Record Statistics for non-i.i.d. Random Variables in the Presence of Rounding-off Effects

Adam Watson

Department of Mathematics
Kings College London

MSc Complex Systems Modelling

Abstract

Previous studies of record events in a time series of random variables is surprisingly limited to those which are independent and identically distributed. Few have gone beyond this setting. This research project combines two recent studies into the statistics of records for random variables that have been rounded to a specified accuracy and those that have been drawn from the same underlying probability distribution with additive linear drift.

For the former, a formula has been added, representing the probability that two adjacent random variables in a given time series are both records. Their relevant correlations have also been touched upon and interestingly, rounding down causes i.i.d random variables to lose their independence. Additionally, the project extends previous work to a non-i.i.d framework.

For the latter, I have analysed the consequences of applying both drift and rounding to a time series. Ultimately, this can be characterised by which universality class of extreme value statistics the probability distribution belongs to. The additive term increases the record rate whilst rounding has the opposite effect. The interrelation between them shows interesting results and one finds the additive drift contribution is the strongest.

Summary

1. Literary Review

Rare events are ones that occur with very low frequency [20] and have had considerable impact across the world over the last century; financial market crashes and global warming changes are examples of this. It therefore comes as a surprise that the field is not saturated, with many more advancements to be made. Mathematically, record events are defined to be entries in a time series that are higher or lower than all previous others [9] and are apparent in all man-made phenomena; climate records, sports and biology, to name a few [11]. The statistics of records for i.i.d random variables has been thoroughly researched and is well understood [9], but for non i.i.d random variables, research is relatively new. One reason for this is that once we assume strong correlations between random variables, we lose tools of classical statistics. Extreme Value Statistics, which has seen a surge in popularity over the last decade, highlights the limitations of an i.i.d assumption. A noticeably larger occurrence of records cannot be explained by this, so new models must be developed. [4]

Therefore, it is important to ask what happens when the underlying time series is correlated [9]. Initial attempts to explain the frequent occurrence of new Olympic records were conducted by Yang in [7]. To consider the record rate at point in time N, a designated number of i.i.d RVs become available at that time step to represent a growing athletic population. The probability of a new record then depends on the ratio between the number of new random variables and ones previously drawn. More recently, in [6] Krug considers the record rate and mean record number where each random variable in an increasing time series is drawn from the same underlying probability distribution but with a broadening shape or increasing variance. This will be referred to later in the project.

Furthermore, applications of the linear drift model [3] in [11], random walks and Lévy flights in [8] are also recent attempts to address this. In the linear drift model, each entry in the time series is a non - i.i.d random variable broken down into component i.i.d part and a constant linear drift. This was applied to American and European weather data in [10] where the record rate change across a time series is proportional to the drift to standard deviation of the underlying distribution ratio. On the other hand, in the random walk model each entry is updated by a jump associated to a symmetric probability distribution. Unlike applications of the LDM model, it was found that record statistics are the same regardless of the underlying jump distribution chosen.

On a different note, research has been conducted on the effect of the possibility of measurement errors [14] and rounding-offs [12]. In all cases, records tend to be discretised due to technological limitations which force us to round off observations, regardless of whether the underlying probability density function is continuous or not. This produces or breaks ties that would not be there originally. Currently, i.i.d random variable cases have only been considered. The effect of ties has a large impact on the number of recorded records in the case of an underlying discrete distribution, whilst record values drawn from a continuous underlying distribution has not been studied extensively [12].

The first part of the project is dedicated to providing a detailed overview of the two papers I will look to overlap: [11] and [12] which entails reproducing simulations and including proofs of derived expressions; as sometimes these can lack clarity across the papers. Before this, I will introduce some classical results from the field which will be useful across the whole project.

2. Introduction

2.1 Classical Extreme Value Statistics

As mentioned previously, the study of extreme value statistics, given random variables are independent and identically distributed, has received considerable attention. I will briefly summarise a few elementary results from the literature. Across the course of the project, a time series of random variables drawn from a certain distribution is considered. Mathematically, one takes a collection of i.i.d random variables $\{X_i\}$ with the same underlying probability density function $f(x)$ and corresponding cumulative density function: $F(x) = \int^x f(y)$. As we are considering records, one quantity of interest is $X_{max} = \max\{X_i : i \leq N\}$, the entry in the N-length time series which is largest. How could one describe the probability distribution of X_{max}?

Suppose $Q_N(x)$ is the cumulative distribution function of X_{max}. Mathematically, this can be expressed as the Prob$[\max\{X_i : i \leq N\} \leq x]$. Our cdf then factorises into N copies of the parent cdf $P(x)$ defined earlier, as each entry in the time series is independent and identically distributed:

$$\begin{aligned} Q_N(x) &= \int^x \int^x \cdots \int^x f(x_1)f(x_2)\cdots f(x_N)dx_1dx_2\cdots dx_N \\ &= \left[\int^x f(y)\right]^N \end{aligned}$$

We can also generalise its asymptotic behaviour by taking $N \to \infty$, $x \to \infty$ and identifying a non-trivial limiting distribution $F(z)$ [15]. This involves finding *centering* and *scaling* constants $a_N \in \mathbb{R}$ and $b_N > 0$ such that:

$$\lim_{N\to\infty} Q_N(a_N + b_N z) = F(z)$$

and this leads us to the **Fisher-Tippet-Gnedenko** Theorem [1]-[2] which states that the limiting distribution $F(z)$ can be one of only three types: Gumbel, Fréchet or Weibull [15]. These three extreme value statistics classes will a constant theme throughout the project. A quick example to clarify this is to take our parent pdf to be exponential $p(x) = \lambda e^{-\lambda x}$. The centering and scaling constants are easily found as followed.

$$\begin{aligned} Q_N(x) = P(x)^N = \left[1 - e^{-\lambda x}\right]^N &= e^{Nlog[1-e^{-\lambda x}]} \\ \approx e^{-Ne^{-\lambda x}} &= e^{-e^{-(\lambda x - logN)}} \end{aligned}$$

which is of the form of our Gumbel pdf $e^{-e^{-z}}$ where $z = \lambda x - logN$ giving $a_N = \frac{logN}{\lambda}$, $b_N = \frac{1}{\lambda}$.

Furthermore, what is the probability of the N'th entry in our time series being an **upper record**? Mathematically, we require $X_N > \max\{X_i : i < N\}$ and therefore:

$$P_N = \int dx\, f(x)\, [F(x)]^{N-1}$$

where the upper and lower limits of the integal represent the support of the corresponding pdf. Inside the integral we have the probability that the Nth RV takes a value x combined with the probability that all previous $N-1$

random variables take a value less than x. Integrating over the whole support considers all possible record values x [13]. A well established result for i.i.d continuous random variables is that $P_N = \frac{1}{N}$, irrespective of the distribution chosen.

Proof. By letting $u = F(x)$, $du = f(x)dx$, $P_N = \int_0^1 u^{N-1}$ and thus equals $\frac{1}{N}$. □

Another quantity of interest within the field is the **record number** R_N: the number of upper records up to and including the N'th entry in the time series. The mean record number $\langle R_N \rangle$ is the sum over the record rates: $\sum_{k=1}^{N} P_k$ which behaves logarithmically for large N. More precisely: $R_N \approx \ln n + \gamma$ where γ is Euler's constant.

The next two chapters will contain model results and assumptions from [11] and [12].

3. Correlations between Record Events in Sequences of Random Variables with a Linear Trend

In the linear drift (LDM) model, each entry in the time series is a non - i.i.d random variable broken down into component i.i.d part and a constant linear drift. One takes a collection of random variables: $\{Y_l\}$ with i.i.d part $\{X_l\}$ and corresponding probability density function $f(x)$, added to a constant linear drift $c > 0$ [11]:

$$Y_l = X_l + cl \tag{3.1}$$

Therefore we can see that the pdf of our random variable Y_l has the same shape as it's component i.i.d part but is shifted by a constant c each time a new random variable is drawn.[11] . Entries in the time series are no longer i.i.d so one can introduce a correlation function $l_{N,N-1}(c)$ which is the joint probability $p_{N,N-1}$ of two record random variables Y_N and Y_{N-1} occurring consecutively normalised by the record marginals: p_N and p_{N-1}:

$$l_{N,N-1}(c) = \frac{p_{N,N-1}}{p_N p_{N-1}} \tag{3.2}$$

3.1 Model Definition

Now consider a collection of independent random variables $\{X_l\}$ with corresponding density function $f_l(x)$ (not necessarily the same across all R.V). We can define the probability that the maximum of N drawn random variables is less than some value x as the product of the analogous cdf's: $\prod_{l=1}^{N} F_l(x)$. This will then allow us to define our record probability p_N, the probability that the N'th RV is a record as:

$$p_N = \int dx f_N(x) \prod_{l=1}^{N-1} F_l(x) \tag{3.3}$$

Again, the product within the integral guarantees that the first $N-1$ random variables are less than a record value x whilst $f_N(x)$ guarantees the Nth random variable is a record. Integrating over the upper and lower bounds of the support of the analogous pdf exhausts all possible record values that can be obtained, giving a probability. Similarly the joint probability that both the N'th and $N-1$'th drawn RV are records can be defined as followed:

$$p_{N,N-1} = \int dx_N f_N(x_N) \int^{x_N} dx_{N-1} f_{N-1}(x_{N-1}) \prod_{l=1}^{N-2} F_l(x_{N-1}) \tag{3.4}$$

By the same logic, the product over the first $N-2$ random variables gives the probability that they are less than some value x_{N-1}. $f_{N-1}(x_{N-1})$ gives the probability of the $N-1$th RV taking a record value x_{N-1}. Crucially, the inner integral limits take into account all record values up to a certain value x_N. Finally, $f_N(x_N)$ gives the probability that the Nth RV obtains a record value x_N and therefore integrating over the whole support generates a joint probability.

The above formulas are a general case for N independent but not necessarily identically distributed random variables. Using **Eq. 3.1** the aim is to remove dependence on the index l by noting that $f_l(y) = f(x + cl)$. Plugging this into **Eq. 3.3**. and **Eq. 3.4**. we obtain the desired components of the correlation function:

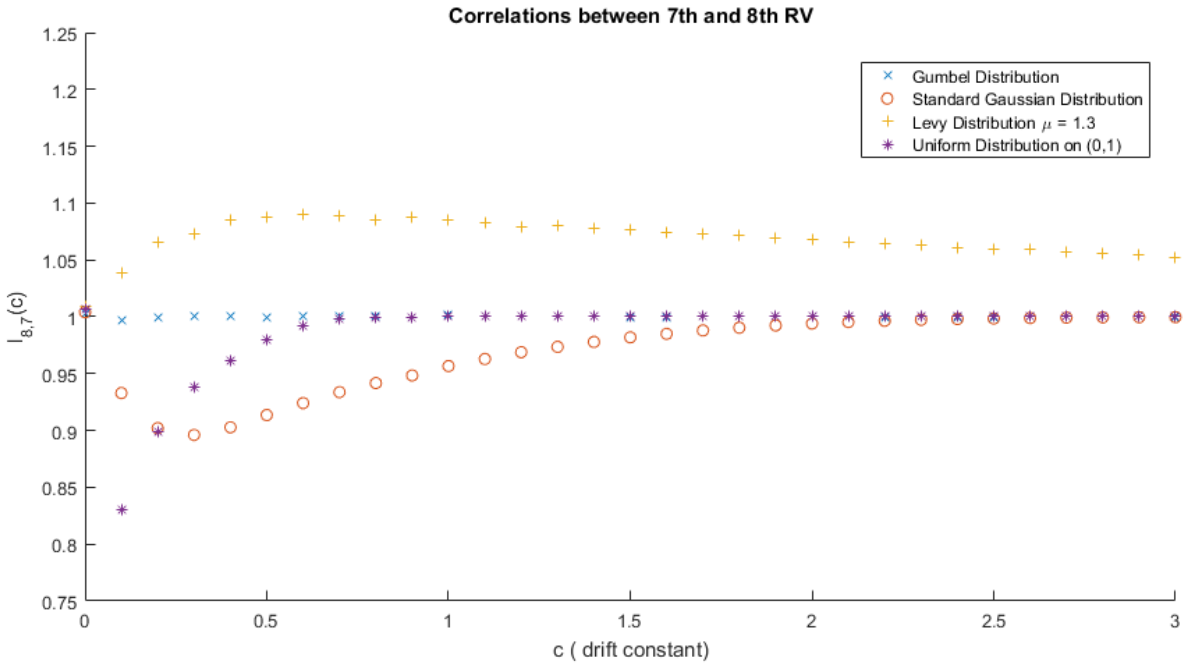


Figure 3.1: Correlations between the record events occuring in the 7th and 8th time step for different distributions of the underlying i.i.d part [11] using **Eq. 3.2**. Monte Carlo simulations are averaged over 10^6 realisations.

$$
\begin{aligned}
p_N(c) &= \int dy f_N(y) \prod_{l=1}^{N-1} F_l(y) = \int dx f(x) \prod_{l=1}^{N-1} F(x+cl) \\
p_{N,N-1}(c) &= \int dy f(y) \int^{y+c} dx f(x) \prod_{l=1}^{N-2} F(x+cl)
\end{aligned}
\tag{3.5}
$$

3.2 Simulation Results

Wergen $et\ al$ simulated record event correlations between the 7th and 8th random variables for four different underlying probability distributions: Gumbel, Lévy stable with $\alpha = 1.3$, uniform on $(0, 1)$ and the standard Gaussian. For all distributions, as c increases, correlations between record events tend to 1 (stochastic independence) as each random variable drawn is a record. Therefore, the probability of two consecutive records is also one giving an equal joint to marginals probability ratio. Where $c << 1$, in some cases (Lévy stable distribution) record events attract each other i.e $l_{N,N-1}(c) > 1$, whilst in others (Gaussian, uniform) record events repel each other i.e $l_{N,N-1}(c) < 1$. For the Gumbel case we see stochastic independence for all values of c, and the proof of this is contained in **Appendix A.1** [11].

3.3 Small Drift Expansion

Simulation results are only interesting within the small c region as for a large value of c every value in a time series will be bigger than all previous others. For this reason, one can apply a first order series expansion to the probability the N'th random variable is a record in the small c limit:

$$
\begin{aligned}
p_N(c) &\approx \int dx f(x) F(x)^{N-1} + c\frac{N(N-1)}{2} \int dx f(x)^2 F(x)^{N-2} \\
&= \frac{1}{N} + c\frac{N(N-1)}{2} I(N-2)
\end{aligned}
\tag{3.6}
$$

where $I(N) = \int dx f(x)^2 F(x)^N$. Noticeably, this expression is made of up the record rate in the case of an **i.i.d** framework and an additive drift term. Clearly the behaviour of $p_N(c)$ depends which EVS class the distribution is a member of [13].

Proof. A first order Taylor expansion says: $F(x + cl) \approx F(x) + clF'(x) = F(x) + clf(x)$. Therefore $p_N(c) \approx \int dx f(x) \prod_{l=1}^{N-1} [F(x) + clf(x)]$. It is also worth noting that, after ignoring any terms that are of $O(c^2)$, $\prod_{l=1}^{N-1} [F(x) + clf(x)] = F(x)^{N-1} + c\frac{N(N-1)}{2} f(x) F(x)^{N-2}$. This pattern can be identified after expanding the series for $N = 2, 3, 4 \cdots$ and generalising for all N. Combining this expansion with $\int dx f(x)$ gives the first line of **Eq. 3.6**.

To show that the first term of the RHS equals $\frac{1}{N}$ we use the substitution $u = F(x)$, $du = f(x)dx$ and limits $F(\infty) = 1$ and $F(-\infty) = 0$ [9]:

$$
p_N(c = 0) = \int_0^1 du\, u^{N-1} = \frac{1}{N}
$$

This covers all of the components in the final line of **Eq. 3.6**. □

Moving to the **joint record rate** in **Eq. 3.5** it was found in [11].

$$
p_{N,N-1}(c) = \frac{1}{N(N-1)} - c\frac{(N-1)(N-2)-2}{2} I(N-2) + c\frac{(N-1)(N-2)}{2} I(N-3) \tag{3.7}
$$

As the formal proof was left out of the paper and many of the methods used are an extension of the derivation above, I have provided a detailed explanation in **Appendix A3**.

