AMSI SUMMER RESEARCH SCHOLARSHIPS 2025–26

Get a taste for Research this Summer

Record Statistics of Non-Independent and Identically Distributed Random Variables

Angus Stewart
Supervised by Prof. Georgy Sofronov
Macquarie University

February 27, 2026

Jane Street®
AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Contents

1 Abstract

This project investigates record statistics in both independent and identically distributed (i.i.d.) sequences of random variables and a non-i.i.d. setting known as the linear drift model which incorporates a linear trend in the sequence of random variables. After outlining the well-known results for records of i.i.d. sequences, attention is turned to the behaviour of the time of the second record, $T(2)$. In the i.i.d. case, the time of the second record follows a discrete Pareto distribution with an infinite expectation. In the linear drift model with an underlying exponential distribution, no closed-form expression is available for the distribution of the time of the second record and, instead, numerical methods and simulations must be used. The results show that introducing a positive drift makes upper records more frequent, reducing the expectation of $T(2)$ to a finite value, while a negative drift preserves the heavy-tailed behaviour and infinite mean. This demonstrates how even a simple non-i.i.d. modification can substantially alter record-setting dynamics.

2 Introduction

Records arise in many contexts, from Olympic performances in the ongoing Winter Olympics, climate-related weather extremes with their frequency increased by climate change or in financial and insurance applications. Their behaviour in random sequences is closely tied to order statistics and extreme value theory and a broad set of results have been proven particularly for i.i.d. sequences.

However, far less is known about the non-i.i.d. case which requires more dedicated analysis for each specific sequence of random variables, unlike the i.i.d. case with its very broad results. As such the focus of this research was on developing results for the non-i.i.d. linear drift model. The linear drift model incorporates a linear trend into the sequence of random variables making it a more realistic model than the i.i.d. model which assumes a constant mean value. For example, this would allow for a linear trend in temperature increase, as caused by climate change, while investigating maximum daily temperature records.

This report first establishes a number of key results in the simplest case, that of records coming from a sequence of i.i.d. random variables. Then, for one particular non-i.i.d. case, the linear drift model with an underlying exponential distribution, the time of the second record is investigated through numerical approximations and simulations after a closed form expression was unable to be derived. It was found that the time of the second record follows a discrete Pareto distribution in the i.i.d. case exhibiting heavy tailed-ness and an infinite expectation but by introducing a positive linear drift the expectation is reduced to a finite value.

3 Statement of Authorship

The first part of this report draws on two literature reviews, Arnold, Balakrishnan, and Nagraja (1998) and Wergen (2013), of the well known results for records of sequences of i.i.d. random variables. The second part of the report was developed independently with the linear drift model being introduced in Stepanov (2022) but

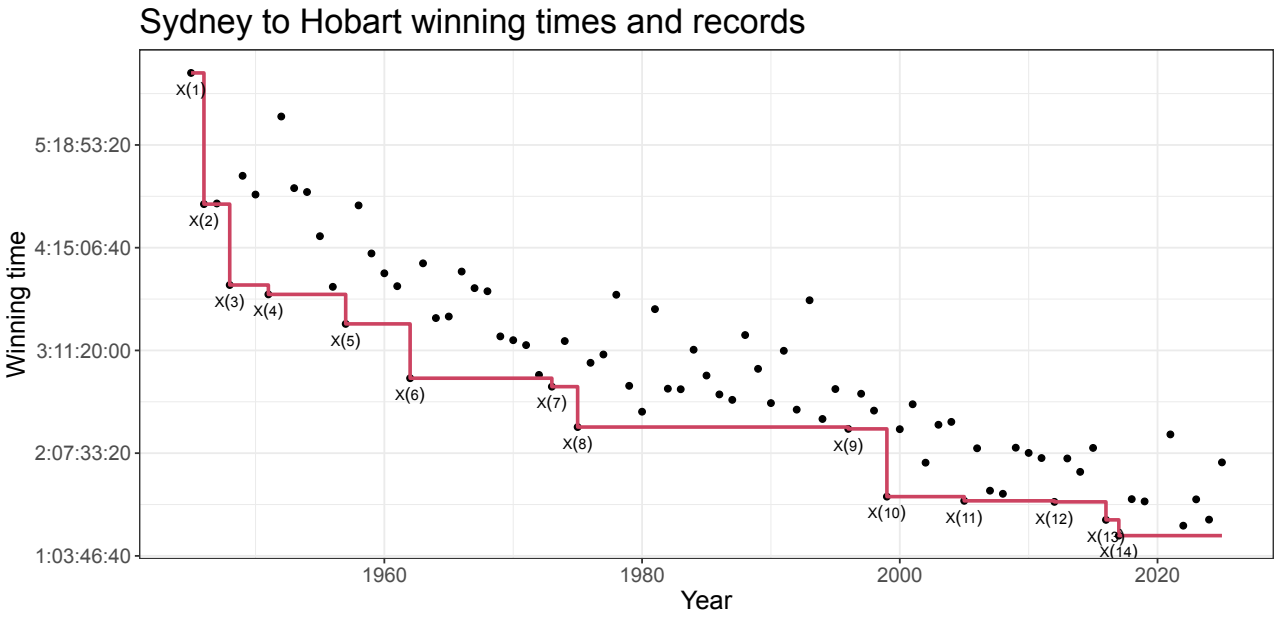

Figure 1: Progression of winning times from the Sydney to Hobart yacht race with the 14 records labelled.

without any mention of its time of the second record where all results were derived by the author of this report themself with assistance from their supervisor.

