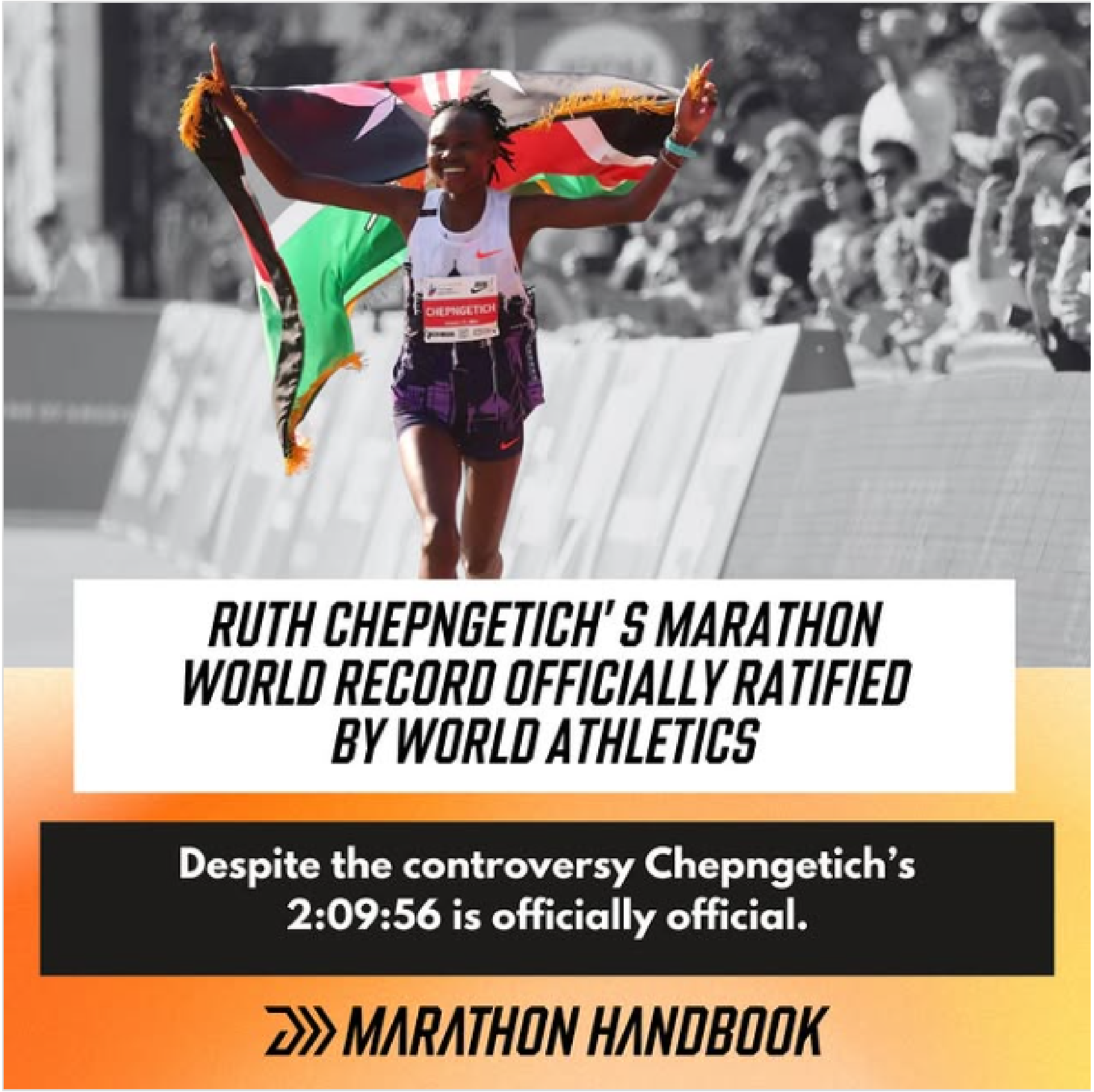
- On October 13, 2024, the Kenyan runner Ruth Chepngetich ran a women's world record for the marathon of 2 hours, 9 minutes, 56 seconds

- This broke the previous world record by nearly 2 minutes and Chepngetich's personal best by nearly 5 minutes