Importantly, Wergen $et\ al$ derive an expansion for the normalised correlation $l_{N,N-1}(c)$ by writing [11]:

$$
l_{N,N-1}(c) = 1 + cJ(N) + O(c^2) \tag{3.8}
$$

where:

$$
\begin{aligned}
J(N) = N(N-1)I(N-2) &+ \frac{N(N-1)^2(N-2)}{2} [I(N-3) - I(N-2)] \\
&- \frac{N^2(N-1)}{2} I(N-2) - \frac{(N-1)^2(N-2)}{2} I(N-3)
\end{aligned}
$$

Proof. Rewriting $p_N(c)p_{N-1}(c)$ and $p_{N,N-1}(c)$ with common denominators:

$$
\begin{aligned}
p_N(c)p_{N-1}(c) &\approx \frac{1 + c\frac{N^2}{2}(N-1)I(N-2) + \frac{c}{2}(N-1)^2(N-2)I(N-3)}{N(N-1)} \\
p_{N,N-1}(c) &\approx \frac{1 + cN(N-1)I(N-2) + \frac{c}{2}N(N-1)^2(N-2)(I(N-3) - I(N-2))}{N(N-1)}
\end{aligned}
\tag{3.9}
$$

Dividing the joint probability by the combination of marginals and using the series approximation: $\frac{1+c(A+B)}{1+c(D+E)} \approx 1 + c[A + B - D - E]$ gives us the above expression for $J(N)$ where A, B, D and E are the component terms in the marginals of **Eq. 3.9**. □

The core part of the paper is based on deriving approximations of the correlation function for the three universal EVS classes. This stems from patterns in simulations. Distributions that belong the to same EVS class tend to exhibit similar small drift behaviour .

3.3.1 Weibull Class

One can begin with the simple example of the Weibull class: the uniform distribution on the interval $[-a, a]$. For $N >> 1$:

$$l_{N,N-1}(c) \approx 1 - \frac{c}{4a}N^2 \tag{3.10}$$

Proof. Using $f(x) = \frac{1}{2a}$ and $F(x) = \frac{x+a}{2a}$, $I(N) = \int_{-a}^{a} dx(\frac{1}{2a})^2(\frac{x+a}{2a})^N = \frac{1}{N+1}\frac{1}{2a}$ follows. Therefore, $I(N-3) - I(N-2) = -\frac{1}{2a}(\frac{1}{(N-1)(N-2)}$ and plugging into our expression for $J(N)$ gives: $J(N) = \frac{N}{2a} + \frac{1}{4a}(-N^2 + N + 1)$. In the large limit order 2 terms dominate giving $J(N) \approx -\frac{N^2}{4a}$. Plugging into the first line of **Eq. 3.8** gives us our answer. □

This expression gives the correct sign of the correlation function as it departs from the i.i.d case (refer back to **Fig. 3.1**) but does not accurately represent their simulations for large N. The expansion in **Eq. 3.6** is only 'valid if $f(x)$ is slowing varying between x and $x + c$' [9] which is not the case for the uniform distribution on a very small range $[-a, a]$. This can be seen in **Fig. C.1** in the Appendix.

Moving to the more general Weibull example with pdf: $f(x) = \xi(1-x)^{\xi-1}$, $\xi > 0$ and $x \in (0, 1)$. It is shown in [9] that:

$$I(N) = \xi\frac{\Gamma(2 - \frac{1}{\xi})\Gamma(N-1)}{\Gamma(N + 1 - \frac{1}{\xi})} \tag{3.11}$$

Proof. For the distribution $f(x) = \xi(1-x)^{\xi-1}$: $I(N) = \int dx \left[\xi(1-x)^{\xi-1}\right]^2 \left((1-x)^{\xi} + 1\right)^{N-2}$, the substitution $1 - x = z^{\frac{1}{xi}}$ gives the following integral in the form of a Beta Distribution.

$$\begin{aligned} I(N) &= \xi \int dz z^{1-\frac{1}{\xi}}(1-z)^{N-2} \\ &= \xi N \frac{\Gamma(2 - \frac{1}{\xi})\Gamma(N-1)}{\Gamma(N + 1 - \frac{1}{\xi})} \end{aligned} \tag{3.12}$$

□

To find an expression for $J(N)$ one notes that in the large N limit: $I(N-2) \approx I(N-3)$ and the second term in the expression **Eq. 3.8** can be ignored. In Mathematica using the in built FullSimplify function allows one to group terms like so:

$$J(N) \approx (2 - 7N + 7N^2 - 2N^3)\xi\frac{\Gamma(2 - \frac{1}{\xi})\Gamma(N-1)}{2\Gamma(N + 1 - \frac{1}{\xi})} \tag{3.13}$$

N^3 is clearly the dominant term here in the asymptotic N limit. Additionally, using Stirling's approximation for the Beta Function: $B(x, y) \sim \Gamma(x)y^{-x}$ with large y and fixed x [13] gives the following expression:

$$l_{N,N-1}(c) \approx 1 - \frac{c}{2}\Gamma(2 - \frac{1}{\xi})N^{1+\frac{1}{\xi}} \tag{3.14}$$

As expected the correlation function departs rapidly from the i.i.d case with increasing L just like the uniform approximation.

3.3.2 Gumbel Class

An important member of the Gumbel class is the exponential distribution with mean a and corresponding pdf $f(x) = a^{-1}e^{\frac{x}{a}}$, $x \geq 0$. Thus: our recurring integral $I(N)$ will take the form:

$$I(N) = \int_0^\infty \frac{1}{a^2} e^{-\frac{2x}{a}}(1 - e^{-\frac{x}{a}})^N = \frac{1}{a}\frac{1}{(N+2)(N+1)}$$

Proof. By making the following substitutions $u = \frac{x}{a}$, $dx = adu$ and $y = e^{-u}$, $dy = -e^{-u}du$ leaves the definite integral $\frac{1}{a}\int_0^1 dy\ y(1-y)^N$. This evaluates to $\frac{1}{a}\frac{1}{(N+2)(N+1)}$. Plugging this into **Eq. 3.6** gives: $P_N = \frac{1}{N} + \frac{c}{2a}$, which has an N-independent c term and gives the approximation:

$$l_{N,N-1}(c) \approx 1 + \frac{c}{2a} \tag{3.15}$$

□

Moving to the standard normal distribution, Wergen $et\ al$ use a saddle point approximation to deal with the difficult format of our ubiquitous integral: $I(N)$. There it was found for large $N >> 1$:

$$I(N) \approx \frac{1}{N^2\sigma}\frac{4\sqrt{\pi}}{e^2}\sqrt{ln\left(\frac{N^2}{8\pi}\right)} \tag{3.16}$$

A clear derivation of this integral can be found in Appendix A.2. Due to the number of rough approximations made in the working, only an estimate was made on the 'effect of drift on correlations' [11]:

$$(l_{N,N-1}(c) - 1) \propto -N\frac{c}{\sigma}\sqrt{ln\left(\frac{N^2}{8\pi}\right)} \tag{3.17}$$

This approximation saturates very quickly with increasing N, a problem similar to the uniform approximation. Overall, the exponential distribution exhibits N *independent* positive correlations in the small c regime, whilst the Gaussian shows strongly N *dependent* negative correlations much like distributions in the Weibull class.

3.3.3 Fréchet Class

The Pareto distribution: $f(x) = \mu x^{-\mu-1}$ is a well-known member of the Fréchet class. In [9] it was found that:

$$I(N) = \mu\frac{\Gamma(2+\frac{1}{\mu})\Gamma(N+1)}{\Gamma(N+\frac{1}{\mu}+3)} \tag{3.18}$$

Proof. The proof stems from solving the integral: $\mu^2\int_1^\infty dx x^{-2-2\mu}(1-x^{-\mu})^N$. Let $y = x^{-\mu}$ and $x^{\mu+1}dy = -\mu dx$. Our new upper and lower limits are 0 and 1, respectively. By noting $-\mu\int_1^0 = \mu\int_0^1$ we are left with an integral: $\mu\int_0^1 x^{-(1+\mu)}(1-y)^N$. The final step involves noting that $x^{-(1+\mu)}$ is equal to $y^{1+\frac{1}{\mu}}$ which leaves us with the exact form of a Beta function $B(x,y)$ where $x = 2+\frac{1}{\mu}$ and $y = N+1$. □

This is gives all the required ingredients to find $J(N)$. A simplification of $J(N)$ in algebra software such as Mathematica using the FullSimplify function gives the awkward expression:

$$J(N) = \frac{(N-1)(-2+N(4-4\mu+N(-3+N+3\mu)))\Gamma(2+\frac{1}{\mu})\Gamma(N-2)}{2\Gamma(N+\frac{1}{\mu}+1)} \tag{3.19}$$

And finally taking the large N limit and evaluating asymptotic behaviour gives: $J(N) = \Gamma(2+\frac{1}{\mu})N^{1-\mu^{-1}}/2$. Note: to find the limiting form of the polynomial expression I used the **mainTerm** function in Mathematica [16](a summary of this can be found in Appendix B2). The overall expression for the correlation function is then.

$$l_{N,N-1}(c) \approx 1 + \frac{c}{2}\Gamma(2 + \frac{1}{\mu})N^{1-\mu^{-1}} \tag{3.20}$$

From this, one would expect positive record correlations which increase at a rate less than linear with N. The analytical prediction matches roughly with numerical simulations. In **Fig. 3.1** one can also see that the Lévy stable distribution shows positive correlations in the same routine. For this reason, it is conjectured that all distributions in the Fréchet class exhibit the same asymptotic correlation behaviour.

3.4 Unified Picture

The paper concludes with a comparison of the 'asymptotics of the correction term $c\frac{N(N-1)}{2}I(N-2)$' with the i.i.d behaviour $P_N \approx \frac{1}{N}$ [9]. The small drift expansion to first order c means that for very large N, the correction term will diverge from the i.i.d case to the value $\lim_{N\to\infty} P_N(c)$. This is equivalent to saying that there exists some timescale $\hat{N}$ whereby $\frac{1}{\hat{N}} \sim cI_{\hat{N}}$ where I define $I_{\hat{N}} = c\frac{\hat{N}(\hat{N}-1)}{2}I(\hat{N}-2)$. The critical value $\hat{N}$ is then found for three EVS classes by plugging in the appropriate $I_{\hat{N}}$:

$$\hat{N} \propto c^{-\frac{\xi}{1+\xi}} \qquad \text{Weibull Class} \tag{3.21}$$

$$\hat{N} \propto c^{-1}|ln(c)|^{\frac{1}{\beta}-1} \qquad \text{Gumbel Class} \tag{3.22}$$

$$\hat{N} \propto c^{-\frac{\mu}{\mu-1}} \qquad \text{Fréchet Class} \tag{3.23}$$

Proof. I will only provide proof for the Weibull Class as the same steps can be used to find the other two expressions. By noting that the correction term in this case is $I_{\hat{N}} = \frac{\xi}{2}\Gamma(2-\frac{1}{\xi})N^{\frac{1}{\xi}} \propto N^{\frac{1}{\xi}}$ the expression $\hat{N} \sim \frac{1}{c}\frac{1}{I_{\hat{N}}}$ gives: $\hat{N} \propto \frac{1}{c}\frac{1}{N^{\frac{1}{\xi}}}$ and finally: $\hat{N} \propto c^{-\frac{\xi}{1+\xi}}$. □

To conclude, this paper defined a correlation measure $l_{N,N-1}(c)$ between random variables drawn from a linear drift framework. Most importantly, distributions from the Weibull class show negative correlations, whilst distributions from the Fréchet class show positive correlations. For the Gumbel class, the exponential and standard normal show positive and negative correlations, respectively, which can be characterised on the stretch exponent β of the general class of Gumbel Distributions: $f(x) = A_\beta e^{-|x|^\beta}$ [11].

4. Rounding Effects in Record Statistics

The second element of the project relies on many of the results derived by Wergen $et\ al$ in [12]. Here, they revert back to the i.i.d assumption in order to study the effect of rounding each random variable to a closest integer multiple of Δ. More precisely, if X_i is the i'th i.i.d random variable drawn, then it gets rounded to a value: $X_i^{\Delta} = k\Delta$. Logically, this should depress the record rate as for a new drawn value to be a record it has to cross a higher threshold to prevent it from being rounded down to the previous record integer multiple. Most importantly, rounding to a scale Δ discretises each random variable and for this reason, the record rate $P_N = \int dx f(x) F(x)^{N-1}$ has to be modified with a Riemann sum:

$$\begin{aligned} P_N^{\Delta} &= \sum_k \left[\int_{k\Delta}^{(k+1)\Delta} dx\ f(x) \right] F(k\Delta)^{N-1} \\ &= \sum_k \left[F((k+1)\Delta)) - F(k\Delta) \right] F(k\Delta)^{N-1} \end{aligned} \tag{4.1}$$

Here, $F(k\Delta)^{N-1}$ gives the probability that the first $N-1$ random variables are less than some integer multiple of Δ. Next, this is combined with the probability that the N'th random variable takes a record value between $k\Delta$ and the next integer multiple $(k+1)\Delta$. By summing over over all integers exhausts all possible record values that can be taken when random variables are drawn from a certain distribution.

4.1 Extreme Value Statistics Classes

Results for the strong record rate P_N^{Δ} are given to identify whether distributions belonging to the same Extreme Value Statistic class exhibit similar record behaviour.

4.1.1 Weibull Class

For the uniform distribution on [0,1] with $\Delta = \frac{1}{L}$, $L > 0$ one can set $f(x) = 1$, $F(x) = X$ and plug into **Eq. 4.1** to give:

$$P_N^{\Delta} = \Delta^N \sum_{k=1}^{\frac{1}{\Delta}-1} k^{N-1} \tag{4.2}$$

Proof. $P_N^{\Delta} = \sum_k^{L-1} \left[\int_{k\Delta}^{(k+1)\Delta} dx\ 1 \right] (k\Delta)^{N-1} = \sum_k^{L-1} [(k+1)\Delta - k\Delta](k\Delta)^{N-1} = \Delta^N \sum_{k=1}^{\frac{1}{\Delta}-1} k^{N-1}$ □

A useful trick employed is replacing the Riemann sum with an integral when the discretisation scale is very small i.e $\Delta << 1$. Thus we then have:

$$P_N^{\Delta} \approx \Delta^N \int_{k=1}^{\frac{1}{\Delta}-1} dk\ k^{N-1} = \frac{1}{N} \left[(1-\Delta)^N - \frac{\Delta^N}{N} \right] \approx \frac{1}{N} (1-\Delta)^N$$

where $\frac{\Delta^N}{N}$ can be ignored as it is negligibly small. Most importantly, the strongly rounded record rate is not $\frac{1}{N}$ but instead decays at an exponential rate to 0 for increasing N and thus, we'll see a smaller number of records over time.

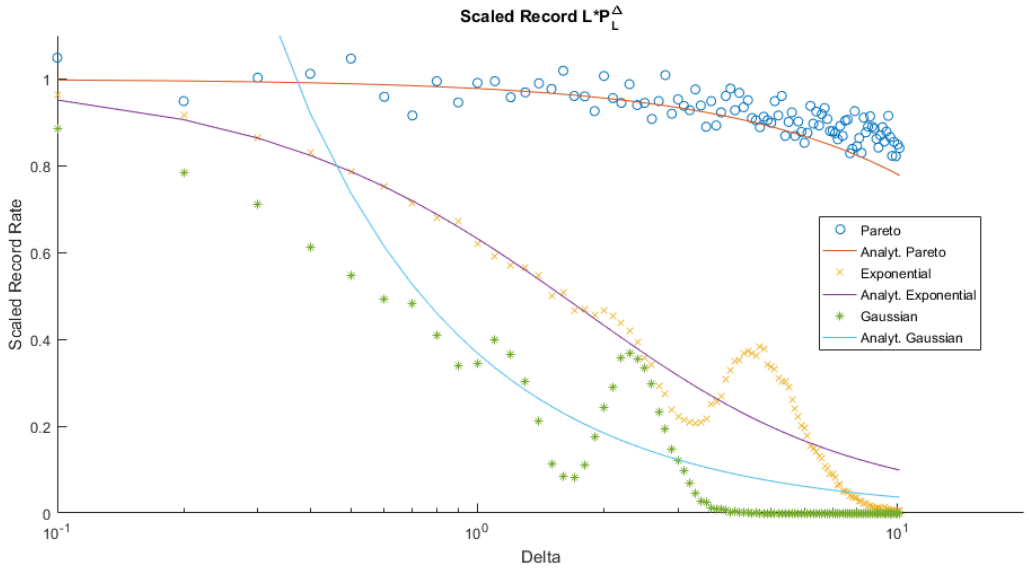


Figure 4.1: Scaled Record Rate $N * P_N^{\Delta}$ for the Pareto Distribution $\mu - 1.2$, Exponential Distribution $\lambda = 1$ and Standard Normal Distribution. The scatter plots represent numerical simulations averaged 10^5 times and the analytical predictions stem from approximating **Eq. 4.1** for the 3 different distributions. Matlab code can be found in **Appendix B.3**.