4 Example and Definitions

Let us start with a motivating example, the progression of lower records of the winning times from the Sydney to Hobart yacht race as shown in Figure 1. We have a sequence of continuous random variables in discrete time steps,

$$\{X_1, X_2, \ldots, X_n\},$$

where in this case each time step is one year and $n = 80$ since the race has been held for the past 80 years.

We define any X_k in the sequence as a lower record if,

$$X_k < \min\{X_1, X_2, \ldots, X_{k-1}\},$$

and, equivalently, as an upper record if

$$X_k > \max\{X_1, X_2, \ldots, X_{k-1}\}.$$

Outside of the Sydney to Hobart example any reference to a record will refer to an upper record.

Define $N(n)$ as the number of records that have been observed after n many observations of the original sequence of random variables. The sequence of record values coming from the initial sequence is denoted as,

$$\{X(1), X(2), \ldots, X(N(n))\},$$

and the time of these records in the original sequence is denoted as

$$\{T(1), T(2), \ldots, T(n)\},$$

such that the n^{th} record occurs at time $T(n)$ in the original random sequence or,

$$X_{T(n)} = X(n),$$

noting that we will always have $T(1) = 1$ since the first observation is always the first record value.

The Sydney to Hobart's first winner was the yacht *Rani* with a time of 6 days, 14 hours and 22 minutes meaning $T(1) = 158.37$ hours and the most recent record holder is *Comanche* with a time of 1 day, 9 hours, 15 minutes and 24 seconds so $T(14) = 33.26$ hours. The progression of the race record is shown by the red line in Figure 1 and exhibits typical features of record sequences with sporadic jumps varying from quite large decreases in the early years with much smaller and less frequent improvements in the modern era. The race record has been set 14 times throughout its 80 year history so $N(14) = 80$.

To determine whether each X_k is a record the indicator variable will be useful,

$$\xi_k = \begin{cases} 1 & X_k \text{ is a record,} \\ 0 & \text{otherwise.} \end{cases}$$

Then define P_k as the probability that X_k is a record event. That is,

$$P_k = \mathbb{P}(\xi_k = 1).$$

This is the first of the properties of records that will be investigated in the following section of the report.

5 Records of a Sequence of i.i.d. Random Variables

Throughout this section we deal with a sequence of i.i.d. random variables and its associated records. From this single assumption, a large variety of results can be derived with no information on the underlying distribution of the individual observations.

From the identical nature of each random variable in the sequence, they all have the same probability density function (pdf) and cumulative distribution function (cdf). This will be notated as $f(x)$ and $F(x)$, respectively, throughout this section.

5.1 Probability of a Record Occurring

We have that the probability of a record occurring at time k is $\frac{1}{k}$. This can be seen intuitively because of the identical nature of each observation. This means that after k observations each of them has an equal chance of being the largest seen so far giving $P_k = \frac{1}{k}$. Alternatively, this can also be more rigorously shown by integrating across the joint distribution of the observations.

Proof

$$
\begin{aligned}
P_k &= \mathbb{P}\left[X_k > \max\left\{X_1, X_2, \ldots, X_{k-1}\right\}\right] \\
&= \int_{-\infty}^{\infty} f(x)F^{k-1}(x)\,dx \\
&\qquad \boxed{\begin{array}{ll} u = F^{k-1}(x) & v = F(x) \\ du = (k-1)F^{k-2}(x)f(x)\,dx & dv = f(x)\,dx \end{array}} \\
&= 1 - 0 - (k-1)\int_{-\infty}^{\infty} F^{k-1}(x)f(x)\,dx \\
&= 1 - (k-1)P_k, \\
kP_k &= 1, \\
P_k &= \frac{1}{k}.
\end{aligned}
$$

5.2 Records are Independent

Although we have that the original sequence of random variables is independent, indeed it is both independent and identically distributed, we have not yet established any such result for the sequence of record values. The sequence of records is independent if the probability of both X_k and X_l being records, denoted as $P_{k,l}$, is equal to P_kP_l. This is hard to show for general k, l (Arnold, Balakrishnan, and Nagraja 1998) but for k and $k+1$ it is relatively simple and is done so below guided by Wergen (2013).