- Many commentators have raised questions about illegal drug use

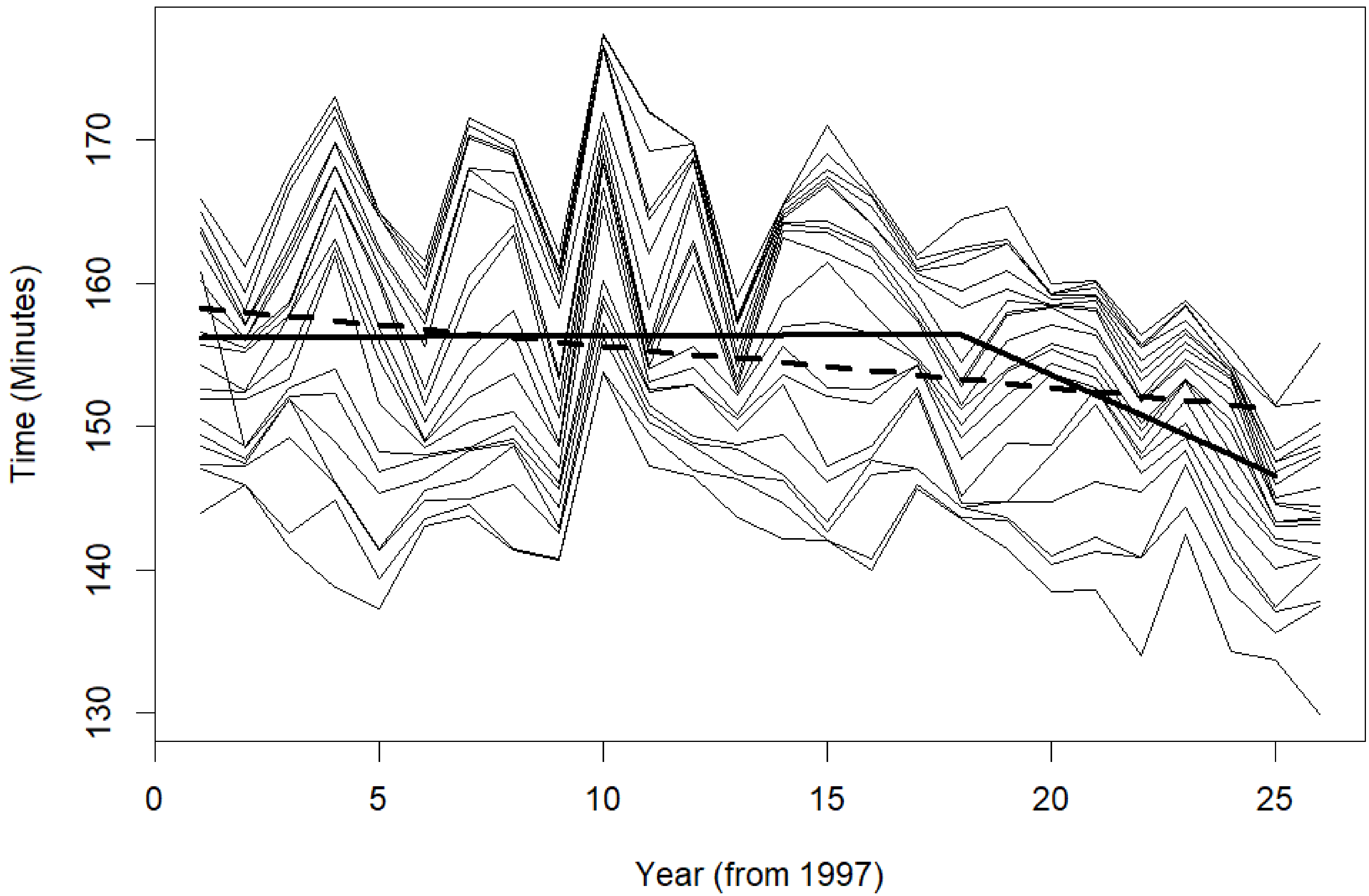
- However, other commentators defended the new record as fully plausible, and World Athletics ratified the record

- This project proposes a statistical analysis to assess the plausibility of this performance.


CHEPNGETICH
RUTH CHEPNGETICH' S MARATHON WORLD RECORD OFFICIALLY RATIFIED BY WORLD ATHLETICS
Despite the controversy Chepngetich's 2:09:56 is officially official.
MARATHON HANDBOOK

II. Statistical Analysis

- We downloaded data from the Chicago Marathon website, best 20 women's times for each year from 1998–2024
- No race 2020 — we just left out that year
- 2007 is a possible outlier but we discuss that later
- Clear downward trend over the 27 years but strong suspicion of a changepoint in the mid-2010s
- Particular interest in a changepoint in 2016 as Nike Vaporfly shoes were only introduced in 2017



Time (Minutes)
130
140
150
160
170
0
5
10
15
20
25
Year (from 1997)

Statistical Model

- Generalized Extreme Value (GEV) distribution for sample minima

$$\Pr\{Y \le y\} \;=\; G(y;\mu,\sigma,\xi) \;=\; 1-\exp\left[-\left\{1+\xi\cdot\frac{\mu-y}{\sigma}\right\}_{+}^{-1/\xi}\right]$$

  where Y is the variable of interest (in this case, the minimum running time in a marathon), y is some real number where the distribution is to be evaluated, and μ, σ and ξ are the parameters of the distribution ($x_+ = \max(x,0)$)

- Location parameter μ
- Scale parameter σ
- Shape parameter ξ — when $\xi = 0$ the formula reduces to

$$G(y;\mu,\sigma,0) \;=\; 1-\exp\left[-\exp\left\{-\frac{\mu-y}{\sigma}\right\}\right]$$

  However for running times data we often find $\xi \approx -0.5$ corresponding to a distribution with a very short left-hand tail

- In practice often use model of form $Y_t \sim G(\cdot;\mu_t,\sigma_t,\xi_t)$ where parameters depend on year t

r largest (smallest) order statistics model (Coles 2001, equation (3.15))

- Let $z^{(1)} \leq z^{(2)} \leq \ldots \leq z^{(r)}$ denote the r best (i.e. smallest) running times in a given year. We make the following assumption (derived from extreme value theory) about the joint density:

$$\begin{aligned} f_r(z^{(1)}, \ldots, z^{(r)}) &= \exp\left\{ -\left[1 + \xi \left(\frac{\mu - z^{(r)}}{\sigma} \right) \right]^{-1/\xi} \right\} \cdot \\ &\quad \cdot \prod_{k=1}^{r} \left\{ \sigma^{-1} \left[1 + \xi \left(\frac{\mu - z^{(k)}}{\sigma} \right) \right]^{-1/\xi - 1} \right\}. \end{aligned}$$

  Assume $1 + \xi \left(\frac{\mu - z^{(k)}}{\sigma} \right) > 0$ for $k = 1, \ldots, r$.

