Using a peak over threshold approach with the Generalised Pareto distribution to project the breaking of world records in the near future

Summary Sheet

Extreme value theory is a branch of statistics that can be used to predict the occurrence of rare events, such as extreme flooding, large insurance losses, crashing of the stock market, or human life expectancy. Because a world record is an extreme event, we will use extreme value theory to predict the probability of breaking the world record in the near future. Almost all mathematical models aiming to predict ultimate world records are based on the development of the world record over many years and extrapolating the trend to the future. It uses only a limited number of past world records and therefore gives rather unreliable estimates and predictions. This is ***not*** the approach we are taking. Instead, we utilise as many top performances of elite athletes in the last few years as possible throughout various athletic events to predict the probability of breaking the world record in the near future. Since our mathematical model is based on a large set of data in a short up-to-date period of time, it is effective in predicting the probability of breaking a world record in the near future. Therefore, using our model minimises the affect that factors such as technological advancement, changes in training and anti-doping laws have on the probability of breaking a world record, since it is based on data from a small period of time (three years). Our mathematical model is inspired by the approach taken by Einmahl and Magnus (2008) in their paper "Records in Athletics through Extreme Value Theory.

Our method involves collecting data from the official website of the International Association of Athletics Federation (IAAF) for the three years prior to an event being held. This data is then ordered from best to worst performance, and data points below a lower threshold are removed by the *Peak over threshold approach*. This lower threshold is determined through comparing the worst performance for each of the three years and selecting the 'best' of these (Einmahl & Magnus, 2008). This data is then used to construct a probability density histogram giving the probability of an athlete's top performance being located within each score range 'bin'. The Generalised Pareto function is then fitted to the probability density histogram. This allows the trend seen in the data to be extrapolated to estimate the probability of an athlete breaking the record in the following year. We assume the worst-case scenario that if the record is broken within the following year, it will be broken at the committee's event. The binomial distribution is then used to factor in the number of athletes competing in the event.

The resulting probability is then considered to be the probability that an athlete will break the world record at the committee's event. Through analysing IAAF data for multiple events, our team found that there was a strong correlation between the calculated probability and the record being broken (a probability greater than 3%) or remaining (a probability less than 3%).

Thus, an organising committee may use our method to calculate the probability of the world record being broken at any one of their events and purchase insurance accordingly. Similarly, insurance companies may also use the model to calculate the amount of money they should charge per event in order to maximise profit.

As the model only relies on archived data, it may be tested by an organising committee or an insurance company on years where the outcome is now known, allowing the validity of the model to be confirmed for each event type before being used to a make a decision which will ultimately have an impact on the organisation's finances.

Question 1: Calculating Average Cost

$$Average\ Cost = \frac{Amount\ of\ Bonus}{Number\ of\ times}$$

$$= \frac{25\ 000}{n}$$

Where ***n*** is the number of times the event is repeated before the record is broken. From the above example, ***n = 25***, and hence:

$$Average\ Cost = \frac{25\ 000}{25}$$

$$Average\ Cost = \$1000$$

Question 2: Calculating the Cost of Insurance

The average cost that the insurance company charges the committee yearly exists to ensure that at a minimum, they fully cover the amount of money that they have to pay out if a record is broken. In this example, an average cost of $1000 is charged every year as the record is likely to be broken once every 25 years and the fee for the world record breaker is $25 000.

However, an additional cost is added to the average cost so that the insurance company can make a reasonable profit and cover all other added costs. If the insurance company is seeking to maximise profit, there is a range of variables to consider involving supply and demand. However, if the company utilises our model proposed in Q5, they can alter the additional cost such that they can lower it in years where they know the record will likely not be broken. In lowering their cost for these years, they will attract more subscribers. Yet, in using of our model they can also determine when it is likely that the world record will be broken, and as a result can choose to raise their additional cost to the point of ridiculousness that year, yet still below the $25 000 bonus. This will either dissuade customers from insuring that year, where the record will likely be broken (meaning the insurance company avoids paying out), or alternatively if a committee still chooses to insure and pay their extravagant additional costs, the insurance company minimises the otherwise huge loss that they could have potentially made charging their normal additional cost.

Question 3: Determining Whether Insurance Should Be Purchased

(a):

The most important variables that the committee must consider is the current quality of their pool and the number of athletes racing. Through the application of our model outlined in question 5, the quality of the current pool (reflected by the estimated probability of the current pool breaking the WR), can be compared to the quality of pools which did or did not break the world record. The number of athletes competing in their event can then be factored in through a binomial distribution to culminate in a projected percentage probability of the record being broken at the committee's event. Ultimately this probability can then be compared to the probability prior to previous successes and failures to assess the need to purchase insurance for any given event. Through the analysis of three separate events, outlined in question 5, we found that a percentage probability of 3% or greater indicates that the likelihood of the WR being broken is high, meaning insurance should

be purchased. If the estimated percentage probability is less than 3%, it is unlikely that the world record will be broken and insurance should not be purchased.

(b):

On years where the model from question 5 has been applied and there is a significant probability of the world record being broken (as outlined by the model's decision making scheme), insurance should be taken out. On years where the probability is insignificant, insurance should not be taken out. Based on the assumption that the model's prediction is accurate, when the world record is broken, the organising committee only has to pay for the cost of insurance. This means that the total amount payed by the organising committee is less than if they always took out insurance or always self-insured.

Question 4: Application of the Model to Various Events

An issue the committee will face is the variability in the probability of a world record being broken between certain event types. For example, the image to the left shows the progression of the Mens' 100m world record since 1988, and illustrates that the record was broke repeatedly and regularly up until 2009. The image on the right displays that the world record for long jump has been stagnant since the 1960's and has not been broken for over two decades.