.

Due to the small support of the standard uniform distribution, scaling effects will be strongest compared to other distributions.

4.1.2 Gumbel Class

One simple example for the Gumbel class is the exponential distribution with $\lambda = 1$: $f(x) = e^{-x}$. Noting that $F((k+1)\Delta) - F(k\Delta) = 1 - e^{-(k+1)\Delta} - (1 - e^{-k\Delta}) = e^{-k\Delta}(1 - e^{-\Delta})$, replacing the sum with an integral for $N >> 1$ and using the substitution $v = k\Delta$, **Eq 4.1** transforms into:

$$\begin{aligned} P_N^{\Delta} &= \sum_{k=1}^{\infty} e^{-k\Delta}(1 - e^{-\Delta})(1 - e^{-k\Delta})^{N-1} \approx \int_{k=1}^{\infty} dk\, e^{-k\Delta}(1 - e^{-\Delta})(1 - e^{-k\Delta}) \\ &= \frac{1}{\Delta}(1 - e^{-\Delta}) \int dv\, e^{-v}(1 - e^{-v})^{N-1} = \frac{1}{N\Delta}(1 - e^{-\Delta}) \end{aligned} \tag{4.3}$$

where the final step comes from the fact that the record rate $P_N = \frac{1}{N}$ for continuous density $f(v) = e^{-v}$. In **Fig. 4.1**, this analytical prediction is in agreement with numerical findings. Furthermore, in the small Δ regime, a first order expansion of **Eq. 4.3** gives $P_N^{\Delta} \approx \frac{1}{N}(1 - \frac{\Delta}{2})$. Clearly, the presence of rounding depresses the record rate (asymptotic decay as $\frac{1}{N}$) much less than the Weibull class (exponential decay). This is unsurprising because the support of the exponential distribution is much broader.

Similarly, for the Gaussian distribution $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ with $N >> 1$, **Eq. 4.1** transforms into:

$$P_N^{\Delta} = \frac{1}{2} \int dk \left[\text{erfc}\left(\frac{k\Delta}{\sqrt{2}}\right) - \text{erfc}\left(\frac{(k+1)\Delta}{\sqrt{2}}\right) \right] F(k\Delta)^{N-1}$$

The following expression was derived for the rounded record rate using the substitution $x = k\Delta$, an infinite series expansion of the complimentary error function erfc and the Laplace method:

$$P_N^{\Delta} \approx \frac{1}{N\Delta}\left[\ln\left(\frac{N^2}{2\pi}\right)\right]^{-1} \tag{4.4}$$

Due to the similarity in the derivation steps to that of **Appendix A.2**, I have excluded the proof method. Most noticeably, the rounded record probability saturates only slightly faster than the record rate $P_N = \frac{1}{N}$ (in contrast to the previous class) and shows similar behaviour to that of the exponential distribution. This can be seen in **Fig 4.1** and it is easy to see that the analytical prediction is only useful in the $\Delta >> 1$ regime as, for $\Delta << 1$, **Eq. 4.4** diverges rapidly.

4.1.3 Fréchet Class

For the Fréchet class: the Pareto distribution $f(x) = \mu x^{-\mu-1}$, with $x > 1$ and $\mu > 0$ is given. Using $F((k+1)\Delta) - F(k\Delta) = (k\Delta)^{-\mu} - ((k+1)\Delta)^{-\mu}$ and plugging into **Eq. 4.1** gives:

$$P_N^{\Delta} \approx \int_1^{\infty} dk \left[(k\Delta)^{-\mu} - ((k+1)\Delta)^{-\mu}\right](1-(k\Delta)^{-\mu})^{N-1}$$

With a final approximation:

$$P_N^{\Delta} = \frac{1}{N}\left[1 - \frac{\Delta}{2}\mu\Gamma\left(2+\frac{1}{\mu}\right)N^{-1/\mu}\right] \tag{4.5}$$

The rounding effect here is very small compared to other distributions and it can be seen that as $N \to \infty$: $P_N^{\Delta} \to P_N$. This is in almost exact agreement with the numerical results **Fig 4.1**.

4.2 Small Δ Expansion

Much like the previous LDM paper, Wergen *et al* provide an analysis of all EVS classes within the small Δ regime. Clearly, the effects of rounding with a small scale has much more real life significance than a larger scale. For example, $\Delta = 5$, means that for a random variable to be a record it has to cross a threshold of 5 each time, a very unlikely event for a probability distribution, say the standard Gaussian. The expression derived in **Eq. 4.1** is simplified to the following:

$$\begin{aligned} P_N^{\Delta} &= \sum_k \left[\int_{k\Delta}^{(k+1)\Delta} dx\ f(x)\right] F(k\Delta)^{N-1} \\ &= \int dx f(x) F(\lfloor x \rfloor_{\Delta})^{N-1} \\ &\approx \frac{1}{N} - N\int dx (x - \lfloor x \rfloor_{\Delta}) f(x)^2 F(x)^{N-2} \end{aligned} \tag{4.6}$$

Proof. Here, the floor function is the 'largest integer multiple of Δ that is smaller than x' [12] and is a convenient way of re-writing $k\Delta$. Mathematically, this the same as saying that $k\Delta \leq x < (k+1)\Delta$ which removes the need for a sum over k. I will elaborate on the proof for the last line. The first order expansion around x gives us: $F(\lfloor x \rfloor_{\Delta})^{N-1} \approx [F(x) + (\lfloor x \rfloor_{\Delta} - x)f(x)]^{N-1}$. After expanding this approximation and ignoring $(\lfloor x \rfloor_{\Delta} - x)^{N-1} f(x)^{N-1}$ terms we are left with:

$$\begin{aligned} P_N^{\Delta} &= \int dx f(x)\left[F(x)^{N-1} - N(x - \lfloor x \rfloor_{\Delta}) f(x) F(x)^{N-2}\right] \\ &\approx \frac{1}{N}\left[1 - \frac{\Delta}{2}N^2 I_N\right] \end{aligned} \tag{4.7}$$

where $I_N = \int dx f(x)^2 F(x)^{N-2}$ Again the first term of the equation is exactly $\frac{1}{N}$ (derived earlier). Next, the authors apply the approximation $x - \lfloor x \rfloor_{\Delta} \approx \frac{\Delta}{2}$. □

5. Methodology

The first weeks of the project will be devoted to becoming familiar with the material referenced and reproducing known results to become comfortable with simulations. To choose what line of research to undertake after this I first identified three typical questions from the field: what is the probability of a record occurring at step N in a time series? What is the probability of adjacent records in a time series? What is the average number of records up to a certain step N in the time series?

Firstly, where non-i.i.d RVs are discretised through rounding, research into the first question hadn't been performed. Therefore, I will attempt to extend to the work done by Krug [6] with this in mind. Secondly, in the case of i.i.d random variables in the presence of rounding off effects, the second question is unanswered which inspired a derivation for a joint rounded record rate formula. Finally, none of the three questions had been addressed where random variables contain an additive drift term but are also rounded. Therefore, I will attempt to answer the first and third question within this framework with the second question, time permitting. Overall, a chronological ordering of research is placed below.

1. Define a rounded record rate P_L^{Δ} for non-i.i.d RVs in a time series.
2. Applying this definition to RVs with changing distribution shape parameters across a time series. Find approximations for a few well-known distributions.
3. Define a joint rounded record rate $P_{L,L-1}^{\Delta}$ for non-i.i.d RVs in a time series. Attempt to quantify a correlation term $A_{L,L-1}^{\Delta}$ which links the joint record rate $P_{L,L-1}^{\Delta}$ its marginals $P_L^{\Delta} P_{L-1}^{\Delta}$.
4. Define a rounded record rate $P_L^{\Delta}(c)$ for non-i.i.d RVs in the presence of additive drift c. Perform analysis of the interconnectedness between c and Δ for different distributions.
5. Derive a small c and small Δ regime expansion and apply this to the EVS classes.
6. Look at the rounded record number statistic $\left\langle R_L^{\Delta} \right\rangle$ with regards to linear drift RVs.

Classical techniques from Extreme Value Statistics will be used in the derivations of analytical expressions and statistical physics methods (Laplace, saddle point) will be employed for approximations.

6. Rounding effects on Independent but Non-i.i.d Random Variables

The first progression in combining rounding effects with a non-i.i.d assumption involves deriving an expression for the case of independent but not necessarily identically distributed random variables. Logically one can write this like so:

$$P_L^{\Delta} = \sum_k \left[\int_{k\Delta}^{(k+1)\Delta} dx \; f_L(x) \right] \prod_{j=1}^{L-1} F_j(k\Delta) \tag{6.1}$$

Again, the probability that the first $L-1$ random variables are smaller than a given value $k\Delta$ factorises into individual cumulative probabilities: $F_j(k\Delta)$. The L'th RV then must take a value in a higher threshold than $k\Delta$, which can be seen through the upper and lower limits in the integral. Finally, by summing over the possible integer scalar multiple's of Δ in the support of the distribution, one arrives at **Eq. 6.1**. The results can be seen in **Fig 6.1** which confirm the above analytical expression.

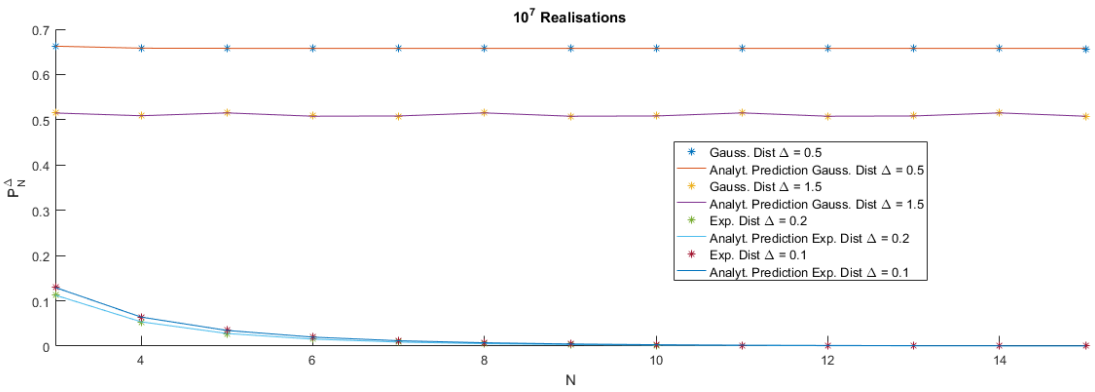


Figure 6.1: Simulations for changing parameters for the Gaussian (increasing mean μ) and Exponential Distribution (increasing mean λ). 10^7 realisations were used. Analytical predictions (plotted line curves) match exactly confirming the existence of **Eq. 6.1**.

Moreover, one can focus on applying this to broadening (or sharpening) distributions whereby random variables are drawn from the same underlying distribution but with increasing (or decreasing) shape parameters.

Studies on this have been performed in [6] without the presence of rounding off effects. Importantly, it was noted that in the study of the effect of global warming records, change or variability in the climate of a country has more relevance than averages. This explains the investigation of the change in shape of the distribution. By finally combining this with the presence of rounding effects ties in nicely with the conclusions found in [12], whereby 'rounding effects should be carefully accounted for when using record statistics to detect secular trends in data.'

Here, the probability density function $f_L(x)$ of the L'th entry X_L is of the form:

$$f_L(X) = \lambda_L f(\lambda_L X) \tag{6.2}$$

where $f(X)$ is a probability distribution originating from one of the three EVS classes. Similarly, its cumulative density function $F_L(x) = F(\lambda_L X)$ for underlying cdf $F(X)$. For example, $\lambda_L = \frac{1}{L}$ and $f(X) = e^{-X}$ corresponds to an exponential distribution with a broader distribution for each RV. In general, it is assumed that the λ_N follows a power-law form: $\lambda_N = \lambda_0 N^{-\alpha}$ where $\alpha > 0$ ($\alpha < 0$) corresponds to a broadening (sharpening) distribution [6].

Applying this to **Eq. 6.1** gives:

$$P_L^{\Delta} = \sum_k \left[\int_{k\Delta}^{(k+1)\Delta} dx\ \lambda_N f(\lambda_L x) \right] \prod_{J=1}^{L-1} F(\lambda_J k\Delta) \tag{6.3}$$

The substitution: $z = \lambda_N x$, $dx = dz/\lambda_N$ and using $\lambda_L = L^{-\alpha}$ gives:

$$\begin{aligned} P_L^{\Delta} &= \sum_k \left[\int_{\lambda_L k\Delta}^{\lambda_L (k+1)\Delta} dz\ f(z) \right] \prod_{j=1}^{L-1} F(\frac{\lambda_J}{\lambda_L} k\Delta) \\ &= \sum_k \left[F\left(\frac{(k+1)\Delta}{L^{\alpha}} \right) - F\left(\frac{k\Delta}{L^{\alpha}} \right) \right] \prod_{J=1}^{L-1} F\left(\frac{k\Delta}{J^{\alpha}} \right) \end{aligned} \tag{6.4}$$

The behaviour of P_L^{Δ} is difficult to solve exactly. By, increasing the probabilistic range over which random variables can be drawn increases the record rate, whilst rounding effects have the opposite effect.

However, drawing random variables from an exponential distribution $f(X) = e^{-X}$ with a mean incrementing for each random variable allows one to find an expression asymptotically. Note: I have taken $\alpha = -1$ here. Using **Eq. 6.4**:

$$P_L^{\Delta} = (1 - e^{-L\Delta}) \sum_k e^{-Lk\Delta} \left[\prod_{j=1}^{L=1} (1 - e^{-jk\Delta}) \right] \tag{6.5}$$

To find a closed form expression one can replace the sum over k with an integral for $\Delta << 1$ and $L >> 1$ to obtain:

$$P_L^{\Delta} = (1 - e^{-L\Delta}) \int dk\ e^{-Lk\Delta} (e^{-k\Delta}; e^{-k\Delta})_{L-1} \tag{6.6}$$

where $(q;q)_{L-1} = \prod_{j=1}^{\infty} (1 - q^j)$ is the q-Pochhammer function and a special case of the Euler Function with $q = e^{-k\Delta}$ [18]. Using the asymptotic result: $(q;q)_{\infty} = \sqrt{(2\pi)/t} \exp\{-\pi^2/6t + t/24\} + o(1)$ where $q = e^{-t}$ and $t = k\Delta$ [5]:

$$\begin{aligned} P_L^{\Delta} &\approx \sqrt{2\pi}(1 - e^{-L\Delta}) \int \frac{dk}{\sqrt{k\Delta}} e^{-Lk\Delta} e^{-\frac{\pi^2}{6k\Delta} + \frac{k\Delta}{24}} \\ &= \sqrt{2\pi}(1 - e^{-L\Delta}) \frac{1}{\Delta} \int \frac{dx}{\sqrt{x}} e^{-Lx} e^{-\frac{\pi^2}{6x} + \frac{x}{24}} \\ &= \sqrt{2\pi}(1 - e^{-L\Delta}) \frac{1}{\Delta} \int dx \exp\left\{ -Lx - \frac{\pi^2}{6x} + \frac{x}{24} - \frac{1}{2}\ln(x) \right\} \end{aligned} \tag{6.7}$$

Taking $g(x) = -Lx - \frac{\pi^2}{6x} + \frac{x}{24} - \ln(\sqrt{x})$ to use in our Laplace method integral approximation: $g'(x) = -L + \frac{\pi^2}{6x^2} + \frac{1}{24} - \frac{1}{2x} = 0$ gives a global maximum:

$$\begin{aligned} \hat{x} &= 2\left(\frac{-3+\sqrt{9+(24L-1)\pi^2)}}{24L}\right) \\ &\approx \frac{\pi}{\sqrt{24L}} \end{aligned} \tag{6.8}$$

Plugging this maximum into $g''(x) = -\frac{\pi^2}{3x^3} + \frac{1}{2x^2}$ gives a complicated expression:

$$\begin{aligned} P_L^{\Delta} &\approx \sqrt{2\pi}\frac{(1-e^{-L\Delta})}{\Delta}\sqrt{\frac{-2\pi}{g''(\hat{x})}}e^{g(\hat{x})} \\ &= \sqrt{2\pi}\frac{(1-e^{-L\Delta})}{\Delta}\sqrt{\frac{6*\pi^2}{\pi 24L^{\frac{3}{2}}-36L}}\exp\left[\frac{-L\pi}{\sqrt{24L}} - \frac{\pi}{6}\sqrt{24L} + \frac{\pi}{24\sqrt{24L}} - \ln\left(\frac{\pi}{\sqrt{24L}}\right)\right] \\ &= \sqrt{2\pi}\frac{(1-e^{-L\Delta})}{\Delta}\frac{\sqrt{24L}}{\pi}\sqrt{\frac{6*\pi^2}{\pi(24L)^{\frac{3}{2}}-36L}}\exp\left[\frac{-L\pi}{\sqrt{24L}} - \frac{\pi}{6}\sqrt{24L} + \frac{\pi}{24\sqrt{24L}}\right] \\ &\approx \sqrt{L}\frac{4\sqrt{3}}{\sqrt{\pi}}\frac{(1-e^{-L\Delta})}{\Delta}\sqrt{\frac{6\pi}{(24L)^{\frac{3}{2}}}}\exp\left\{-\frac{5\sqrt{L}\pi}{2\sqrt{6}}\right\} \end{aligned} \tag{6.9}$$

The mainTerm function (**Appendix B.2**) was used in the final step approximation to tidy up the function. This gives results that are fairly accurate compared to numerical simulations which can be seen in **Fig. 6.2**.