- This reduces to the GEV density $\frac{\partial G(y; \mu, \sigma, \xi)}{\partial y}$ with $y = z^{(1)}$, when $r = 1$.
- Trend models: keep σ, ξ constant but allow $\mu = \mu_t$ for various possible functions μ_t
- The plan: fit a model to the data 1998–2023, use it to predict performances in 2024

Model (a): Constant location parameter

$\mu_t = \mu$, constant

Parameter	Estimate	S.E.	z-ratio	p-value
μ	140.059	0.539	259.95	0
$\log \sigma$	1.261	0.048	26.34	6.1×10^{-153}
ξ	−0.511	0.037	−13.82	1.9×10^{-43}
NLLH	393.713			

Table of parameter values for model (a), years 1–25, $r = 10$

Model (b): Linear trend in time

$\mu_t = \beta_0 + \beta_1 (t - t_0)$, t_0 a centering time (which we take equal to the intended changepoint)

Parameter	Estimate	S.E.	z-ratio	p-value
β_0	139.335	0.517	269.602	0
$\log \sigma$	1.122	0.055	20.373	2.9×10^{-92}
ξ	−0.578	0.041	−13.953	2.0×10^{-44}
β_1	−0.211	0.036	−5.945	2.8×10^{-9}
NLLH	384.124			

Table of parameter values for model (b), years 1–25, $r = 10$

This shows that the linear trned model significantly improves on the model with constant μ, σ, ξ, but it does not prove that this is the "best" model

Model (c): Changepoint model

$\mu_t = \beta_0 + \beta_1(t - t_0) + \beta_2(t - t_0)_+$, t_0 changepoint (initially $t_0 = 19$, corresponding to year 2016)

Parameter	Estimate	S.E.	z-ratio	p-value
β_0	143.139	0.950	150.702	0
$\log \sigma$	1.031	0.064	16.172	8.0×10^{-59}
ξ	−0.617	0.047	−13.128	2.3×10^{-39}
β_1	0.109	0.082	−1.336	0.182
β_2	−1.731	0.447	−3.874	1.1×10^{-4}
NLLH	376.137			

Table of parameter values for model (c), years 1–25, $r = 10$, $t_0 = 19$

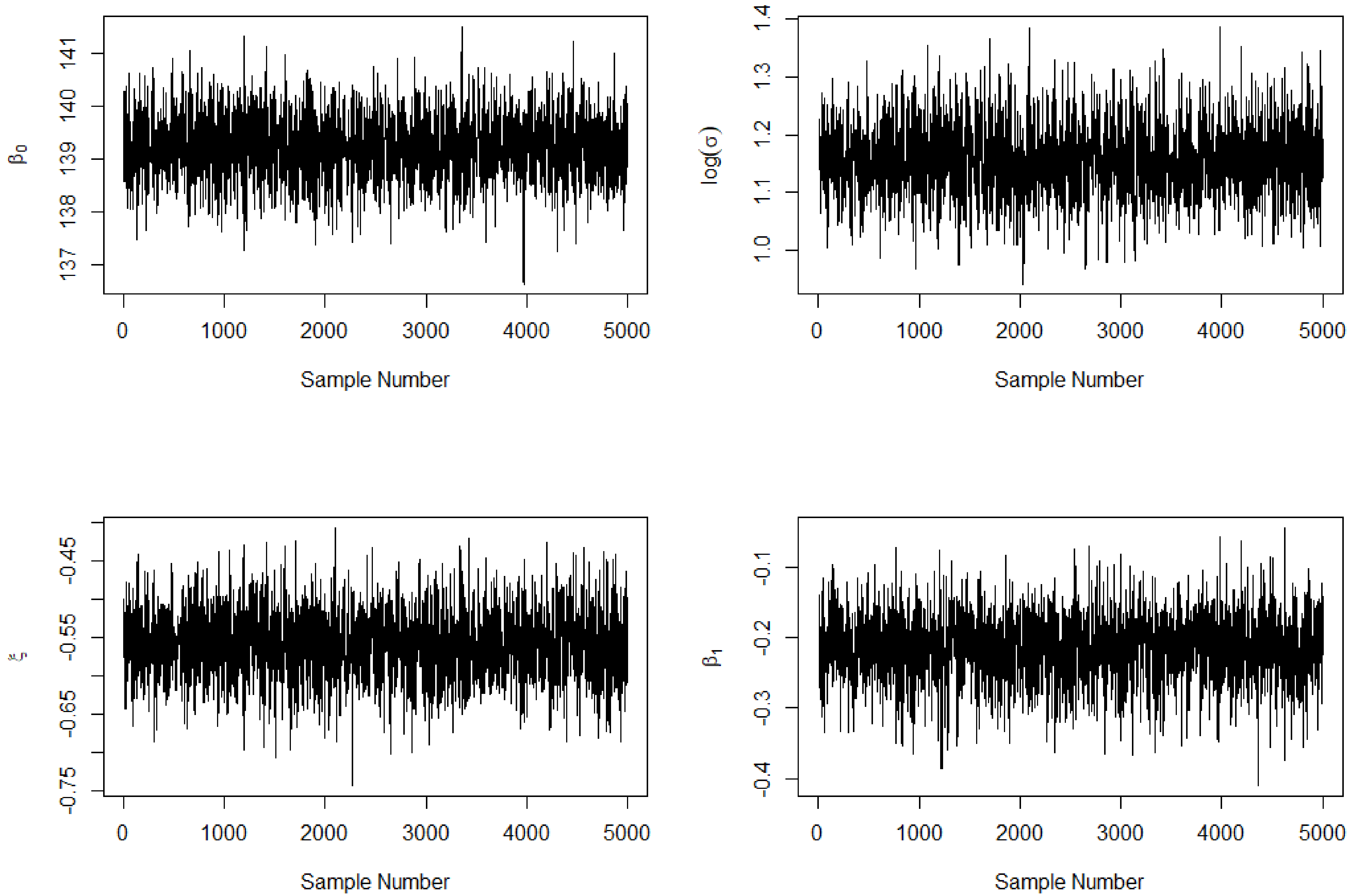
Changepoint model significantly better than linear trend