Proof

$$
\begin{aligned}
P_{k,k+1} &= \mathbb{P}\left[X_k > \max\left\{X_1, X_2, \ldots, X_{k-1}\right\} \cap X_{k+1} > X_k\right] \\
&= \int_{-\infty}^{\infty} f(x_{k+1}) \int_{-\infty}^{x_{k+1}} f(x_k)F^{k-1}(x_k)\,dx_k\,dx_{k+1} \\
&\qquad \boxed{\begin{array}{l} u = F(x_k) \Rightarrow u' = f(x_k) \\ x_k = x_{k+1} \Rightarrow u = F(x_{k+1}) \\ x_k = -\infty \Rightarrow u = 0 \end{array}} \\
&= \int_{-\infty}^{\infty} f(x_{k+1}) \int_{0}^{F(x_{k+1})} u^{k-1}\,du\,dx_{k+1} \\
&= \frac{1}{k}\int_{-\infty}^{\infty} f(x_{k+1})F^{k}(x_{k+1})\,dx_{k+1} \\
&= \frac{1}{k}P_{k+1} \\
&= P_kP_{k+1}.
\end{aligned}
$$

5.3 Distribution of Record Values

It also turns out that an expression relying only on the pdf and cdf of the underlying distribution followed by each X_i can be obtained for the distribution of record values of a particular record number (Arnold, Balakrishnan, and Nagraja 1998). Let this distribution's pdf be denoted $f_k(x)$ and we can write $X(k) \sim F_k$ meaning the k^{th} record value is distributed according to the cdf F_k.

Let us begin by taking a sequence of i.i.d. random variables coming from a standard exponential distribution which has a pdf given by $f_X(x) = e^{-x}$ for $x > 0$, that is $X \sim \text{Exp}(1)$. The key property of the exponential distribution used here is its "memoryless" property where if it is given that x is above some threshold value, say $\tilde{x}$, then the distribution of that random variable will still be exponentially distributed, that is $f_{X|X>\tilde{x}}(x) = e^{-(x-\tilde{x})}$.

Using this and convoluting $f_{k-1}(x)$ and $f_k(x)$ (adding the $(k-1)$th record and a random exponential sample), one can obtain that

$$f_k(x) = \frac{1}{(k-1)!} x^{-k} e^{-x},$$

for the exponential distribution.

By expressing an arbitrary $f(x)$ in terms of an exponential distribution, this can then be expanded to a formula for most common distributions,

$$f_k(x) = \frac{1}{(k-1)!} \left(-\ln(1-F(x))\right)^k f(x).$$

5.4 Time of Records Occurring Form a Markov Chain

We now turn to what the bulk of this Summer research has been focused on. Recall that $T(n)$ is the time of observation of the n^{th} record in the original sequence $\{X_1, X_2, \ldots, X_n\}$. More formally, we have that

$$T(1) = 1, \quad T(n+1) = \min\left\{j : j > T(n), X_j > X_{T(n)}\right\},$$

such that $X(n) = X_{T(n)}$.

Now consider the sequence $\{T(1), T(2), \ldots, T(n)\}$ which forms a time-homogeneous Markov chain (Stepanov 2022) with a discrete time component and discrete state space. Thus, the distribution of $T(n)$, when the n^{th} record occurs, depends only on the distribution of the $(n-1)^{\text{th}}$ record and is given by the following.

Proof

$$\begin{aligned}
\mathbb{P}\left(T(n) = k \,|\, T(n-1) = j\right) &= \mathbb{P}\left(\xi_{j+1} = 0, \xi_{j+2} = 0, \ldots, \xi_{k-1} = 0, \xi_k = 1\right) \\
&= \left(1 - \frac{1}{j+1}\right)\left(1 - \frac{1}{j+2}\right)\left(\ldots\right)\left(1 - \frac{1}{k-1}\right)\left(1 - \frac{1}{k}\right) \\
&= \frac{j}{j+1} \times \frac{j+1}{j+2} \times \cdots \times \frac{k-2}{k-1} \times \frac{1}{k} \\
&= \frac{j}{(k-1)k}.
\end{aligned}$$

5.5 Time of the Second Record

Let us now consider the time of the second record, this is $T(2)$. Using the independence assumption together with the probability of each observation being a record makes finding its distribution simple. First, we find its probability mass function (pmf) or just the probability of $T(2)$ taking any particular value (restricted to integers of 2 or greater).

Proof

$$\begin{aligned}
\mathbb{P}(T(2)=t) &= \mathbb{P}(\xi_1=1,\,\xi_2=0,\,\ldots,\,\xi_{t-1}=0,\,\xi_t=1)\\
&= \mathbb{P}(\xi_1=1)\mathbb{P}(\xi_2=0)\ldots\mathbb{P}(\xi_{t-1}=0)\mathbb{P}(\xi_t=1)\\
&= (1)\left(1-\frac{1}{2}\right)\left(\ldots\right)\left(1-\frac{1}{t-1}\right)\left(\frac{1}{t}\right)\\
&= \frac{1}{2}\times\frac{2}{3}\times\frac{3}{4}\times\cdots\times\frac{t-2}{t-1}\times\frac{1}{t}\\
&= \frac{1}{(t-1)t}\\
&= \frac{1}{t-1}-\frac{1}{t} \quad \text{for } t=2,3,\ldots.
\end{aligned}$$

We can then prove the claim made in my research proposal that the expected waiting time for a record to occur (apart from the trivial first one) is infinite (Stepanov 2022).