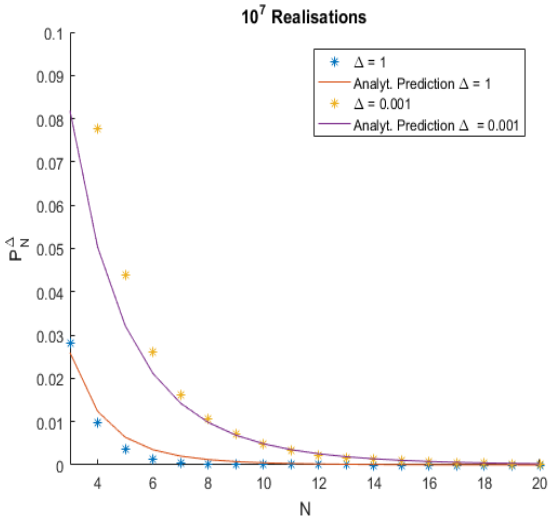


Figure 6.2: The rounded record rate P_L^{Δ} whereby the shape of the exponential distribution is sharpened for each RV drawn. Monte Carlo simulations (scatter plot) have been averaged 10^6 times. The inverse transform method [19] was used to generate exponential random variables with changing shape parameter. Analytical **Eq. 6.8** predicts the decay of the record rate quite accurately even with a number of approximations made in the derivation.

With an increasing exponential mean λ, the range over which random variables are drawn decreases more and more, depressing the probability of there being a record for relatively large L. This is then combined with a depressing rounding effect and overall can be seen with the analytical expression which decreases exponentially in L.

Returning to **Eq 6.1**, for the Weibull class of bounded distributions one can take the case where random variables are drawn from a uniform distribution on $(0, j)$ with j incrementing across the time series. As the width of the distribution increases for each RV, the rounded record rate should logically be higher than the previous $U(0, 1)$ i.i.d case derived in Chapter 4. Our expression is then:

$$P_L^{\Delta} = \frac{\Delta^L}{L!} \sum_{k=1}^{\frac{L}{\Delta}-1} k^{L-1} \tag{6.10}$$

Proof. Using: $f_j(x) = \frac{1}{j}$ and $F_j(x) = \frac{x}{j}$ one obtains: $P_L^{\Delta} = \frac{\Delta}{L} \sum_k \prod_{j=1}^{L-1} \frac{k\Delta}{j} = \frac{\Delta^L}{L} \sum_k \frac{k^{L-1}}{(L-1)!} = \frac{\Delta^L}{L!} \sum_k k^{L-1}$. □

For a small discretisation scale $\Delta << 1$ one replaces the sum by an integral and finds:

$$\begin{aligned} P_L^{\Delta} &= \frac{\Delta^L}{L!} \int_{k=1}^{\frac{L}{\Delta}-1} dk\, k^{L-1} = \frac{\Delta^L}{L!} \frac{1}{L} \left[\frac{(L-\Delta)^L}{\Delta^L} - 1 \right] \\ &\approx \frac{1}{L} \frac{(L-\Delta)^L}{L!} \end{aligned} \tag{6.11}$$

Which is only valid where $L - \Delta \leq 1$, a contradiction as we've examined the small Δ regime. The problem stems from this. The sum in **Eq 6.1** is over integers between $k = 1$ to $\frac{L}{\Delta} - 1$ to include all lattice sites on a $(0, L)$ range. This means if we have L RVs drawn from a uniform distribution with increasing width $(0, 1)$ to $(0, L)$ then the summation range increases for each new RV drawn. Therefore, this analysis is only valid on changing distribution parameters with the same support region throughout the time series.

Finally, for the Fréchet class, taking the Pareto distribution $f(x) = \mu x^{-\mu-1}$, $F(x) = 1 - x^{-\mu}$ and $\alpha = 1$, **Eq. 6.4** transforms into:

$$\begin{aligned} P_L^{\Delta} &= \sum_k \left[\left(\frac{k\Delta}{L} \right)^{-\mu} - \left(\frac{(k+1)\Delta}{L} \right)^{-\mu} \right] \prod_{J=1}^{N-1} \left[1 - \left(\frac{k\Delta}{J} \right)^{-\mu} \right] \\ P_L^{\Delta} &= \sum_k \left[\left(\frac{L}{k\Delta} \right)^{\mu} - \left(\frac{L}{(k+1)\Delta} \right)^{\mu} \right] \prod_{J=1}^{N-1} \left(1 - \left(\frac{J}{k\Delta} \right)^{\mu} \right) \end{aligned} \tag{6.12}$$

By now considering the asymptotic record rate (replacing the sum with an integral) with substitution $x = \frac{L}{k\Delta}$, $dk = -\frac{L}{x^2\Delta} dx$, replacing appropriate integral limits.

$$P_L^{\Delta} \approx \frac{L}{\Delta} \int_0^{\frac{L}{\Delta}} \frac{dx}{x^2} \left[x^{\mu} - \left(\frac{Lx}{(L + \Delta x)} \right)^{\mu} \right] \prod_{j=1}^{L-1} \left[1 - \left(\frac{J}{L} x \right)^{\mu} \right] \tag{6.13}$$

Unfortunately, I was unable to simply this expression further to obtain a useful approximation. Because the Pareto distribution has a relatively small effect on the presence of rounding off effects, broadening the distribution will result in P_L^{Δ} saturating at a much *slower* rate than the i.i.d case P_L.

6.1 Joint Rounded Record Rate

In Chapter 3, **Eq 3.4** gave the joint probability of two consecutive independent but not identically distributed RV records occurring. One can attempt to extend this to the rounded case above. Firstly, in the case of **i.i.d random variables** the classical result that a joint probability decomposes into its marginals does not hold:

$$P^{\Delta}_{L,L-1} \neq P^{\Delta}_{L} P^{\Delta}_{L-1}$$
$$= \sum_{k} \left[\int_{k\Delta}^{(k+1)\Delta} dx\ f(x) \right] F(k\Delta)^{L-1} \star \sum_{j} \left[\int_{j\Delta}^{(j+1)\Delta} dx\ f(x) \right] F(j\Delta)^{L-2}$$

where k and j represent integer summation over the support of the pdf. Interestingly, in the presence of rounding effects adjacent random variables are in some way correlated with similar properties. An easy example to show this is by considering a large Δ regime whereby almost all RV in a time series are rounded down to the same threshold thus producing ties. For this reason, one can predict with great certainty that adjacent $N-1$ and N random variables will rounded down to the same threshold and therefore are no longer independent. Initial simulations suggest that the disparity between the joint rounded record rate and its marginals increases as Δ increases.

To tackle this, one ansatz could be something like below:

$$P^{\Delta}_{L,L-1} = \sum_{k} \left[\int_{(k+1)\Delta} dx\ f(x) \right] \left[\int_{k\Delta}^{(k+1)\Delta} dx\ f(x) \right] F(k\Delta)^{L-2} \tag{6.14}$$

i.e the probability that first $L-2$ random variables take a value less than some scalar multiple of Δ: $k\Delta$ combined with the probability that the $L-1$'th RV is then rounded to the next threshold greater than $k\Delta$, and finally multiplied by the probability that the L'th RV, is greater than the previous record threshold $(k+1)\Delta$. Summing over all sites k leaves a joint probability.

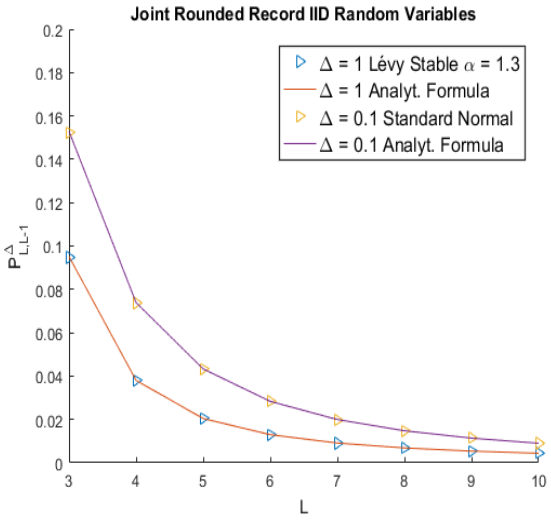


Figure 6.3: The joint rounded record rate $P^{\Delta}_{L,L-1}$ for the Lévy Stable $\alpha = 1.3$ Distribution and standard Gaussian Distribution with i.i.d Random Variables. The scatter plot (right-pointing triangles) are from Monte Carlo simulations which have been averaged 10^6 times. The thick lines are a plot of **Eq. 6.14**. Based on the analytical and numerical match for different distributions and Δ, it is very likely this is an exact formula for the joint rounded record rate.

Testing this formula has been exact for the Lévy Stable $\alpha = 1.3$ Distribution and standard Gaussian Distribution with i.i.d Random Variables. This can be seen in **Fig. 6.3**. I have performed extensive numerical simulations for different distributions under changing Δ regimes and the formula holds.

Furthermore, in the case of **non-i.i.d** RV: I replaced $F(k\Delta)^{L-2}$ with $\prod_{j=1}^{L-2} F_j(k\Delta)$ (this logic is exactly the same as the derivation for **Eq. 6.1**). Testing this formula has been exact for the Lévy Stable Distribution and Normal

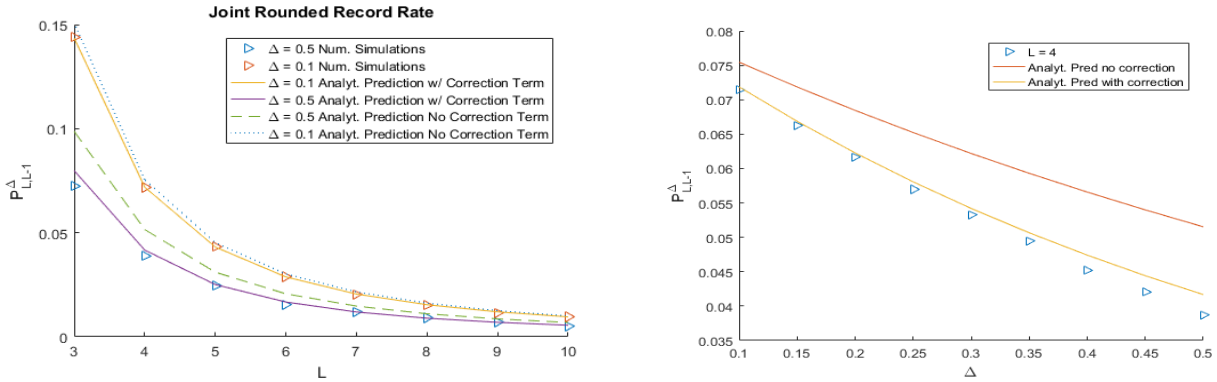
Distribution with changing index stability α and increasing variance, respectively and varying Δ's.

Moreover, one can conjecture that the joint rounded record rate is equal to the multiplication of its marginals with an inclusion of an additive term or cofactor. For example:

$$P^{\Delta}_{L,L-1} = A^{\Delta}_{L,L-1} \star P^{\Delta}_{L} P^{\Delta}_{L-1} \tag{6.15}$$

where A is some function of Δ and L and could be either a positive or negative **correlation**. A simple example can highlight this: take the standard exponential distribution with $f(x) = e^{-x}$ and $F(x) = 1 - e^{-x}$ in the small $\Delta << 1$ regime:

$$\begin{aligned} P^{\Delta}_{L,L-1} &= \sum_{k} \left[1 - (1 - e^{-(k+1)\Delta})\right] \left[e^{-k\Delta} - e^{-(k+1)\Delta}\right] (1 - e^{-k\Delta})^{L-2} \\ &= e^{-\Delta}(1 - e^{-\Delta}) \sum_{k} e^{-2k\Delta} (1 - e^{-k\Delta})^{L-2} \\ &\approx e^{-\Delta} \frac{(1 - e^{-\Delta})}{\Delta} \int_{1-e^{-\Delta}}^{1} du (1-u) u^{L-2} \approx e^{-\Delta} \frac{(1 - e^{-\Delta})}{\Delta} \frac{1}{L(L-1)} \\ &\approx e^{-\Delta} (1 - \frac{\Delta}{2}) \frac{1}{L(L-1)} \end{aligned} \tag{6.16}$$



(a) (b)

Figure 6.4: **Fig.(a)** shows the Joint Rounded Record Rate for standard exponentially distributed random variables in a time series from $L = 3$ to 10. Numerical Simulations for $\Delta = 0.1$ and $\Delta = 0.5$ are shown in the scatter plot. The thick lines show the analytical prediction after considering the correction term $A^{\Delta}_{L,L-1}$ whilst the dashed lines are the analytical formula $P^{\Delta}_{L} P^{\Delta}_{L-1}$ without correction. **Fig.(b)** isolates a specific random variable in the time series (in this case $L = 4$) and varies Δ from 0.1 to 0.5. The analytical prediction post correction shows more accurate results.

where the final step comes from a first order expansion of $\frac{(1-e^{-\Delta})}{\Delta}$ in the small Δ regime. I have replaced the sum with an integral and made the substitution $u = (1 - e^{-k\Delta})$ in the third step. Furthermore, I have only included dominant terms in the large L limit when solving the integral in the same step.

Using **Eq. 4.3** one also finds that $P^{\Delta}_{L} P^{\Delta}_{L-1} \approx \frac{1}{L(L-1)}(1 - \frac{\Delta}{2})^2$. Plugging into the conjecture $P^{\Delta}_{L,L-1} = A^{\Delta}_{L,L-1} \star P^{\Delta}_{L} P^{\Delta}_{L-1}$ one finds:

$$A^{\Delta}_{L,L-1} \approx \frac{e^{-\Delta}}{(1 - \frac{\Delta}{2})} \tag{6.17}$$

Proof. Plugging in expressions for $P^{\Delta}_{L,L-1}$ and $P^{\Delta}_{L}P^{\Delta}_{L-1}$ gives $e^{-\Delta}(1-\frac{\Delta}{2})\frac{1}{L(L-1)} \approx A^{\Delta}_{L,L-1}\frac{1}{L(L-1)}(1-\frac{\Delta}{2})^2$ with $A^{\Delta}_{L,L-1}$ following from a simple rearrangement. □

The results of testing this is summarised in **Fig. 6.4** and shows better agreement with simulations. Interestingly, $A^{\Delta}_{L,L-1}$ is independent of L and this matches with the N-independent correlation term in **Eq. 3.15**.

Furthermore, for the Weibull class one can use the uniform distribution: $f(x) = 1$ where $x \in [0, 1]$.

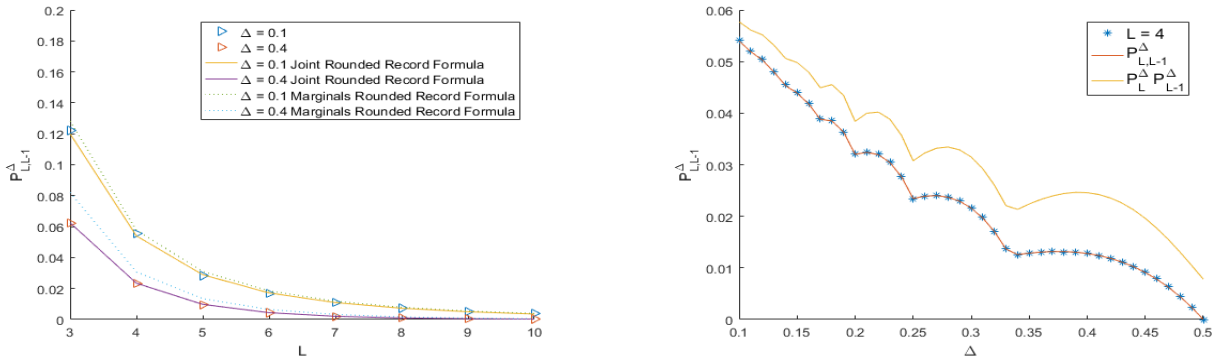
$$P^{\Delta}_{L,L-1} = \Delta^{L-1}\left[H^{(2-L)}_{\frac{1}{\Delta}-1} - \Delta\left(H^{(1-L)}_{\frac{1}{\Delta}-1} + H^{(2-L)}_{\frac{1}{\Delta}-1}\right)\right] \tag{6.18}$$

where the $H^{(r)}_n = \sum_{i=1}^{n}\frac{1}{i^{(r)}}$ is the n'th Harmonic number of order r [17].