However, in fact, this is not the best value of t_0. The NLLH is smallest when $t_0 = 13$ (interpretation: it's not all due to the shoes)

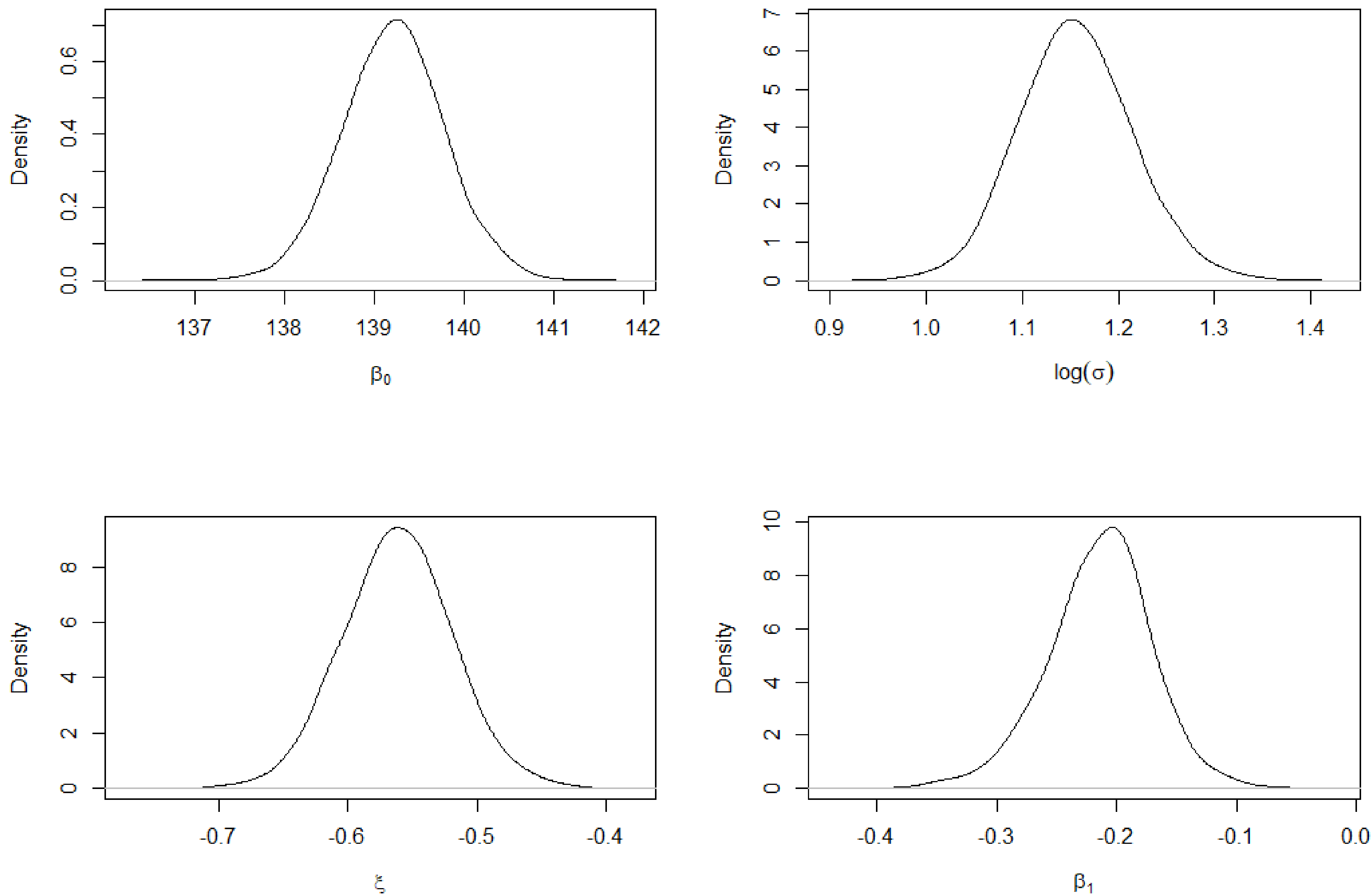
Motivation for a Bayesian Analysis

Two major parameters of interest:

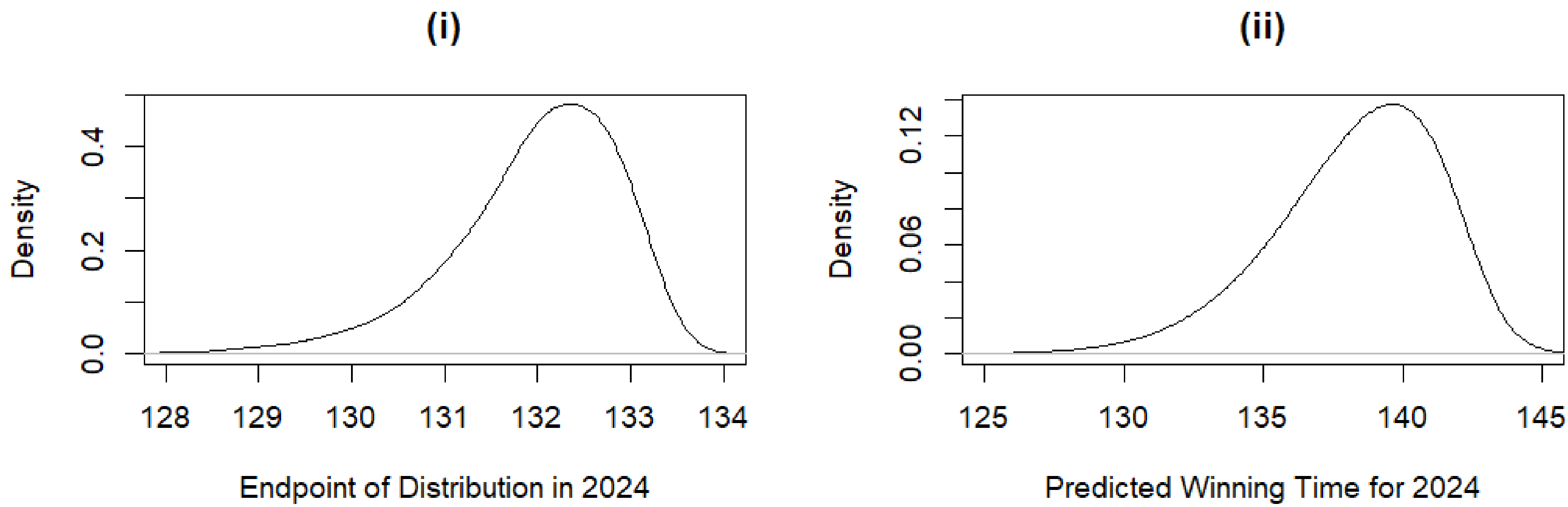
- Endpoint of the distribution: for GEV(μ, σ, ξ) with $\xi < 0$, the lower end-point of the distribution is at $y = \mu + \frac{\sigma}{\xi}$.
- Obvious modification for a trend model $y_t = \mu_t + \frac{\sigma}{\xi}$ but notice this does depend on t
- Do this for year 26 (2024 calendar year — note that we left out 2020 completely) to get projected "best possble time" for 2024
- However a more realistic comparison would be based on the *predicted winning time* for 2024 — sampling from the GEV distribution with fitted $\mu_t,\ \sigma,\ \xi$.
- In this context, a Bayesian calculation allows us to incorporate the incertainty in the GEV parameters
- Run an MCMC



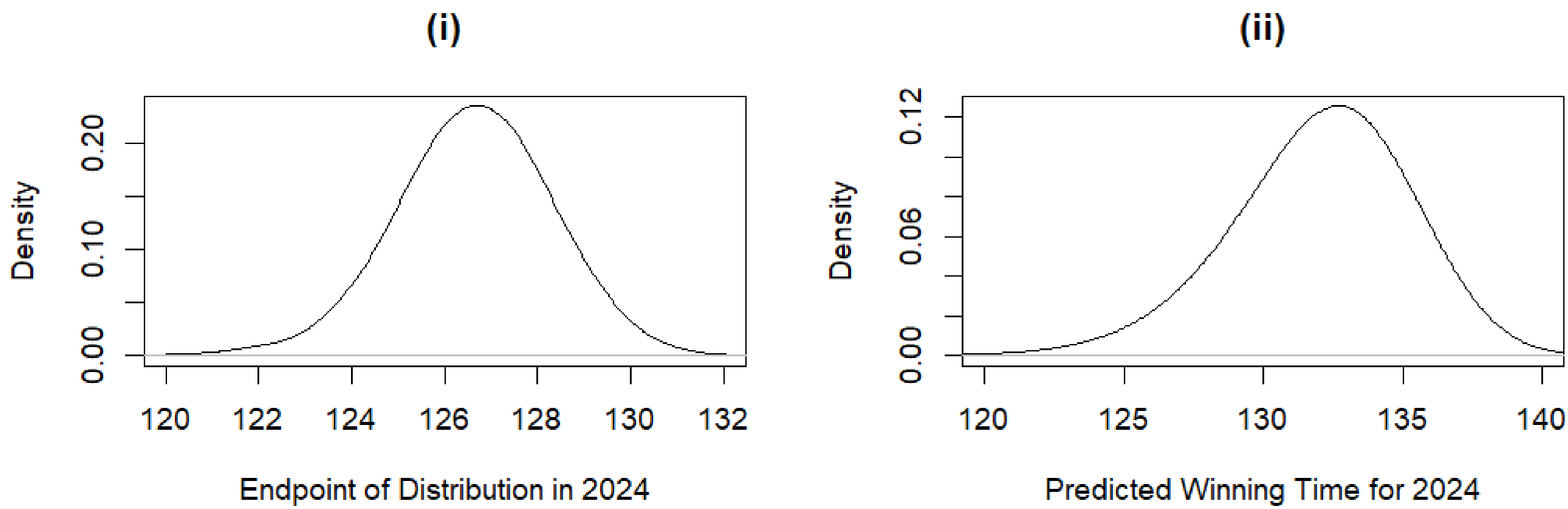
Trace plots for model (b)



Posterior density plots for model (b)



Posterior density plots for the endpoint and the predicted 2024 winning time under model (b). The (estimated) probability that the winning time is < 130 minutes is about 0.011 — "unlikely but not impossible"



Posterior density plots for the endpoint and the predicted 2024 winning time under model (c). The (estimated) probability that the winning time is < 130 minutes is about 0.25 — maybe even too large to be believable?