Proof

$$\begin{aligned}
\mathbb{E}[T(2)] &= \sum_{t=2}^{\infty}\frac{t}{(t-1)t}\\
&= \sum_{t=2}^{\infty}\frac{1}{t-1}\\
&= \sum_{t=1}^{\infty}\frac{1}{t}\\
&= \infty.
\end{aligned}$$

Having an infinite mean seems paradoxical for $T(2)$ when a whole 50% of the time $T(2)$ will take the value 2, as 50% of the time we will have $X_2 > X_1$ in this i.i.d. case. This indicates the extreme heavy tailed-ness of $T(2)$ with enough extreme results to outweigh the high probability of $T(2) = 2$ and make the expectation infinite.

It should also be checked that the above pmf of $T(2)$ is a valid pmf by ensuring that it sums to 1.

Proof

$$\begin{aligned}\sum_{t=2}^{\infty}\mathbb{P}\left(T(2)=t\right) &= \sum_{t=2}^{\infty}\frac{1}{(t-1)t}\\ &= \sum_{t=2}^{\infty}\left[\frac{1}{t-1}-\frac{1}{t}\right]\\ &= 1-\frac{1}{2}+\frac{1}{2}-\frac{1}{3}+\frac{1}{3}-\dots\\ &= 1.\end{aligned}$$

We can also find the cdf of $T(2)$.

Proof

$$\begin{aligned}F_{T(2)}(t) &= \mathbb{P}\left(T(2)\leq t\right)\\ &= \sum_{x=2}^{t}\left[\frac{1}{x-1}-\frac{1}{x}\right]\\ &= 1-\frac{1}{2}+\frac{1}{2}-\dots+\frac{1}{t-1}-\frac{1}{t}\\ &= \begin{cases}1-\frac{1}{\lfloor t\rfloor} & \text{for } t\geq 2,\\ 0 & \text{for } t<2.\end{cases}\end{aligned}$$

At first glance this cdf appears very similar to that of a Pareto$(1,1)$ but unlike a true Pareto distribution it is a discrete distribution, hence calling it a discrete Pareto distribution. This distribution shares a number of properties with the usual Pareto distribution including heavy tailed-ness which was seen when the expectation of $T(2)$ was shown to be infinite earlier in this section.

5.6 Generation of Records

Say we wished to generate a sequence of records of a certain length using the naive approach of simply generating a sequence $\{X_1, X_2, \dots, X_n\}$ and taking any naturally occurring records. How many random samples would we expect to have to generate? It turns out that the expected number of records in a sequence of n many i.i.d. random samples is the n^{th} harmonic number or approximately $\log n$ (Weisstein 2026).

Proof

$$\begin{aligned}\mathbb{E}[N(n)] &= \sum_{j=1}^{n} P_j \\ &= \sum_{j=1}^{n} \frac{1}{j} \\ &= H_n \\ &\approx \log n + \gamma, \quad \text{where } \gamma \text{ is the Euler-Mascheroni constant.}\end{aligned}$$

This makes for a very inefficient method of simulation with an expected number of 4.6 records after 100 samples and still only 6.9 records after 1000 samples.

More efficient methods of record generation utilise the fact that sequences of records form a Markov process of discrete time and continuous state space with the following distribution,

$$\mathbb{P}(X(n+1) \leq y \,|\, X(n) = x) = \frac{F(y) - F(x)}{1 - F(x)} \quad \text{for } x < y.$$

6 A Sequence of Non-i.i.d. Random Variables: the Linear Drift Model

The linear drift model is a sequence of random variables $\{X_1, X_2, \ldots, X_k\}$ such that

$$X_k = Y_k + ck, \quad \text{where } Y_k \text{ is an i.i.d. sequence, } c \in \mathbb{R}.$$

It is just one example of a sequence of non-i.i.d. random variables and will be considered throughout this section. It incorporates an overall linear trend to the sequence making it far more useful in modelling real world records unlike the simplistic i.i.d. model. This does complicate its analysis although the assumption of independence for the initial sequence has not been relaxed meaning many results can still be found. A further simplification will also be made in that only an underlying exponential distribution will be considered, where $Y_k \sim \text{Exp}(1)$.

First the probability of the n^{th} observation being a record is derived. Then the distribution of the time of the second record is derived which was significantly more difficult than in the i.i.d. case where the sequence of records was independent. This is no longer the case in the linear drift model because the order of observation now matters, it is time dependent. As such, no closed form expression was found so instead numerical simulations and calculations are utilised with all all R code for these numerical approximations contained in Appendix 8.1.

6.1 Probability of nth Observation Being a Record

To find the probability of the n^{th} observation being a record, we can simply integrate over all possible values of X_n or really its underlying Y_n. This results in a formula requiring knowledge of Y_k's pdf and cdf (Wergen 2013).