Proof. Plugging $F(x) = x$ into **Eq 6.14** gives:

$$\begin{aligned} P^{\Delta}_{L,L-1} &= \sum_{k}^{\frac{1}{\Delta}-1}(1-(k+1)\Delta)((k+1)\Delta - k\Delta)(k\Delta)^{L-2} \\ &= \Delta^{L-1}\sum_{k}[1-(k+1)\Delta]\,k^{L-2} \\ &= \Delta^{L-1}\sum_{k}\left[k^{L-2} - k^{L-1}\Delta - k^{L-2}\Delta\right] \end{aligned} \tag{6.19}$$

Replacing sums with expression for the Harmonic number gives **Eq. 6.18**. □



(a) (b)

Figure 6.5: **Fig.(a)** shows the Joint Rounded Record Rate for uniform distributed random variables on $[0, 1]$ in a time series from $L = 3$ to 10. Numerical Simulations for $\Delta = 0.1$ and $\Delta = 0.4$ are shown in the scatter plot. The thick line shows the analytical prediction after considering the correction term $A^{\Delta}_{L,L-1}$ whilst the dashed line is the analytical formula $P^{\Delta}_{L}P^{\Delta}_{L-1}$ without correction. **Fig.(b)** isolates a specific random variable in the time series (in this case $L = 4$) and varies Δ from 0.1 to 0.5.

Using **Eq 4.2** from the original rounding paper the multiplication of marginal probabilities follows easily: $P^{\Delta}_{L}P^{\Delta}_{L-1} = \Delta^{L}\sum_{k=1}^{\frac{1}{\Delta}-1}k^{L-1} * \Delta^{L-1}\sum_{k=1}^{\frac{1}{\Delta}-1}k^{L-2} = \Delta^{2L-1}H^{(1-L)}_{\frac{1}{\Delta}-1}H^{(2-L)}_{\frac{1}{\Delta}-1}$. Next, making use of **Eq 6.15** allows one to find a correlation term. As no approximations have been made, this should represent an exact formula.

$$A^{\Delta}_{L,L-1} = \frac{\Delta^{-L}\left[H^{(2-L)}_{\frac{1}{\Delta}-1} - \Delta\left(H^{(1-L)}_{\frac{1}{\Delta}-1} + H^{(2-L)}_{\frac{1}{\Delta}-1}\right)\right]}{H^{(1-L)}_{\frac{1}{\Delta}-1}H^{(2-L)}_{\frac{1}{\Delta}-1}} \tag{6.20}$$

Proof.

$$A^{\Delta}_{L,L-1} = \frac{P^{\Delta}_{N,N-1}}{P^{\Delta}_{N} P^{\Delta}_{N-1}} = \frac{\Delta^{L-1}\left[H^{(2-L)}_{\frac{1}{\Delta}-1} - \Delta\left(H^{(1-L)}_{\frac{1}{\Delta}-1} + H^{(2-L)}_{\frac{1}{\Delta}-1}\right)\right]}{\Delta^{2L-1} H^{(1-L)}_{\frac{1}{\Delta}-1} H^{(2-L)}_{\frac{1}{\Delta}-1}}$$
$$= \frac{\Delta^{-L}\left[H^{(2-L)}_{\frac{1}{\Delta}-1} - \Delta\left(H^{(1-L)}_{\frac{1}{\Delta}-1} + H^{(2-L)}_{\frac{1}{\Delta}-1}\right)\right]}{H^{(1-L)}_{\frac{1}{\Delta}-1} H^{(2-L)}_{\frac{1}{\Delta}-1}}$$

□

In the next chapter, the same framework will be kept from Chapter 6 with the addition of a constant linear drift term c.

7. Rounding effects on Linear Drift Random Variables

With the confidence that **Eq. 6.1** is very likely to be an exact expression, one can move onto the analytical derivation in the case where random variables consist of an i.i.d component with additive linear drift. Both c and Δ play a different role in increasing and suppressing the rounded record rate, respectively.

Now writing RVs in the form $Y_l = X_l + cl$ one can attempt define a new rounded record rate based on the fact each value Y_l gets rounded to a value $\lfloor Y_l \rfloor_\Delta = \left\lfloor \frac{Y_l}{\Delta} \right\rfloor \star \Delta = \left\lfloor \frac{X_l + cl}{\Delta} \right\rfloor \star \Delta$. One ansantz could be:

$$\begin{aligned} P_L^\Delta(c) &= \sum_k \left[\int_{k\Delta}^{(k+1)\Delta} dx f(x) \right] \prod_{l=1}^{L-1} F(k\Delta + cl) \\ &= \sum_k \left[F((k+1)\Delta) - F(k\Delta) \right] \prod_{l=1}^{L-1} F(k\Delta + cl) \end{aligned} \tag{7.1}$$

Once again, this expression follows the same logic as the rounding effects on independent but not identically distributed RV. One has to replace $F_j(k\Delta)$ in **Eq. 6.1** with $F(k\Delta + cl)$, as the same underlying probability distribution is used across all RV. Problems arise with this, however, as we'll see.

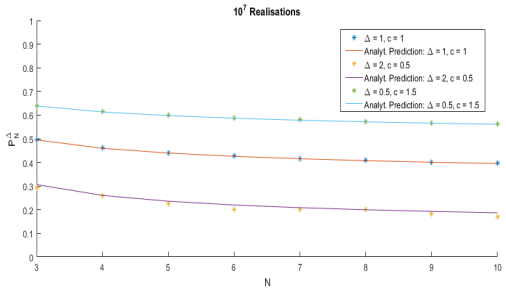


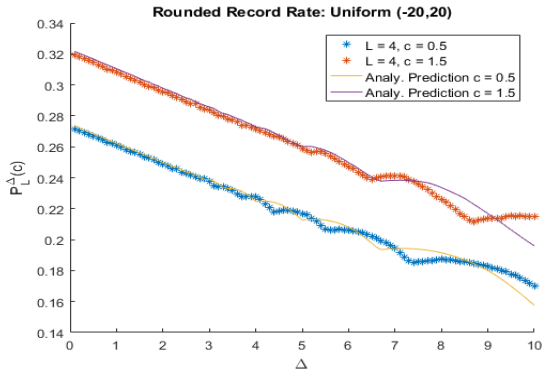
Figure 7.1: Numerical simulations (10^7 realisations) for Levy-Stable ($\mu = 1.3$) symmetrically distributed rounded random variables with constant linear drift. Analytical predictions for $P_L^\Delta(c)$ give excellent results across different c and Δ regimes. For $\Delta = 2$ however one starts to see discrepancies for $L = 6$.

Initial simulations using different distributions (**Fig. 7.1** and **Fig. 7.3**) show that **Eq. 7.1** is a very accurate expression where c and Δ are within a small sensible regime. Here, results match numerical simulations exactly. However, for increasing c and Δ the analytical prediction doesn't hold. One idea to explain this is: suppose c is large relative to Δ. The product in **Eq. 7.1**:

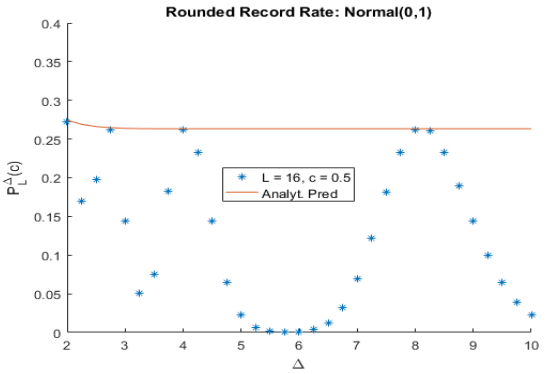
$$\prod_{l=1}^{L-1} F(k\Delta + cl) = F(k\Delta + c)F(k\Delta + 2c) \cdots F(k\Delta + (L-1)c)$$

gives the probability that the first RV is less than the first threshold plus c $(k\Delta + c)$ with the probability that the second RV is less than the first threshold plus $2c$, and so on. The inner integral then implies the L'th RV takes a **record** value between $k\Delta$ and $(k+1)\Delta$. However, for large c, $k\Delta + c$, $k\Delta + 2c \cdots$ will obtain a value larger than

the threshold between $k\Delta$ and $(k+1)\Delta$. Already, it seems there are some overlaps in probabilities here and the theoretical logic is wrong.



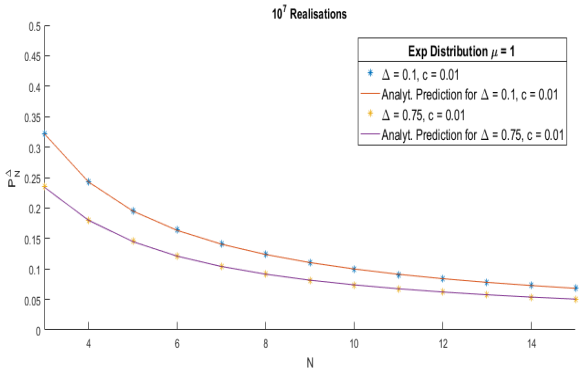
(a) $P_L^\Delta(c)$ with Uniform(-20,20) distributed RV. Simulations are averaged 10^6 times.



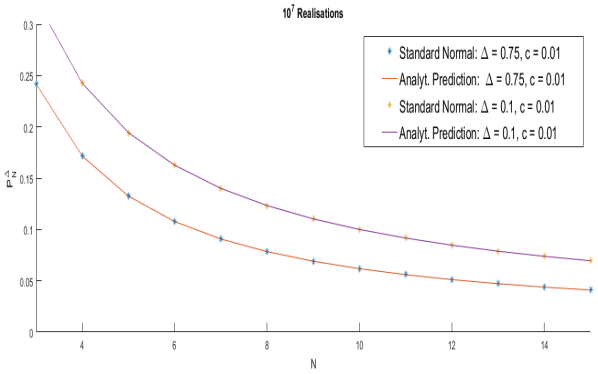
(b) $P_L^\Delta(c)$ with standard normally distributed RV. Simulations are averaged 10^6 times.

Figure 7.2: In Figure (a) , **Eq. 7.1** is unable to correctly predict the peaks and troughs exhibited in the large Δ regime for both $c = 0.5$ and $c = 1.5$. As expected, for a smaller Δ, simulations are very accurate. In Figure (b) numerical simulations show violent peaks and troughs and it is interesting that $P_L^\Delta(c)$ can oscillate to a max of around 0.25 after increasing Δ. **Eq. 7.1** exhibits no oscillations in this regime and seems to instead give the maximum probability for each Δ.

However, after numerical simulations with large c and Δ for varying distributions, the ansantz in **Eq. 7.1** matched nicely. Surprisingly, the biggest discrepancies in simulations and **Eq 7.1** have been in the large Δ regime which can be seen in **Fig. 7.2**.



(a) Numerical simulations (10^7 realisations) for exponentially distributed rounded random variables with constant linear drift. Analytical predictions give exact results in the small Δ and c regimes.



(b) Numerical simulations (10^7 realisations) for standard normally distributed rounded random variables with constant linear drift. Again, analytical predictions give excellent result in the small Δ and c regimes

Figure 7.3

7.1 Small Δ and c Regime

The following section is dedicated to deriving an expression for the rounded record rate $P_L^\Delta(c)$ in the small Δ and c regime. The expression will look like a combination of **Eq. 3.6** and **Eq. 4.7**. Analytical results are compared with numerical simulations in **Fig 7.3**. One explanation for the accuracy of **Eq 7.1** in this regime is by noting that :

$$\left\lfloor \frac{X_l + cl}{\Delta} \right\rfloor \star \Delta \approx \left\lfloor \frac{X_l}{\Delta} \right\rfloor \star \Delta + \left\lfloor \frac{cl}{\Delta} \right\rfloor \star \Delta \approx \left\lfloor \frac{X_l}{\Delta} \right\rfloor \star \Delta + c \star l = k\Delta + c * l$$

This approximation is very accurate which can be seen when generating random variables with the two formulas:

Table 7.1: Generating 2x5 exponential($\mu = 1$) random variables with $c = 0.1$ and $\Delta = 0.1$

Table 7.2: $\left\lfloor \frac{X_l + cl}{\Delta} \right\rfloor \star \Delta$

1.0000	0.9000	0.4000	1.0000	0.7000
2.1000	1.0000	1.1000	1.7000	0.6000

Table 7.3: $\left\lfloor \frac{X_l}{\Delta} \right\rfloor \star \Delta + c \star l$

1.0000	0.9000	0.4000	1.0000	0.7000
2.1000	1.0000	1.1000	1.7000	0.6000

To tackle this derivation, one might consider re-writing **Eq. 7.1** in the form:

$$P_L^{\Delta}(c) = \sum_k \phi_k(c, \Delta)$$

and perform a multi-variate Taylor expansion around the point $(c = 0, \Delta = 0)$. One example of the problem I faced is as followed: suppose we wish to perform a multi-variate $(0, 0)$ expansion on the function: $F(k\Delta + cl)$. Using a mathematical algebra software such as Mathematica and ignoring $O(\Delta^2)$, $O(c^2)$ and $O(\Delta c)$ terms we obtain the following:

$$\begin{aligned} F(k\Delta + cl) &\approx F(0) + \Delta k F'(0) + O\left(\Delta^2\right) + c\left(lF'(0) + klF''(0)\Delta + O\left(\Delta^2\right)\right) + O\left(c^2\right) \\ &\approx F(0) + \Delta k F'(0) + c\left(lF'(0)\right) \end{aligned}$$

which gives a particularly unhelpful expression to work with. To maintain simplicity, one can start with the original expression seen previously and continue as followed.

$$P_L^{\Delta}(c) \approx \sum_k \left[\int_{k\Delta}^{(k+1)\Delta} dx f(x) \right] \prod_{l=1}^{L-1} F(k\Delta + cl)$$

Next, consider the small $\Delta << 1$ regime. In [12], $\lfloor x \rfloor_{\Delta}$ was defined to be the 'largest integer multiple of Δ less than x'. Again, by replacing $k\Delta$ for this expression for $k\Delta \leq x < (k+1)\Delta$, and noting that the small Δ regime allows one to return back to the pre-discretisation case:

$$P_L^{\Delta}(c) \approx \int dx f(x) \prod_{l=1}^{N-1} F(\lfloor x \rfloor_{\Delta} + cl) \tag{7.2}$$

A first order expansion of the cumulative distribution function $F(\lfloor x \rfloor_{\Delta} + cl)$ in the small $c << 1$ regime gives:

$$F(\lfloor x \rfloor_{\Delta} + cl) \approx F(\lfloor x \rfloor_{\Delta}) + clf(\lfloor x \rfloor_{\Delta})$$

Similarly, a first order expansion around the point x allows one to write:

$$F(\lfloor x \rfloor_{\Delta}) \approx F(x) + (\lfloor x \rfloor_{\Delta} - x) f(x)$$

And finally: $f(\lfloor x \rfloor_{\Delta}) \approx f(x) + (\lfloor x \rfloor_{\Delta} - x) f'(x)$. Using the approximation $\frac{\Delta}{2} \approx (x - \lfloor x \rfloor_{\Delta})$ in [12], and putting everything together:

$$
\begin{aligned}
P_L^{\Delta}(c) &\approx \int dx f(x) \prod_{l=1}^{L-1} \left[F(\lfloor x \rfloor_{\Delta} + cl)\right] \approx \int dx f(x) \prod_{l=1}^{L-1} \left[F(\lfloor x \rfloor_{\Delta}) + clf(\lfloor x \rfloor_{\Delta})\right]. \\
&\approx \int dx f(x) \prod_{l=1}^{L-1} \left[F(x) + (\lfloor x \rfloor_{\Delta} - x)f(x) + cl\left\{f(x) + (\lfloor x \rfloor_{\Delta} - x)f'(x)\right\}\right] \\
&\approx \int dx f(x) \prod_{l=1}^{L-1} \left[F(x) - \frac{\Delta}{2}f(x) + cl\left\{f(x) - \frac{\Delta}{2}f'(x)\right\}\right]
\end{aligned} \tag{7.3}
$$