Sensitivity analysis I

Start Year	Endpoint				Winning Time			
	r=5	r=10	r=15	r=20	r=5	r=10	r=15	r=20
1998	0.17	0.034	0.031	0.091	0.004	0.011	0.013	0.014
2002	0.273	0.045	0.067	0.172	0.006	0.011	0.015	0.016
2005	0.211	0.07	0.121	0.213	0.015	0.022	0.024	0.024
2008	0.433	0.481	0.552	0.627	0.055	0.063	0.062	0.053
2005x	0.198	0.061	0.068	0.141	0.01	0.018	0.021	0.019
1998x	0.188	0.03	0.027	0.084	0.004	0.011	0.013	0.014
1998y	0.181	0.028	0.031	0.078	0.004	0.011	0.014	0.014

Posterior probability that the endpoint or the winning time is under 130 minutes, based on model (b), for four values of r and different starting years.
2005x: analysis starting in 2005 but omitting 2007.
1998x: analysis with changepoint in 2013 instead of 2016.
1998y: analysis with changepoint in 2013 instead of 2016.

Sensitivity analysis II

Start Year	Endpoint				Winning Time			
	r=5	r=10	r=15	r=20	r=5	r=10	r=15	r=20
1998	0.931	0.983	0.992	0.996	0.097	0.25	0.306	0.315
2002	0.967	0.988	0.996	0.996	0.168	0.324	0.392	0.34
2005	0.918	0.964	0.978	0.981	0.182	0.283	0.283	0.256
2008	0.596	0.73	0.771	0.847	0.079	0.116	0.132	0.125
2005x	0.92	0.973	0.976	0.98	0.151	0.262	0.297	0.258
1998x	0.686	0.722	0.796	0.874	0.037	0.069	0.092	0.106
1998y	0.477	0.469	0.584	0.656	0.029	0.057	0.065	0.066

Same as previous table, but based on model (c)

Sensitivity analysis III

I also did some investigation of varying the prior distribution for the Bayesian analysis, but this did not make much difference (we may return to this point later in the course)

However, the results contrast sharply with some earlier analyses based on the performances by Chinese women runners in 1993

Athletics world records blow as Wang Junxia 'admits' being part of Chinese state-sponsored doping regime

Ben Bloom

READ MORE ABOUT:
IAAF (International Association of Athletics Federations), China

04 February 2016 10:47pm GMT

Gift this article free

Wang Junxia (right) currently holds the 10,000m and 3,000m world records Credit: REUTERS

Appl. Statist. (1997)
46, *No.* 1, *pp.* 123–128

Letter to the Editors

Statistics for Exceptional Athletics Records

Robinson and Tawn (1995) analysed data from 1972 to 1992 in the women's 1500 m and 3000 m running events, with a view to establishing whether the remarkable performances achieved by Chinese athletes in 1993 were statistically inconsistent with previous performances, a conclusion that might be taken as evidence of illegal drug use. Particular attention was paid to the performance of Wang Junxia, who improved the 3000 m record from 502.62 s to 486.11 s. For this, they fitted a model to the five best performances by different athletes in each year. They then constructed 90% confidence intervals, under several variants of the model, for x_{ult}, a parameter representing the long-term limit of performance. Although the analysis provides some grounds for regarding Wang's time as extremely unusual, in all cases the reported confidence interval for x_{ult} included her record time, and to this extent the evidence is inconclusive.

In this letter, I argue that a simpler model, based on fitting part of the data without any trend component, produces very similar results to those of Robinson and Tawn with rather less effort. My main point, however, is that if we look at Wang's time from the point of view of prediction intervals for the observed value, rather than confidence intervals for the hypothetical x_{ult} parameter, then we indeed obtain strong evidence that the performance was a severe outlier.

RLS (1997)