Proof

$$\begin{aligned}
P_n &= \mathbb{P}\left[X_n > \max\left\{X_1, X_2, \ldots, X_{n-1}\right\}\right] \\
&= \mathbb{P}\left[Y_n + nc > \max\left\{Y_1 + c, Y_2 + 2c, \ldots, Y_{n-1} + (n-1)c\right\}\right] \\
&= \mathbb{P}\left[Y_1 + c < Y_n + nc\right] \mathbb{P}\left[Y_2 + 2c < Y_n + nc\right] \ldots \mathbb{P}\left[Y_{n-1} + (n-1)c < Y_n + nc\right] \\
&= \int_{\mathbb{R}} f(y)\, \mathbb{P}\left[Y_1 < y + (n-1)c\right] \mathbb{P}\left[Y_2 < y + (n-2)c\right] \ldots \mathbb{P}\left[Y_{n-1} < y + c\right] dy \\
&= \int_{\mathbb{R}} f(y) \prod_{k=1}^{n-1} F(y + ck)\, dy.
\end{aligned}$$

When $Y_k \sim \text{Exp}(1)$ as it is throughout this section, we obtain

$$P_n = \int_{\max\{0,-c\}}^{\infty} e^{-y} \prod_{k=1}^{n-1} \left(1 - e^{-y-ck}\right) dy,$$

with the lower limit accounting for the exponential distribution positive support. This can be further "simplified" for $c \geq 0$ to

$$P_n = \int_0^1 \frac{(y, e^{-c})_n}{1-y}\, dy,$$

where $(a, q)_n$ is the q-Pochhammer symbol with $(a, q)_n := \prod_{k=0}^{n} \left(1 - aq^k\right)$ (Abramowitz and Stegun 1970). However, this is done to derive the asymptotic record rate where limiting results of the q-Pochhammer symbol can be used. Note that this asymptotically constant record rate exists whenever Y_k has a finite first moment like the exponential distribution (Ballerini and Resnick 1985).

For investigations in the non-asymptotic case a recursive formula for P_n would be extremely useful. However, this has not yet been possible to derive, instead only the following semi-recursive formula (still slightly more efficient for computation) was found.

Derivation

Firstly, we trivially have that $P_1 = 1$. We can calculate P_2 as follows using the integral product formula from the start of this section and direct integration.

$$\begin{aligned}
P_2 &= \int_{\max\{0,-c\}}^{\infty} e^{-y} \left(1 - e^{-y-c}\right) dy \\
&= \int_{\max\{0,-c\}}^{\infty} e^{-y} - e^{-2y-c}\, dy \\
&= \left[-e^{-y} + \frac{1}{2} e^{-2y-c}\right]_{\max\{0,-c\}}^{\infty} \\
&= e^{\max\{0,-c\}} - \frac{1}{2} e^{-2\max\{0,-c\}-c} \\
&= \begin{cases} 1 - \frac{1}{2} e^{-c} & c \geq 0, \\ \frac{1}{2} e^{c} & c < 0. \end{cases}
\end{aligned}$$

Now a semi-recursive form for P_n can be found for $n \geq 3$. Ideally a pure recursive form would be found

making later calculation of $T(2)$ far more efficient but this is sufficient for now.

$$\begin{aligned}
P_n &= \int_{\max\{0,-c\}}^{\infty} e^{-y} \prod_{k=1}^{n-1} \left(1 - e^{-y-ck}\right) \, dy \\
&= \int_{\max\{0,-c\}}^{\infty} \left(1 - e^{-y-(n-1)c}\right) e^{-y} \prod_{k=1}^{n-2} \left(1 - e^{-y-ck}\right) \, dy \\
&= \int_{\max\{0,-c\}}^{\infty} e^{-y} \prod_{k=1}^{n-2} \left(1 - e^{-y-ck}\right) \, dy - e^{-(n-1)c} \int_{\max\{0,-c\}}^{\infty} e^{-2y} \prod_{k=1}^{n-2} \left(1 - e^{-y-ck}\right) \, dy \\
&= P_{n-1} - e^{-(n-1)c} \int_{\max\{0,-c\}}^{\infty} e^{-2y} \prod_{k=1}^{n-2} \left(1 - e^{-y-ck}\right) \, dy.
\end{aligned}$$

6.2 Numerically Calculating the Time of the Second Record

To derive the pmf of the time of the second record, we must be very careful about what is and is not independent. It is only the sequence of X's and the underlying Y's that are independent, unlike in the i.i.d. case where the sequence of record values also had independence. That is what allowed us to simply multiply the individual probabilities of each observation being a record to obtain the distribution of the time of the second record.

For the linear drift model, the probability instead needs to be calculated explicitly with another integral formula using the underlying distribution's pdf and cdf.