Expanding this product and ignoring $O(c^2)$ and $O(\Delta^2)$ terms gives (see **Appendix B.4** for code used in this derivation):

$$
\begin{aligned}
\prod_{l=1}^{L-1} [\cdots] \approx F(x)^{L-1} &- \frac{\Delta}{2}(L-1)f(x)F(x)^{L-2} \\
&+ c\frac{L(L-1)}{2}\left[F(x)^{L-2}f(x) - (L-1)\Delta F(x)^{L-3}f(x)^2 - \frac{\Delta}{2}f'(x)F(x)^{L-2}\right]
\end{aligned}
$$

By further ignoring $O(c * \Delta)$ terms and putting together with the whole integral in **Eq. 7.2** we arrive at the expression:

$$
\begin{aligned}
P_L^{\Delta}(c) &\approx \frac{1}{L} - \frac{\Delta}{2}(L-1)\int dx f(x)^2 F(x)^{L-2} + c\frac{L(L-1)}{2}\int dx F(x)^{L-2} f(x)^2 \\
&= \frac{1}{L} - \frac{\Delta}{2}(L-1) \star I_{2,L-2} + c\frac{L(L-1)}{2} I_{2,L-2} \\
&= \frac{1}{L} + I_{2,L-2}\left[c\frac{L(L-1)}{2} - \frac{\Delta}{2}(L-1)\right]
\end{aligned} \tag{7.4}
$$

where $I_{p,r} = \int dx f(x)^p F(x)^r$. This is noticeably a combination of the small drift **Eq 3.6** in [11] and scaling regime **Eq. 4.7** in [12]. One interesting question is, at which point in the time series does the drift dominate the scaling term and L time dependency is lost. A simple rearrangement of **Eq. 7.4** minus the time independent case $P_L^{\Delta}(0)$ gives: $c\frac{L(L-1)}{2} I_{2,L-2} > \frac{\Delta}{2}(L-1)$ and therefore:

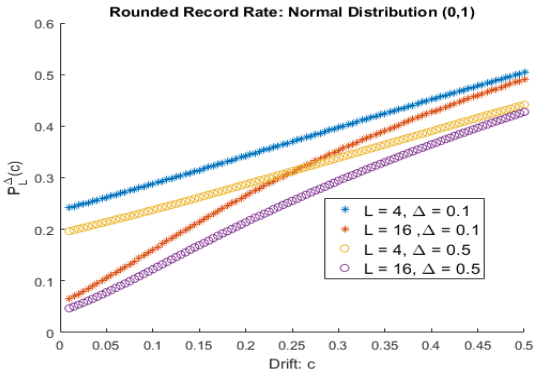
$$
L > \frac{\Delta}{c} \tag{7.5}
$$

As this is no longer depends on the underlying probability distribution, this approximation is a gross simplification. However, the logic is still correct: the higher the $\frac{\Delta}{c}$ ratio, the later in the time series do we see convergence in $P_L^{\Delta}(c)$. **Appendix C.3** highlights this.

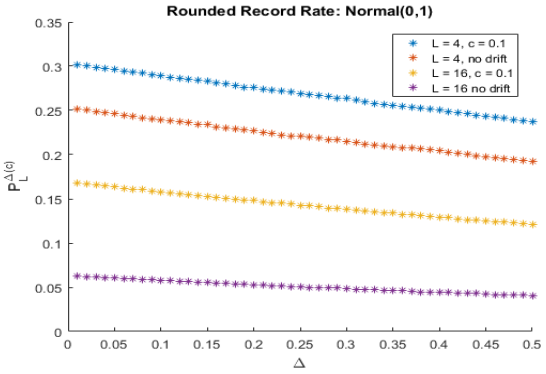
Using this approximation, one can derive an asymptotic expression for the mean record number where $L >> 1$: $\left\langle R_L^{\Delta}(c)\right\rangle = \sum_{k=1}^{L} P_L^{\Delta}(c)$.

$$
\begin{aligned}
\left\langle R_L^{\Delta}(c)\right\rangle &\approx \sum_{m=1}^{L}\left[\frac{1}{m} + I_{2,m-2}\left[c\frac{m(m-1)}{2} - \frac{\Delta}{2}(m-1)\right]\right] \\
&= \langle R_L\rangle + \sum_{m=1}^{L}\left[I_{2,m-2}\left[c\frac{m(m-1)}{2} - \frac{\Delta}{2}(m-1)\right]\right]
\end{aligned} \tag{7.6}
$$

where $\langle R_L\rangle \approx \ln L + \gamma$ with $\gamma \approx 0.577...$ in the time independent case [12]. Code to generate this can be seen in **Appendix B.5**. A summary of the scaling and drift effects on the uniform, normal and Lévy distribution can be seen in **Fig 7.3-7.5**.

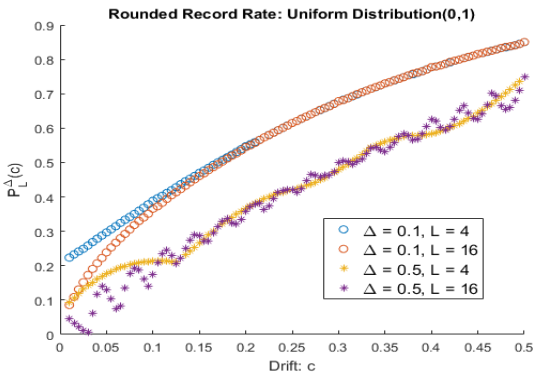


(a) P_L^{Δ} with changing c and fixed Δ's

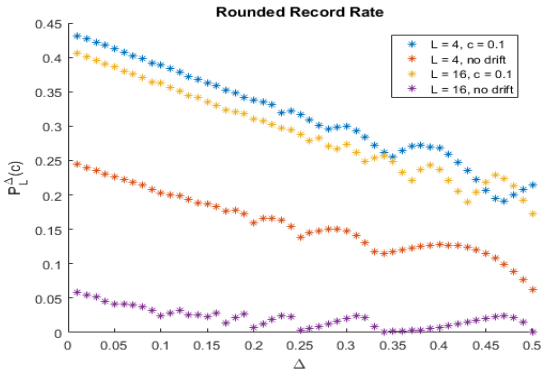


(b) P_L^{Δ} with changing Δ and comparison between drift ($c = 0.1$) and no drift.

Figure 7.4: Comparison of the effect additive drift and scaling has on the standard normal distribution. In Figure (a) $P_L^{\Delta}(c)$ is increasing with drift c. Where the drift term is small $c << 0.1$ the difference between $L = 4$ and $L = 16$ is large for both $\Delta = 0.1$ and $\Delta = 0.5$ which implies that P_L^{Δ} is L-time-*dependent* in this reason. As c overtakes the fixed rounding value, $P_L^{\Delta}(c)$ for $L = 16$ converges to $L = 4$ which implies it is time-*independent* in this region. In Figure (b), the distance $P_L^{\Delta}(c = 0.1) - P_L^{\Delta}(c = 0)$ increases for larger L. Again, we see no convergence in L where $\Delta > c$. The rounded record rate decays at a much slower rate compared to the fast increases in probability seen in Figure(a) which is to be expected: logically adding $c * L$ to the L'th RV is sure to have a stronger effect than rounding to a multiple integer of Δ.

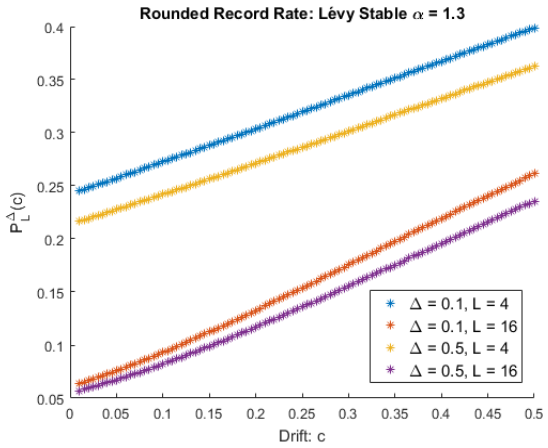


(a) P_L^{Δ} with changing c and fixed Δ's

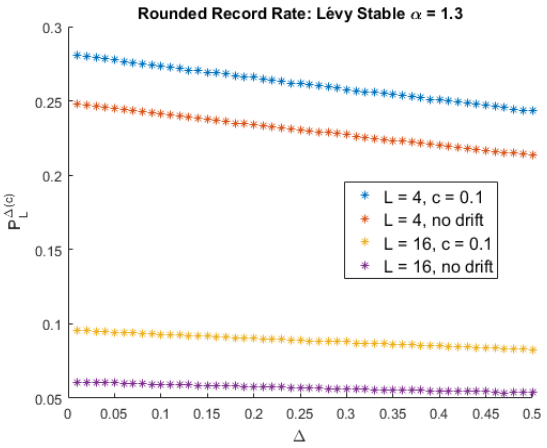


(b) P_L^{Δ} with changing Δ and comparison between drift ($c = 0.1$) and no drift.

Figure 7.5: Comparison of the effect additive drift and scaling has on the standard uniform distribution. In Figure (a) $P_L^{\Delta}(c)$ is generally increasing with drift c. For $\Delta = 0.1$, the cases where $L = 4$ and $L = 16$, converge rapidly. Interestingly, as both Δ and L increase one starts to see regular peaks and troughs. Logically, with an increasing c term and fixed RV, P_L^{Δ} shouldn't decrease. This was a particularly interesting phenomenon and one I'd like to eventually understand. In Figure(b), the distance $P_L^{\Delta}(c = 0.1) - P_L^{\Delta}(c = 0)$ rapidly increases for larger L. Overall, the rounded record rate has a greater reaction to additive drift than distributions from the Gumbel class due to the small width of its distribution.



(a) P_L^{Δ} with changing c and fixed Δ.



(b) P_L^{Δ} with changing Δ and fixed c.

Figure 7.6: Comparison of the effect additive drift and scaling has on the Lévy Stable Distribution with $\alpha = 1.3$. In Figure (a) $P_L^{\Delta}(c)$ shows different behaviour to the previous distributions from other EVS classes. For both Δ's, the cases where $L = 4$ and $L = 16$, no longer converge in the small Δ regime and are therefore *time-dependent*. In Figure(b), $P_L^{\Delta}(c)$ decays at a similar rate seen in Figure (a); something not seen in the previous two classes. From this one can conjecture that complimentary and saturating effects of c and Δ are less obvious in distributions from the Fréchet class.

7.2 Weibull Class

For the uniform distribution on $[-a, a]$ with distribution $f(x) = \frac{1}{2a}$ we arrive at the following approximation:

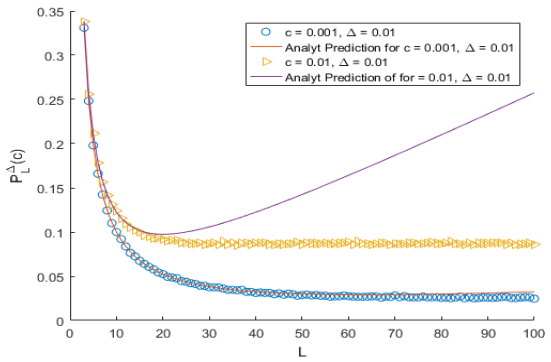
$$P_L^{\Delta} \approx \frac{1}{L}\left[1 - \frac{\Delta}{4a}L + \frac{c}{4a}L^2\right] \tag{7.7}$$

Proof. A simple use of $I_{2,L-2} = \frac{1}{2(L-1)}$ gives us our expression. □

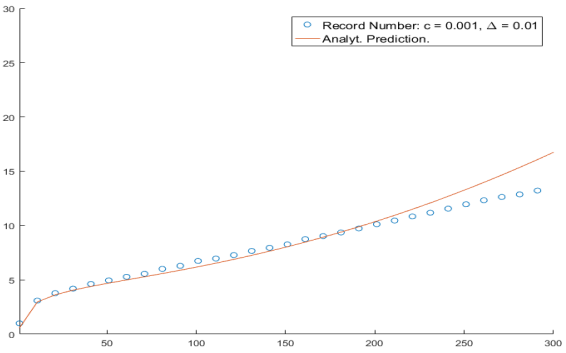
Clearly this is not a particularly useful approximation as for large L the c term dominates and quickly diverges to a large positive number. This is because the expansion in **Eq. 7.4** is only 'valid if f(x) is slowing varying between $k\Delta$ and $k\Delta + cl$ [9] which is not the case for the uniform distribution on $[-a, a]$. Similarly, the mean record number can be expressed as:

$$\begin{aligned}\left\langle R_L^{\Delta}(c)\right\rangle &\approx \ln(L) + \gamma + \sum_{m=1}^{L}\left[c\frac{m}{4} - \frac{\Delta}{4}\right] \\ &= \ln(L) + \gamma + c\frac{L(L-1)}{8} - \Delta\frac{L}{4} \\ &= \ln(L) + \gamma + L\left[c\frac{L-1}{8} - \frac{\Delta}{4}\right]\end{aligned} \tag{7.8}$$

Comparison of **Eq. 7.7** and **Eq. 7.8** with numerical results can be seen in **Fig. 7.7**.



(a) $P_L^{\Delta}(c)$ with different values of L and two different cases of fixed c and Δ. Monte carlo simulations (circle points) have been averaged 10^5 times.



(b) $\langle R_L^{\Delta}(c) \rangle$ with different values of L and one different case of fixed c and Δ. Monte Carlo simulations (circle points) have been averaged 10^5 times.

Figure 7.7: In **Fig. (a)** $P_L^{\Delta}(c)$ uses an underlying uniform distribution on the width $[-a, a]$ with $a = 1$. For the first case where $\Delta > c$, the approximation matches very accurately with numerical simulations. However, for larger L, the expression becomes very inaccurate due to the sharp divergence of our drift term. This can be seen around $L = 100$ where the analytical prediction begins to increase above simulations. Furthermore, for certain fixed values of c and Δ, $P_L^{\Delta}(c)$ becomes independent in time L. This can be seen in the flattening out of simulation plots. It seems there exists an $\hat{L}$ where the scaling effect is cancelled out completely by drift. In **Fig. (b)**, the same underlying distribution is used. The analytical expression derived in **Eq. 7.8** matches accurately with numerical simulations run for $c = 0.001$, $\Delta = 0.01$ up to $L = 200$ where it then diverges due to the same problem mentioned above.

7.3 Gumbel Class

For the standard Gaussian distribution $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ one can use **Eq. 3.16**: $I_{2,L-2} \approx \frac{1}{N^2} \frac{4\sqrt{\pi}}{e^2} \sqrt{\ln\left(\frac{N^2}{8\pi}\right)}$ to give:

$$\begin{aligned} P_L^{\Delta}(c) &\approx \frac{1}{L} + \left[c\frac{L(L-1)}{2} - \frac{\Delta}{2}(L-1) \right] \frac{1}{L^2} \frac{4\sqrt{\pi}}{e^2} \sqrt{\ln\left(\frac{L^2}{8\pi}\right)} \\ &\approx \frac{1}{L} + \left[\frac{c}{2} - \frac{\Delta}{2L} \right] \frac{4\sqrt{\pi}}{e^2} \sqrt{\ln\left(\frac{L^2}{8\pi}\right)} \end{aligned} \tag{7.9}$$

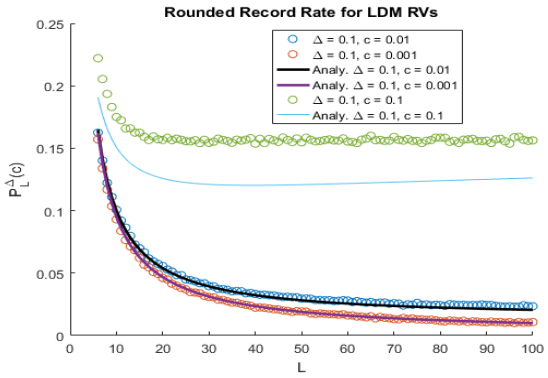
Similarly, the mean record number can be expressed as:

$$\begin{aligned} \langle R_L^{\Delta}(c) \rangle &\approx \ln(L) + \gamma + \sum_{m=1}^{L} \left[\frac{1}{m^2} \frac{4\sqrt{\pi}}{e^2} \sqrt{\ln\left(\frac{m^2}{8\pi}\right)} \left[c\frac{m(m-1)}{2} - \frac{\Delta}{2}(m-1) \right] \right] \\ &= \ln(L) + \gamma + \frac{2\sqrt{\pi}}{e^2} \sum_{m=1}^{L} \left[\sqrt{\ln\left(\frac{m^2}{8\pi}\right)} \left[c - \frac{\Delta}{m} \right] \right] \end{aligned} \tag{7.10}$$

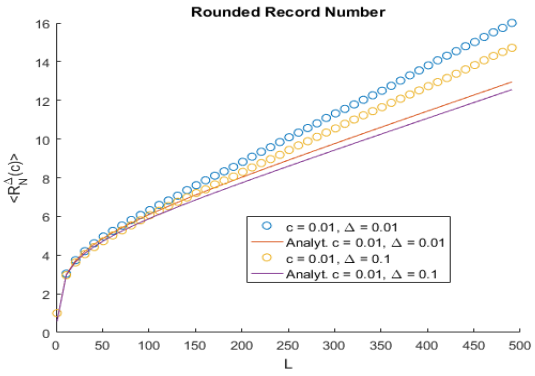
Comparison of **Eq. 7.9** and **Eq. 7.10** with numerical results can be seen in **Fig. 7.8** .