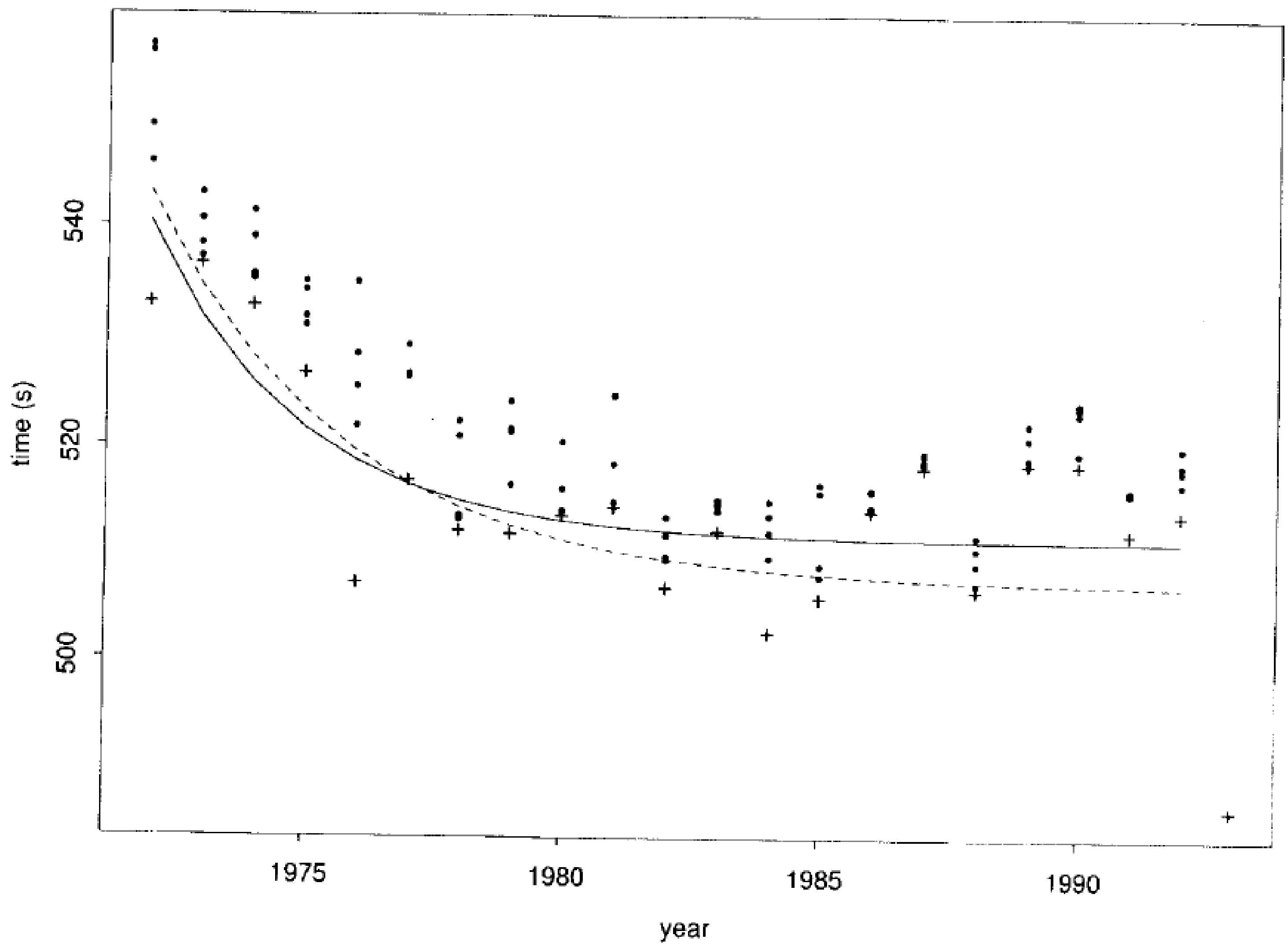


Fig. 1. Five best annual times for the 3000 m: +, annual minima; ———, regression function (1.5), estimated by using model A of Table 1; --------, regression function (1.5), estimated by using model C of Table 1

Robinson and Tawn (1995)

Prediction intervals

For the following analysis it is convenient to reparameterize the GEV distribution in three-parameter Weibull form:

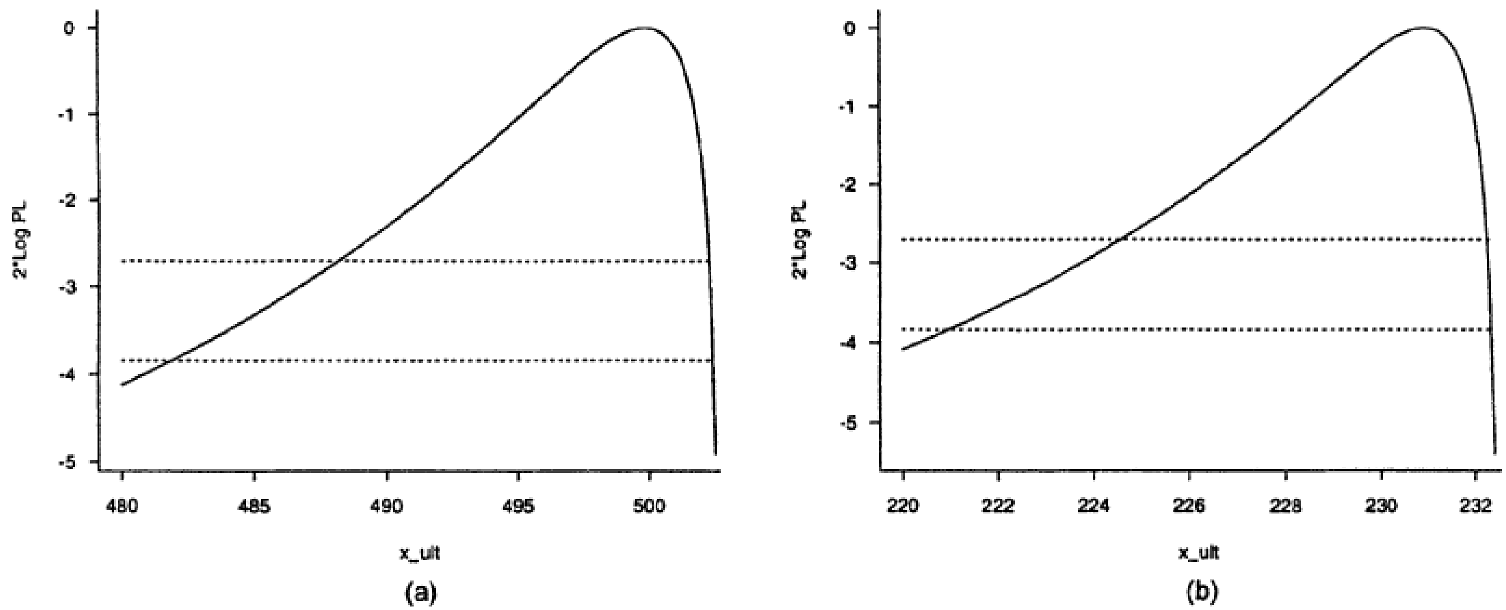


Fig. 1. Profile likelihood plots: (a) 3000 m data; (b) 1500 m data

RLS (1997)

For the 3000 m data based on 1980–92, the posterior probability that $\theta < 486.11$ is 0.076—small, but hardly negligible. In contrast, when averaged according to the posterior distribution of the unknown parameters, the conditional probability of an observed record of less than 486.11, given that a record occurs at all, is 0.00047, which is much smaller.

Plots of the posterior and predictive density, shown in Fig. 2, are indeed of a very different

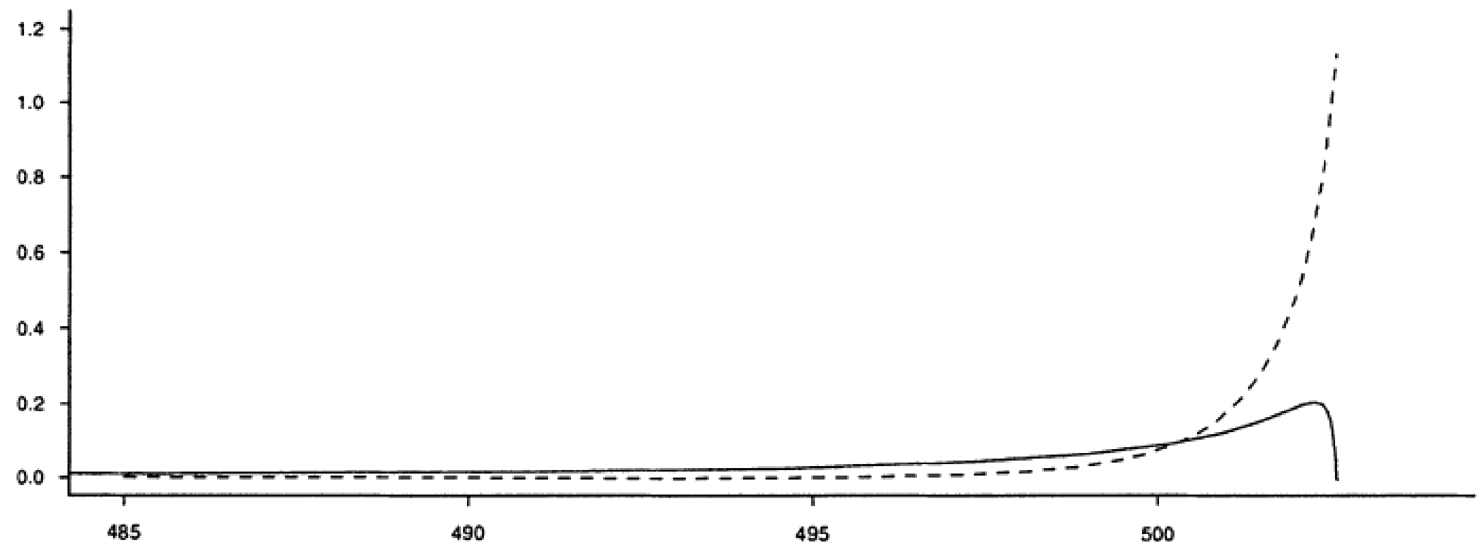


Fig. 2. Bayesian densities: ——— , posterior density; - - - - , predictive density

RLS (1997)

Summary and Conclusions

- Trends in the data are clearly significant, and there is also strong evidence that a changepoint model fits better than a linear trend model, though not necessarily based on 2016 as the changepoint
- For estimated probabilities that the winning time for 2024 is below 130 minutes, conditional on previous times up to 2023, is never exceptionally small (smallest is 0.004, but most of our estimated probabiities are quite a bit larger than that), so we cannot say that Chepngetich's performance is "too good to be true"
- The results contrast strongly with similar analyses for Chinese women runners in 1993
- From an extreme value theory perspective, the key points are:
  - The generalized extreme value distribution (GEV) as a canonical probability model for extremes
  - Extension to the joint distribution of r largest or smallest events
  - Use of trend (covariate) models to handle changes in the distribution (a big issue in climate applications)
  - Inference methods — maximum likelihood and Bayesian, though there are other methods that have been popularized

Overview of the Course

- Basic theory: derivation of extreme value distributions as limiting distributions of extremes in random samples
- How good is the approximation? Rates of convergence, higher order expansions
- Extremes in stationary sequences; how do we cope with dependence?
- Statistical theory: maximum likelihood, Bayesian, other
- Multivariate extremes
- Spatial extremes
- Applications