Derivation

$$\begin{aligned}
\mathbb{P}\left(T(2) = t\right) &= \mathbb{P}\left(X_1 \geq X_2, X_3, \ldots, X_{t-1} \cap X_1 < X_t\right) \\
&= \mathbb{P}\left(Y_1 + c \geq Y_2 + 2c, Y_3 + 3c, \ldots, Y_{t-1} + (t-1)c \cap Y_1 + c < Y_t + tc\right) \\
&= \int_{\mathbb{R}} f(y) \mathbb{P}\left(Y_t > y - (t-1)c\right) \prod_{i=2}^{t-1} \mathbb{P}\left(Y_i \leq y - (i-1)c\right) \, dy \\
&= \int_{\mathbb{R}} f(y) \left\{1 - F(y-(t-1)c)\right\} \prod_{i=2}^{t-1} F\left(y - (i-1)c\right) \, dy.
\end{aligned}$$

It is best to leave the integral in this form and not substitute in the explicit form of $f(y)$ and $F(y)$ for the exponential distribution. This is because this will impact on the bounds of integration for any ranges of y that result in the pdf or cdf being 0. Leaving it in this form and using proper pdf and cdf functions in the R code which return 0 for any negative input, for the exponential distribution, makes it far simpler.

The above pmf was implemented in R code which is included in Appendix 8.1. Evaluating the distribution of $T(2)$ is not excessively computationally intensive and does not need to be done for infinite t values. This is because the pmf is a decreasing function, the later observations will always have a lower chance of being the second record simply because they come later in the sequence. Eventually, up to the limits of numerical precision, for large enough t, $P(T(2) = t) \approx 0$ and computation of the distribution can stop.

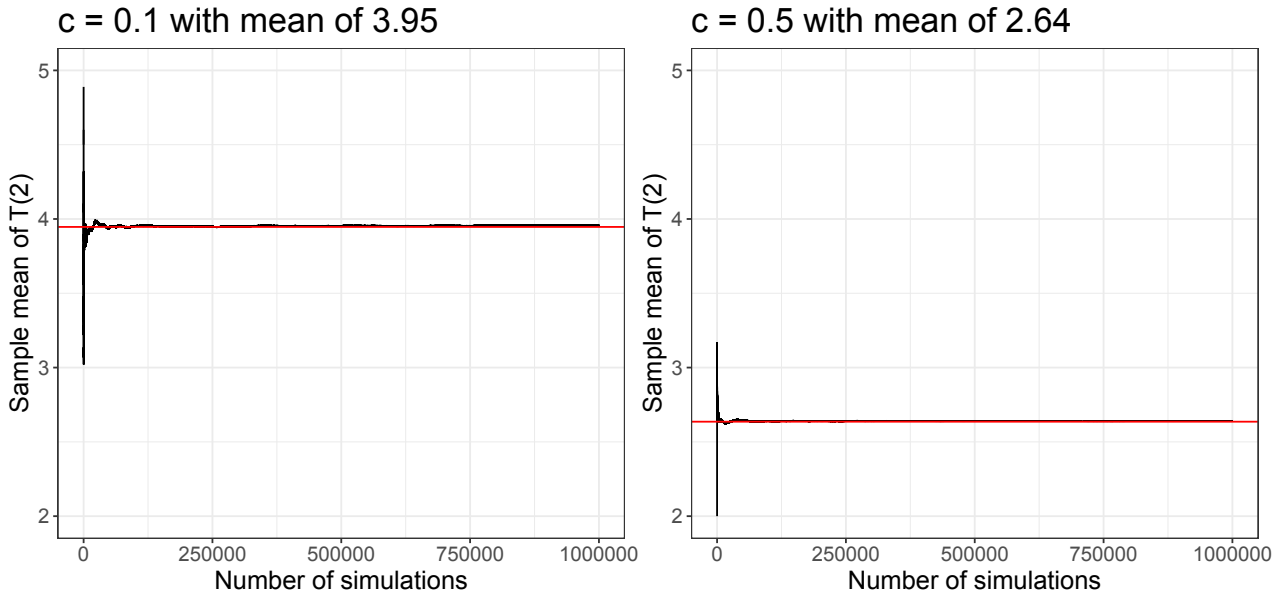

Figure 2: Simulation study of the time of the second record under the linear drift model.

6.3 Simulation Study

Figure 2 shows the results from the simulation study where the underlying Y_k distribution is an exponential distribution with mean 1. The R code used for this simulation is contained in Appendix 8.2. An outline of the simulation's procedure is as follows,

1. Generate the random sequence $\{X_1, X_2, \ldots, X_n\}$ until a second record is observed, i.e. when $X_n > X_1$.
2. The observed value of n is the observed value of $T(2)$ for that single simulation.
3. Repeat steps 1 and 2 and calculate the sample mean of $T(2)$ as the number of simulations increase.

This is a naive approach of generation that is highly inefficient particularly because of the heavy tailed nature of $T(2)$ meaning some simulations of $\{X_1, X_2, \ldots, X_n\}$ are hundreds of thousands of observations long until eventually the second record is observed. A potentially more efficient method would have been to utilise the numerically derived distribution of $T(2)$ and sample directly from that using the inverse cdf method but this would require numerical calculation of the inverse cdf.

When c is positive and an increasing trend is introduced to the sequence of random variables, the upper records become more frequent. This reduces the expectation of $T(2)$, on average less observations are required until the second record is observed, and makes it finite rather than infinite as in the i.i.d. case, when $c = 0$. This can be seen in Figure 2 where for positive values of c we see the sample mean rapidly converges to some fixed value as expected under the law of large numbers which holds because we now have that $T(2)$ is finite.