7.4 Fréchet Class

Again a subset of of the Fréchet class we consider is the Pareto distribution: $f(x) = \mu x^{-1-\mu}$ with $F(x) = (1 - x^{-\mu})$, $\mu > 1$ and $x \geq 1$. In [9] it was shown that $I_{2,L-2} = \mu \frac{\Gamma(L-1)\Gamma\left(2+\frac{1}{\mu}\right)}{\Gamma\left(L+\frac{1}{\mu}+1\right)}$. All together this gives a rounded record rate:

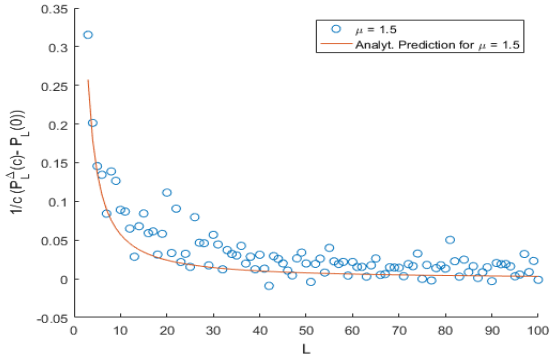


(a) P_L^{Δ} with changing RV L for 10^6 realisations.

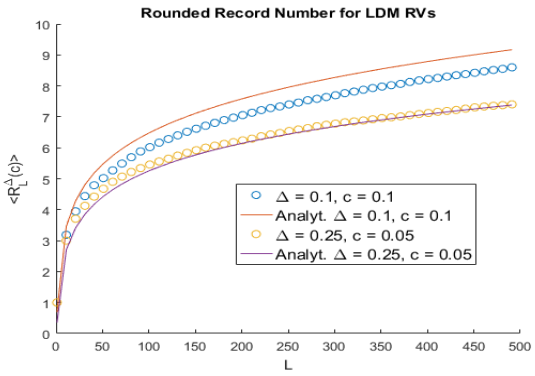


(b) $\left\langle R_L^{\Delta} \right\rangle$ with changing RV L for 10^6 realisations.

Figure 7.8: Monte Carlo simulations for the Linear Drift Model in the presence of rounding off effects using a **standard normal** distribution. Figure (a) shows $P_L^{\Delta}(c)$ in the small regime for different Δ and c. The circles represent simulations, with the lines the analytical results. For $\Delta > c$ one sees very accurate predictions, but as the gap closes the analytical prediction underestimates $P_L^{\Delta}(c)$. In Figure (b) the $\left\langle R_L^{\Delta}(c) \right\rangle$ (code in Appendix B.4) results match very accurately for $L \leq 100$ but diverge for larger L due to the number of approximations made in calculating $I_{2,L-2}$.



(a) P_L^{Δ} with $\mu = 1.5$ and fixed scaling $\Delta = 0.0001$ and $c = 0.01$



(b) $\left\langle R_L^{\Delta}(c) \right\rangle$ with $\mu = 1.5$.

Figure 7.9: Monte Carlo simulations for the Linear Drift Model in the presence of rounding off effects using a **Fréchet** class power law distribution: $f(x) = \mu x^{-\mu-1}$. Figure shows the difference between $P_L^{\Delta}(c)$ in the small regime $c = 0.01$ and the time and rounding independent case $P_L^{\Delta} = \frac{1}{L}$ with $c = 0$. This is then normalised by $\frac{1}{c}$ to make the scale more apparent. The circles represent the average over 10^6 simulations, with the lines the analytical results. For μ near to 1, results match very accurately but diverge for larger mean μ. However, for larger N the discrepancy between simulation and analytical results reduces. In Figure (b) the rounded record number $\left\langle R_L^{\Delta}(c) \right\rangle$ is given. For $\Delta > c$, the analytical formula in **Eq. 7.12** matches very accurately but as we can see for $\Delta = 0.1$, $c = 0.1$ the prediction overshoots the simulations. As we have seen, distributions in the Fréchet class are more immune to drift and rounding effects which is why the rounded record number increases at a slower rate than distributions from the previous two classes.

$$
\begin{aligned}
P_L^{\Delta}(c) &\approx \frac{1}{L} + \left[c\frac{L(L-1)}{2} - \frac{\Delta}{2}(L-1) \right] \mu \frac{\Gamma(L-1)\Gamma\left(2+\frac{1}{\mu}\right)}{\Gamma\left(L+\frac{1}{\mu}+1\right)} \\
&\approx \frac{1}{L} + \mu\Gamma\left(2+\frac{1}{\mu}\right) L^{-\frac{1}{\mu}} \left[\frac{c}{2} - \frac{\Delta}{2} L^{-1} \right]
\end{aligned}
\tag{7.11}
$$

which tends to the record rate P_N in the large L limit. For the record number:

$$\begin{aligned}\left\langle R_L^{\Delta}(c)\right\rangle &\approx \ln(L)+\gamma+\mu\Gamma\left(2+\frac{1}{\mu}\right)\sum_{m=1}^{L} m^{-\frac{1}{\mu}}\left[\frac{c}{2}-\frac{\Delta}{2}m^{-1}\right] \\ &= \ln(L)+\gamma+\frac{\mu}{2}\Gamma\left(2+\frac{1}{\mu}\right)\left(cH_L^{\left(\frac{1}{\mu}\right)}-\Delta H_L^{\left(1+\frac{1}{\mu}\right)}\right)\end{aligned} \tag{7.12}$$

Comparison to the numerical results with extended comments can be seen in **Fig. 7.9.**

8. Discussion and Conclusion

Overall this research project consists of two elements. I have extended the theoretical studies on rounding down independent and identically distributed random variables in [12] by defining a rounded record rate P_L^{Δ} where random variables are not identically distributed; a simple extension and nothing ground breaking. Its correctness was confirmed using extensive numerical simulations in Matlab with random variables drawn from a distribution with changing shape parameters. The next logical progression was to extend the theoretical work on broadening (sharpening) distribution in [6] to this. Using the classic Laplace method I was able to find a fairly accurate approximation: P_L^{Δ} for exponential random variables with sharpening distribution. With more time, I would like to find a useful approximation using an underlying Pareto distribution and compare the rate of saturation between it and the exponential approximation.

Furthermore, my research showed that when rounding down i.i.d random variables and thus creating ties, independence between adjacent random variables is lost. Mathematically: $P_{L,L-1}^{\Delta} \neq P_L^{\Delta} P_{L-1}^{\Delta}$ which inspired the derivation of a joint rounded record rate in Chapter 6. Simulations suggested that the difference between the joint record rate and multiplication of its marginals is dependent on Δ and L. The greater the rounding effect, the larger this difference and because of this I attempted to quantify a correlation term $A_{L,L-1}^{\Delta}$ and plugged in two example distributions. Analytical predictions here were successful. A future development on this is to isolate a small Δ regime and find a general approximation for $P_{L,L-1}^{\Delta}$ with i.i.d $P_{L,L-1}^{\Delta} = \frac{1}{L(L-1)}$ component and a negative term that depends on Δ and the underlying probability distribution used. By taking the ratio between this and its marginals approximation: $P_L^{\Delta} P_{L-1}^{\Delta}$ would give a general formula for $A_{L,L-1}^{\Delta}$.

The second element of the project was research into the interconnectedness between the opposite effects additive drift c and Δ have on a time series of random variables. I had great difficulty deriving an *exact* expression for the rounded record rate in the presence of drift: $P_L^{\Delta}(c)$, so instead, I isolated a small c and Δ as the derived formula was extremely accurate in this regime. Following the same methods and logic as previous authors, my aim was to find an expression that consisted of the i.i.d record rate: $P_L = \frac{1}{L}$ plus positive c and negative Δ terms. Interestingly, the expression derived developed into a combination of the small c-positive term in [9] and the negative Δ term in [12]. It also correctly predicted that additive drift had a stronger impact on the rounded record rate. Numerical vs analytical predictions were tested for distributions belonging to the three different EVS classes with varying success.

Where noticeable differences existed, this was caused by a chain of approximations; for example the formula depended on the integral $I_{2,L-2}$ which was derived for the well-known distributions in previous literature [9]. This integral was particularly hard to calculate for a Gaussian distribution, so many strong corrections were made in the process. One future development on the above is a multivariate Taylor expansion of the general expression for $P_L^{\Delta}(c)$; something more sophisticated than my approach. Finally, a simple application of $\left\langle R_L^{\Delta}(c) \right\rangle = \sum_{j=1}^{L} P_j^{\Delta}(c)$ allowed one to find approximations for the progression of record numbers in a time series, again with varying success. Unfortunately, I ran out of time to derive a correlation function $\Pi_{L,L-1}^{\Delta}(c) = \frac{P_{L,L-1}^{\Delta}(c)}{P_L^{\Delta}(c) P_{L-1}^{\Delta}(c)}$. This would have consisted of first finding a joint rounded record rate in the presence of additive drift and I would be interested to see how it compares across distributions.

9. Self-Assessment

I will start with what I should have done differently to help myself during the project. Firstly, in hindsight, I would have spent less time re-deriving some of the approximations in the core papers: [11] and [12] which occasionally lacked clarity. This would have freed up more time for actual research. Secondly, I have learnt that an analytical derivation needs to be checked extensively with simulations. For example, it was only till late into the project did I notice that the formula in **Eq. 7.1** was not correct in the very large Δ limit... a big set back. Finally, although this was out of my control due to time restrictions, illustrating the real life applicability of ties caused by rounding down linear drift random variables would have added an extra dimension to the project.

To the positives, having never studied rare events or extreme value theory, I'm pleasantly surprised with the outcome of this project. I have worked very hard throughout this summer and have enjoyed understanding and analysing some of the puzzling results that numerical simulations have thrown out. Given that there is a surprisingly small amount of literature concerning records in a non-i.i.d framework this project represents an addition to the field. For this reason, I believe it is worthy of a strong merit or higher.

Appendices

A. Analytical Derivations

A.1 Gumbel Distributed Random Variables: Stochastic Independence

The proof of this is contained in [11]. Take a RV with an underlying Gumbel distribution: $f(x) = \exp(-e^{-x} - x)$ and $F(x) = \exp(-e^{-x})$. Properties of the distribution give: $F(x+a) = \exp(-e^{-x-a}) = F(x)^{e^{-a}}$. Now: $P_N(c) = \int dx f(x) \prod_{j=1}^{N-1} F(x)$. A substitution of $F(x) = u$ and $a = cl$ gives:

$$P_N(c) = \int_0^1 du\, u^{\sum_{j=1}^{N-1} e^{-cl}} = \frac{1}{\sum_{j=1}^{N-1} e^{-cl}}$$

Similarly, for the joint record rate:

$$\begin{aligned} P_{N,N-1}(c) &= \int_0^1 du \int_0^{u^{e^{-c}}} dy\, y^{\sum_{j=1}^{N-2} e^{-cl}} = \frac{1}{\sum_{j=1}^{N-2} e^{-cl}} \int_0^1 du\, u^{\sum_{j=1}^{N-1}} \\ &= \left[\frac{1}{\sum_{j=1}^{N-1} e^{-cl}}\right]\left[\frac{1}{\sum_{j=1}^{N} e^{-cl}}\right] = P_{N-1}(c) P_N(c) \end{aligned}$$

Which implies stochastic independence for LDM Gumbel distributed random variables.

A.2 $I(N)$ solved for Gaussian Distribution

In the papers [9] and [11] the derivation for approximating the integral $I(N)$ lacks clarity. For this reason, I've provided a step by step proof below.

Proof. I first start by writing $I(N-2) = \frac{1}{2\pi}\frac{1}{(2\pi)^{N-2}} \int_{-\infty}^{\infty} e^{-x^2} \left[\int_{-\infty}^{x} dx e^{-\frac{x^2}{2}}\right]^{N-2}$. This can be evaluated for large N be using a saddle point approximation in the format $I(N-2) = \text{const} \int_{-\infty}^{\infty} dx e^{g(x)}$ where $g(x) = -x^2 + (N-2)\ln\left[\int_{-\infty}^{x} dx e^{-\frac{x^2}{2}}\right]$. Saddle point maximum then satisfies: $g'(x) = 0$, giving:

$$-2\hat{x} + N\frac{e^{\frac{-\hat{x}^2}{2}}}{\int_{-\infty}^{x} dx e^{-\frac{x^2}{2}}} = 0 \tag{A.1}$$

The denominator of this fraction is $\approx \sqrt{2\pi}$ [9] in the large N and $\hat{x}$ limit leaving the following expression:

$$\hat{x} = \frac{N-2}{\sqrt{8\pi}} e^{\frac{-\hat{x}^2}{2}} \tag{A.2}$$

Squaring both sides and rearranging gives: $e^{\hat{x}^2}\hat{x}^2 = \frac{N^2}{8\pi}$ and one finds the saddle point maximum is in the form of Lambert W-function with $W(z)e^{W(z)} = z$ giving:

$$\hat{x} = \sqrt{W\left[\frac{(N-2)^2}{8\pi}\right]} \tag{A.3}$$

To complete the expression for $I(N-2)$, one must plug $\hat{x}$ into $g(\hat{x})$ and $g''(\hat{x})$ and make suitable approximations. To make more sense of this, note that, after $\sqrt{2\pi}$ substitution: $g(x) = -x^2 + N \star \ln(\sqrt{2\pi}) - 2ln(\sqrt{2\pi})$. In the large N limit, the first ln term disappears and using $ln(\sqrt{2\pi}) \approx 1$ one arrives at $g(x) = -\hat{x}^2 - 2$. Furthermore, the second derivative of $g(x)$ is, using the substitution **Eq. A.2**:

$$\begin{aligned} g''(\hat{x}) &= -2 + (N-2)\frac{1}{\sqrt{2\pi}}(-\hat{x})e^{\frac{-\hat{x}^2}{2}} \\ &\approx -2 + (N-2)\frac{1}{\sqrt{2\pi}}(-\hat{x})\frac{\sqrt{8\pi}\hat{x}}{(N-2)} \\ &\approx -2(1+\hat{x}^2) \end{aligned} \tag{A.4}$$

In full, our expression in the large N limit becomes the following. Note the use of back substitution of equtions **Eq. A.2** and **Eq. A.3**.

$$I(N-2) \approx \frac{1}{2\pi}\sqrt{\frac{-2\pi}{g''(\hat{x})}}e^{g(\hat{x})} = \frac{1}{2\pi}\sqrt{\frac{\pi}{1+\hat{x}^2}}e^{-\hat{x}^2-2} = \frac{4\sqrt{\pi}}{(N-2)^2e^2}\frac{\hat{x}^2}{\sqrt{1+\hat{x}^2}} \tag{A.5}$$

By making crude assumptions that $\frac{\hat{x}^2}{\sqrt{1+\hat{x}^2}} = \hat{x}$, plugging in this in terms of the Lambert-W function and using the expansion [9]: $W(N) \approx \ln(N) - \ln(\ln(N)) \approx \ln(N)$ in the large N limit:

$$I(N) \approx \frac{4\sqrt{\pi}}{N^2e^2}\sqrt{ln\frac{N^2}{8\pi}} \tag{A.6}$$

□

A.3 Joint Record Rate in the Small c Regime

Given the joint record rate $p_{N,N-1}(c)$ in **Eq. 3.5** one can approximate the product: $\prod_{l=1}^{N-2}[F(x) + clf(x)] = F(x)^{N-2} + c\frac{(N-1)(N-2)}{2}f(x)F(x)^{N-2}$. Moreover, the inner integral $\int^{y+c} dxf(x)$ is approximately $F(y) + cf(y)$ after a first order series expansion.