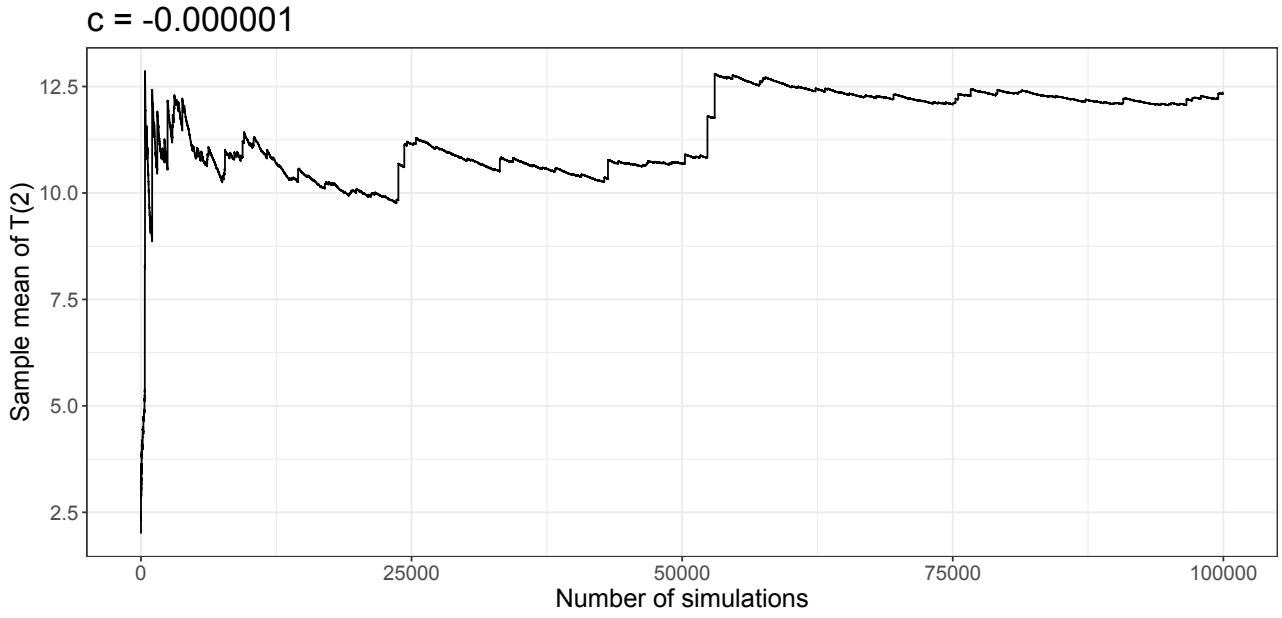

Figure 3: When c is negative the expectation remains infinite.

The simulations agree with the theoretically derived and numerically evaluated mean of $T(2)$ which is the horizontal red line in each plot.

On the other hand when, c is negative and a decreasing trend is introduced to the sequence of random variables, the upper records become less frequent increasing the expectation of $T(2)$. It is also possible that if the first observation X_1 is sufficiently high and c has great enough magnitude there will simply never be a second record. As such the only simulations, as shown in Figure 3, that could be produced had a value of c very close to 0 with the expectation of $T(2)$ still being infinite. Hence, the law of large numbers does not apply and there is no convergence with the plot looking the same as how it would in the i.i.d. case.

7 Discussion and Conclusion

A record is an observation greater than all that precede it and appear in many areas including sports, climate, finance and insurance. The winning times of the annual Sydney to Hobart yacht race was used to introduce the concept and statistical notation of records before well known results for the i.i.d. case were explored.

In particular, the time of the second record, the first non trivial record, was investigated and found to follow a discrete Pareto distribution. This distribution shares the heavy tailed property of the regular Pareto distribution so has an infinite expectation.

Then the linear drift model was introduced using an underlying exponential distribution. The probability of a record occurring at each time n was derived and simplified to a semi-recursive form for more efficient evaluation. Again the time of the second record was investigated using numerical methods to evaluate it and

compare to simulated results. It was found that a positive linear trend reduces the expectation from infinite in the i.i.d. case to a finite value that can be approximated by numerical evaluation of an integral involving the pdf and cdf of the underlying distribution.

Future work could further investigate the distribution of $T(2)$ under the same setup of an exponential distribution in the linear drift model and perhaps find a closed form expression for it. Additionally, other non-i.i.d. models could be investigated to determine whether imposing a trend weaker than linearity is still enough to reduce the expected time of the second record to a finite value.

8 Appendices

8.1 R Code for Numerically Calculating $T(2)$

As done in section 6.2.

```r
c <- 0.1
pmf_T2 <- function(t, c) {
  # start indexing from 0, X0, X1, X2, ...
  if (t < 0 || t %% 1 != 0) stop("t must be a nonnegative integer (time index).")
  # second record cannot occur at time 0
  if (t == 0) return(0)
  if (c<0) stop("c must be positive for this code")

  integrand <- function(y) {
    # y may be a vector
    fy <- dexp(y)                              # f_{Y0}(y)
    # product_{i=1}^{t-1} P(Y_i <= y - i*c)
    if (t == 1) {
      prod_cdfs <- rep(1, length(y))       # empty product = 1
    } else {
      I <- 1:(t-1)
      M <- outer(y, I, function(yy, ii) pexp(yy - ii*c))
      M[M <= 0] <- .Machine$double.xmin    # avoid exact zeros
      log_prod <- rowSums(log(M))
      prod_cdfs <- exp(log_prod)
    }
    tail_prob <- 1 - pexp(y - t*c)         # P(Y_t > y - t*c)
    fy * prod_cdfs * tail_prob
```

```
  }

  integrate(integrand, lower = 0, upper = Inf)$value
}

simulateT2data <- function(c, n = 1000000) {
  # JUST GIVES SAMPLE OF t's
  allT2 <- numeric(n) # list of {t1,t2,...,tn} where each ti is an observation of T(2)
  X1_all <- rexp(n) + c
  for (i in 1:n) {
    T2 <- 2
    X1 <- X1_all[i]
    k <- 2
    Xk <- rexp(1) + k*c
    while (Xk < X1) {
      k <- k + 1
      T2 <- T2 + 1
      Xk <- rexp(1) + k*c
    }
    allT2[i] <- T2
  }
  return(allT2)
}

# compare simulated pmf with theoretical
tibble(t=simulateT2data(c)) %>%
  count(t) %>%
  mutate(proportion_sample = n/sum(n),
         sample_count = n) %>%
  full_join(mutate(tibble(
    t = 2:100 - 1,
    probs_theoretical = map_dbl(t, ~ pmf_T2(.x, c))
  ), t=t+1)) %>%
  filter(probs_theoretical > 0.005) %>%
  ggplot(aes(x=t)) +
  geom_point(aes(y=probs_theoretical, shape="Theoretical", color="Theoretical"), size=3) +
```

```
  geom_point(aes(y=proportion_sample, shape="Simulated", color="Simulated"), size=3) +
  scale_shape_manual(values = c("Theoretical" = 1, "Simulated" = 2)) +
  scale_color_manual(values = c("Theoretical" = "red", "Simulated" = "blue")) +
  labs(shape = "Dataset", color = "Dataset", y = "P(T(2)=t))") +
  theme_bw()
```

8.2 R Code for Simulation Study

As done in Section 6.3.

```
# the original simulation now with predicted mean
mean_T2_0.1 <- sum(pull(tibble(t=1:1000, prob = map_dbl(t, ~ pmf_T2(.x, c))), prob)*(2:1001))
ldmplot <- function(c, mean_T2, n = 1000000) {
  # RETURNS THE PLOT
  allT2 <- numeric(n) # list of {t1,t2,...,tn} where each ti is an observation of T(2)

  for (i in 1:n) {
    T2 <- 2
    X1 <- rexp(1,1) + c
    k <- 2
    Xk <- rexp(1, 1) + k*c

    while (Xk < X1) {
      k <- k + 1
      T2 <- T2 + 1
      Xk <- rexp(1, 1) + k*c
    }

    allT2[i] <- T2
  }

  plot <- tibble(
    obs = 1:n,
    T2 = allT2
  ) %>%
    mutate(running_mean = cummean(T2)) %>%
```

```
    ggplot(aes(x = obs, y = running_mean)) +
    geom_line() +
    geom_hline(yintercept = mean_T2, color="red") +
    labs(
      subtitle = paste0("c = ",
                        c,
                        ". Red line is the numerically calculated mean of ",
                        round(mean_T2, 2)),
      x = "Number of simulations",
      y = "Sample mean of T(2)"
    ) +
    theme_bw() +
    theme(axis.title.x = element_text(size=15),
          axis.title.y = element_text(size=15),
          axis.text = element_text(size=12),
          plot.title = element_text(size=20))

  return(plot)
}
ldmplot(c, mean_T2_0.1)
```

9 References

Abramowitz, Milton and Irene A. Stegun (1970). *Handbook of Mathematical Functions*. New York: Dover Publications Inc.

Arnold, Barry C., N. Balakrishnan, and H. N. Nagraja (1998). *Records*. Wiley-Interscience.

Ballerini, R. and S. Resnick (1985). In: *Journal of Applied Probability* 22, pp. 487–502.

Stepanov, Alexei (2022). "On the Mathematical Theory of Records". In: *Communications in Mathematics*.

Weisstein, Eric W. (2026). *Harmonic Number*. https://mathworld.wolfram.com/HarmonicNumber.html. From *MathWorld–A Wolfram Web Resource*. (Visited on 02/25/2026).

Wergen, Gregory (2013). "Records in Stochastic Processes - Theory and Applications". In: *J. Phys. A: Math. Th.* 46.223001. DOI: https://doi.org/10.48550/arXiv.1211.6005. URL: https://arxiv.org/abs/1211.6005.