$$\begin{aligned} p_{N,N-1}(c) &\approx \int dyf(y)[F(y)+cf(y)]\left[c\frac{(N-1)(N-2)}{2}F(x)^{N-3}f(x) + F(x)^{N-2}\right] \\ &= \int dyf(y)\int^y dxf(x)F(x)^{N-2} + c\int dyf(y)^2F^{N-2}(x) \\ &+ c\frac{(N-1)(N-2)}{2}\int dyf(y)\int^y dxf(x) \\ &= \frac{1}{N(N-1)} - c\frac{(N-1)(N-2)-2}{2}I(N-2) + c\frac{(N-1)(N-2)}{2}I(N-3) \end{aligned} \tag{A.7}$$

To show that the first term of the RHS equals $\frac{1}{N(N-1)}$ we use the substitution $u = F(y)$, $du = f(y)dy$ and the limits $F(\infty) = 1$ and $F(-\infty) = 0$ [9] to obtain:

$$p_{N,N-1}(c=0) = \int_0^1 du\int_0^u d\hat{u}\,\hat{u}^{N-2} = \frac{1}{N-1}\frac{1}{N}$$

B. Matlab and Mathematica Code

B.1 $l_{N,N-1}$: Correlation Between Consecutive Records in the Presence of Constant Linear Drift.

```
1  function corrPlot_pnts = corrPlot(N,L)
2      i = 0; corrPlot_pnts = zeros(1,30);
3      x = exprnd(1,N,L); % Generate N by L matrix of Exp RV.
4
5      for c = 0:0.1:3     % For each c, output a correlation .
6          i = i + 1;
7          corrPlot_pnts(i) = corr_Numerical(c,N,L,x);
8      end
9      % Code to output graph.
10     c = 0:0.1:3;
11     xlabel('c (linear drift)'); ylabel('l_{N,N-1}'); scatter(c, corrPlot_pnts,'*');
12 end
13
14 function corr = corr_Numerical(c,N,L,x)
15 % JRC: JointRecordCounter for consecutive RV. LthRC: Record Counter for the Nth RV.
16 % L_1thRC: Record Counter for the N-1th RV.
17     JRC = 0; LthRC = 0; L_1thRC = 0;
18     H = x + c.*repmat(1:L,N,1); % Apply linear drift
19     % Identify maximum values for comparison.
20     B = max(H(1:N,1:L-1),[],2); C = max(H(1:N,1:L-2),[],2);
21     j = 0;
22         while j < N     % N realisations
23         j = j + 1;
24             if H(j,L) > B(j)
25              LthRC = LthRC + 1;
26                  if H(j,L-1) > C(j)
27                    JRC = JRC + 1;
28                  end
29             end
30             if H(j,L-1) > C(j)
31              L_1thRC = L_1thRC + 1;
32             end
33         end
34 % Joint record rate divided by marginals.
35 corr = N * (JRC / (LthRC * L_1thRC));
36 end
```

B.1.1 Algorithm Summarised

1. The function corrPlot(N,L) takes 2 inputs N and L, where N is the number of simulations to run and L is the corresponding random variable in the time series. For example: corrPlot(100000,8) would plot the correlation probabilities between the 8th and 7th random variables for $c = 0$ to 3 in increments of 0.1, where simulations have been averaged over 10^5 realisations.

2. After initialising variables and array sizes in line 2, L random variables $\{Y_l\}$ are generated with exponentially

distributed ($\lambda = 1$) i.i.d part and additive linear drift c. This is repeated N times and placed in a N by L matrix x.

3. The code is optimised by utilising Matlab's 'max' function which, when applied to a matrix, pulls out the maximum value from each column into a single 1 by L array. Random variable L is a record if it is greater than the maximum of all previous $L-1$ records, thus by taking the 'max' of the transpose of matrix H containing columns 1 to $L-1$, allows for a very efficient comparison between the two. This can be seen in column vectors B and C defined in line 5.

4. Three counters '$LthRC$', 'L_1thRC', 'JRC' are updated after iterating through each element of the column matrices B and C. This is contained in the while loop. '$LthRC$' counter is updated to '$LthRC + 1$' if the L'th RV is the maximum of all previous $L-1$ random variables. Similarly: '$PrevLthRC$' counter is updated to '$PrevLthRC + 1$ if the $L-1$'th RV is the maximum of all $L-2$ previous random variables. If the L and $L-1$ random variables were **both** records then the 'JRC' is incremented.

5. At the end of the N simulations I approximate p_N, p_{N-1} and $p_{N,N-1}$ (Eq. 3.5) by dividing each of the three counters by the number of simulations N. This allows for the correlation function $l_{N,N-1}(c)$ to be evaluated.

6. Next by iterating the above function for values of $c = 0$ to $c = 3$ in incremental steps of 0.1 and plotting the correlation function with c on the x axis allows us to compare the effect of choosing different distributions.

B.2 mainTerm: Extracting Dominant Terms in Asymptotic Limit. [16]

Listing B.1: mainTerm Function

```
1  mainTerm[expr_, x_] := Module[
2  {approxpoly, prec = 1, j = 0, expon, coef},
3  approxpoly = Normal[Series[expr, {x, Infinity , prec }]];
4  While[approxpoly === 0 && j <= 5,
5   j++;
6   prec = 2*prec + j;
7   approxpoly = Normal[Series[expr, {x, Infinity , prec }]];
8   ];
9  approxpoly = PowerExpand@approxpoly;
10 expon = Exponent[approxpoly, x];
11 coef = Coefficient [approxpoly, x, expon];
12 coef*x^expon
13 ]
```

In pseudocode the *mainTerm* function does the following.

1. Generate a power series at infinity to obtain a polynomial.

2. Find largest exponent.

3. Find corresponding coefficient.

B.3 Rounded Record Rate: P_N^{Δ}

```
function RecordProb = StrongRecordRate(N,L,Delta)
    x = normrnd(0,1,N,L);
    H = floor(x ./ Delta)*Delta;
    RecordCounter = 0; j = 0;
    B = max(H(1:N,1:L-1),[],2);
    while j < N
        j = j + 1;
        if H(j,L) > B(j)
            RecordCounter = RecordCounter + 1;
        end
    end
    RecordProb = (RecordCounter / N);
end
```

The above code returns the rounded record rate P_L^{Δ} for i.i.d random variables. The same logic applies from **Appendix B.1** but now each entry in the time series is rounded to some integer multiple of Δ. This can be seen in line 3 through use of the built in floor function.

B.4 Small c and Δ Expansion Approximation

Listing B.2: Small c and Δ Expansion

```
TermToExpand = A - Delta/2*B + c*I (B - Delta/2*G);
P1 = Expand[Product[TermToExpand, {I, 1}]];
P2 = Expand[Product[TermToExpand, {I, 2}]];
P3 = Expand[Product[TermToExpand, {I, 3}]];
P4 = Expand[Product[TermToExpand, {I, 4}]];
NewP1 = P1 /. Delta^b_ /; b >= 2 -> 0;
NewNewP1 = NewP1 /. c^b_ /; b >= 2 -> 0;
NewP2 = P2 /. Delta^b_ /; b >= 2 -> 0;
NewNewP2 = NewP2 /. c^b_ /; b >= 2 -> 0;
NewP3 = P3 /. Delta^b_ /; b >= 2 -> 0;
NewNewP3 = NewP3 /. c^b_ /; b >= 2 -> 0;
NewP4 = P4 /. Delta^b_ /; b >= 2 -> 0;
NewNewP4 = NewP4 /. c^b_ /; b >= 2 -> 0;
FinalP1 = Expand[FullSimplify[NewNewP1]]
FinalP2 = Expand[FullSimplify[NewNewP2]]
FinalP3 = Expand[FullSimplify[NewNewP3]]
FinalP4 = Expand[FullSimplify[NewNewP4]]

Output:

A   + B c         - (B * Delta])/2  - (c G * Delta)/2
A^2 + 3 A B c     - A B Delta       - 3/2 B^2 c Delta    - 3/2 A c G Delta
A^3 + 6 A^2 B c - 3/2 A^2 B Delta - 6 A B^2 c Delta    - 3 A^2 c G Delta
A^4 + 10 A^3 B c - 2 A^3 B Delta - 15 A^2 B^2 c Delta - 5 A^3 c G Delta
```

In pseudocode the function does the following.

1. A represents $F(x)$, B represents $f(x)$ and G represents $f'(x)$ in the final line of **Eq. 7.3**.
2. Take the product of TermToExpand for $L = 2, 3, 4, 5$.
3. For each case: NewPX sends all $O(\Delta^2)$ terms to 0, NewNewPX sends all $O(c^2)$ to 0.
4. Expand and simplify each case.
5. Identify a general pattern across all cases. For example: the first term has general case A^{L-1}.

Note $O(c\Delta)$ terms have been left in here. They are excluded later in the derivation.

B.5 $\left\langle R_L^{\Delta}(c)\right\rangle$: record number for the LDM RV

The following computes the record number $\left\langle R_M^{\Delta}(c)\right\rangle$ for the Mth RV in the time series.

```
function RNPlotFinal = RecordNumberPlot(M,N,c,Delta)
i = 0; RNPlot = zeros(M,(N/10)); k = 0;
x = exprnd(1,M,N) + c.*(repmat(1:N,M,1));
H = floor(x./ Delta).*Delta;
while k < M
    k = k + 1;
    for L = 1:10:N
        i = i + 1;
        RNPlot(k,i) = RecordNumber(L,H,k);
    end
    i = 0;
end
RNPlotFinal = mean(RNPlot);
L = 1:10:N;
scatter(L,RNPlotFinal);
hold on;
xlabel('L'); ylabel('<R_L^{\Delta}(c)>');title('Rounded Record Number for LDM RVs');
end

function RN = RecordNumber(L,H,k)
RecordCounter = 1; j = 1;
RecordIndex = 1;
while j < L
    j = j + 1;
    if H(k,j) > H(k,RecordIndex)
        RecordCounter = RecordCounter + 1;
        RecordIndex = j;
    end
end
RN = RecordCounter;
end
```

C. Additional Figures

C.1 $l_{N,N-1}(c)$: Simulations for a Uniform Distribution on [-1,1]

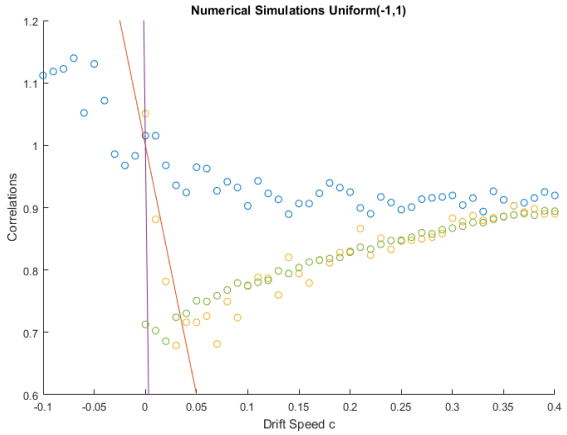


Figure C.1.1: Numerical simulations for the correlation function on the Uniform distribution for $N = 4, 16, 64$ (blue, red and green circles respectively). The analytical predictions (thick plotted lines) are not useful as $l_{N,N-1}(c)$ decays to zero rapidly with increasing N.

C.2 $P_N^{\Delta}(c)$: Rounded Record Rate

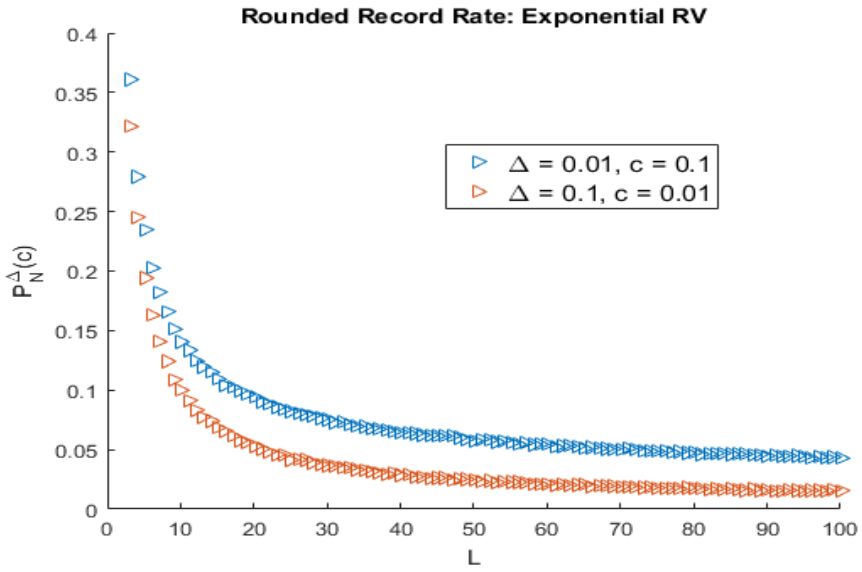


Figure C.2.1: Rounded record rate with drift $P_N^{\Delta}(c)$ using an exponential distribution and different values in the time series. These are averaged over 10^6 realisations. For $\Delta = 0.1, c = 0.01$ the ratio $\frac{\Delta}{c}$ says time-dependency occurs for $L > 10$, and for $\Delta = 0.1, c = 0.01$ the ratio $\frac{\Delta}{c}$ says time-dependency occurs for $L > 1$.

Bibliography

[1] Fisher R A and Tippett L H C. “Limiting forms of the frequency distribution of the largest and smallest member of a sample”. In: *Proc. Cambridge Phil. Soc.* 24 180-190 (1928).

[2] Gnedenko B V. “Sur la distribution limite du terme maximum dune serie aleatoire”. In: *Annals of Mathematics* 44 423-453 (1943).

[3] Ballerini R and Resnick S. “Records from improving populations”. In: *Journal of Applied Probability* 22, 487 (1985).

[4] H.N. Nagaraia Barry C.Arnold N. Balakrishnan. *Records*. Wiley Series in Probability and Statistics. 1998. ISBN: 0471081086.

[5] B. Gordon and R. J. McIntosh. “Some Eighth Order Mock Theta Functions.” In: *J. London Math. Soc.* 62 (2000), pp. 321–335.

[6] Joachim Krug. “Records in a changing world”. In: *Journal of Statistical Mechanics* P10013 (2007).

[7] Yang MCK. “On the Distribution of the Inter-Record Times in an Increasing Population”. In: *Journal of Applied Probability* 12, 148 (2007).

[8] S. N. Majumdar and R. M. Ziff. “Universal Record Statistics of Random Walks and Lévy Flights”. In: *Phys. Rev. Lett.* **101**,050601 (2008).

[9] J. Franke G.Wergen and J. Krug. “Records and sequences of records from random variables with a linear trend”. In: *Journal of Statistical Mechanics: Theory and Experiment* P10013 (2010).

[10] J Wergen G. Krug. “Record-breaking temperatures reveal a warming climate”. In: *Europhys.Lett* 92,30008 (2010).

[11] J. Franke G.Wergen and J. Krug. “Correlations between record events in sequences of random variables with a linear trend”. In: *Journal of Statistical Physics* **144**, 1206 (2011), pp. 1206–1222.

[12] G. Wergen et al. “Rounding effects in record statistics”. In: *Phys. Rev. Lett.* **109**,164102 (2012).

[13] G.Wergen. “Record statistics beyond the standard model - Theory and applications”. In: *PhD Thesis* (2012). DOI: `http://kups.ub.uni-koeln.de/5016/1/main.pdf`.

[14] Y. Edery et al. “Record-breaking statistics for random walks in the presence of measurement error and noise”. In: *Phys. Rev. Lett.* **110**,180602 (2013).

[15] Pierpaolo Vivo. “Large deviations of the maximum of independent and identically distributed random variables”. In: *Cond-Mat.Stat-Mech* 1507.05442 (2015).

[16] Daniel Lichtblau. *mainTerm*. URL: `http://mathematica.stackexchange.com/questions/9965/limiting-form-of-a-polynomial-expression`.

[17] Jonathan Sondow and Eric W. Weisstein. *Harmonic Number*. URL: `http://mathworld.wolfram.com/HarmonicNumber.html`.

[18] Eric W. Weisstein. *q-Pochhammer Symbol*. URL: `http://mathworld.wolfram.com/q-PochhammerSymbol.html`.

[19] Wikipedia. *Inverse Transform: Exponential Distribution*. URL: `https://en.wikipedia.org/wiki/Exponential_distribution`.

[20] Wikipedia. *Rare Events*. URL: https://en.wikipedia.org/wiki/Rare_events.