9.92	1.1		Carl Lewis	United States	Seoul, South Korea	September 24, 1988	OR[note 2][2]
9.90	1.9		Leroy Burrell	United States	New York, USA	June 14, 1991	[2]
9.86	1.0		Carl Lewis	United States	Tokyo, Japan	August 25, 1991	[2]
9.85	1.2		Leroy Burrell	United States	Lausanne, Switzerland	July 6, 1994	[2]
9.84	0.7	9.835	Donovan Bailey	Canada	Atlanta, USA	July 27, 1996	OR[2][8]
9.79	0.1		Maurice Greene	United States	Athens, Greece	June 16, 1999	[2]
9.78	2.0		Tim Montgomery	United States	Paris, France	September 14, 2002	[9][note 4]
9.77	1.6	9.768	Asafa Powell	Jamaica	Athens, Greece	June 14, 2005	[2]
	1.7	9.766	Justin Gatlin	United States	Doha, Qatar	May 12, 2006	[9][10][note 5]
	1.5	9.763	Asafa Powell	Jamaica	Gateshead, England	June 11, 2006	[7]
	1.0	9.762			Zürich, Switzerland	August 18, 2006	[7]
9.74	1.7		Asafa Powell	Jamaica	Rieti, Italy	September 9, 2007	[1]
9.72	1.7				New York, USA	May 31, 2008	[2]
9.69	0.0	9.683	Usain Bolt	Jamaica	Beijing, China	August 16, 2008	OR[2]
9.58	0.9	9.572			Berlin, Germany	August 16, 2009	[1][11][12]

8.28 m (27 ft 2 in)	1.2	Ralph Boston (USA)	Moscow, Soviet Union	16 July 1961[1]
8.31 m (27 ft 3 1/4 in)	-0.1	Igor Ter-Ovanesyan (URS)	Yerevan, Soviet Union	10 June 1962[1]
8.31 m (27 ft 3 1/4 in)	0.0	Ralph Boston (USA)	Kingston, Jamaica	15 August 1964[1]
8.34 m (27 ft 4 1/4 in)	1.0	Ralph Boston (USA)	Los Angeles, United States	12 September 1964[1]
8.35 m (27 ft 4 3/4 in)	0.0	Ralph Boston (USA)	Modesto, United States	29 May 1965[1]
8.35 m (27 ft 4 3/4 in) **at Altitude**	0.0	Igor Ter-Ovanesyan (URS)	Mexico City, Mexico	19 October 1967[1]
8.90 m (29 ft 2 1/2 in) **at Altitude**	2.0	Bob Beamon (USA)	Mexico City, Mexico	18 October 1968[1]
8.95 m (29 ft 4 1/4 in)	0.3	Mike Powell (USA)	Shinjuku, Tokyo, Japan	30 August 1991[1]

The question then remains, how can this variability in probability between events be accounted for in a general mathematical model?

The difficulty of breaking the world record for each event is reflected by the probability calculated by the model as it compares the distribution of results from the current pool (participants from the past three years) to the current world record, and hence variation between events is of no consequence.

As a result, each type of event is treated uniquely by the fundamental operation of the model and hence by simply applying the model our team has constructed, the committee faces no issue in distinction between events. In terms of weighting our factors, our model places complete weight on the quality of the current pool, which is then compared to the quality of past pools when the WR either was or was not achieved.

Question 5: Mathematical Model

Introduction

Extreme value theory is a branch of statistics that can be used to predict the occurrence of rare events, such as extreme flood, large insurance losses, stock market crash, human life expectancy and record breaking. Because a world record is an extreme event, we will use extreme value theory to predict the probability of breaking the world record in the near future. Almost all mathematical models aiming to predict ultimate world records are based on the development of the world record over many years and extrapolating the trend to the future. It uses only a limited number of past world records and therefore gives rather unreliable estimates and predictions. This is ***not*** the approach we are taking. Instead, we utilise as many top performances of elite athletes as possible in the last few years leading up to an event throughout various athletic events to predict the probability of breaking the world record in the near future. Since our mathematical model is based on a large set of data in a short up-to-date period of time, it is effective in predicting the probability of breaking a world record in the near future. Therefore, our model minimises the affect that factors such as *technological advancement*, *development in training practices* and *anti-doping laws* have on the probability of breaking a world record, since it is expected that these factors do not alter significantly in a short period of time. Our mathematical model is inspired by the approach taken by Einmahl and Magnus (2008) in their paper "Records in Athletics through Extreme Value Theory".

The following model produces a probability of a world record being broken at a specific event. The model utilises data from the three years prior to the event in order to estimate the probability that the world record will be broken at that event. Our team has applied this model to multiple events from the past decade to find a correlation between the estimated probability and the actual outcome. Based on the percentage range observed, the team has created a general decision-making scheme to decide whether or not insurance should be purchased based upon the calculated percentage probability. By following the method outlined below, a committee can calculate the probability that a World Record (WR) will be broken at each of their events and can thus determine whether or not insurance should be purchased on an event-by-event basis.

When using this model it is important to ensure that the addressed variables may be accurately determined possibly a month or longer before the event itself. As a result we have chosen to disregard the effects of weather in our model as it is near-impossible to precisely predict precipitation and humidity a month or longer prior to an event. Using data compiled over the previous 3 years to determine the comparative quality of the pool of athletes is much more accurate and reliable than relying on an unpredictable variables such as weather.

STAGE 1: Data Collection

The data collected on the best performance of top athletes will be ordered from best to worst, and only the data beyond a certain threshold will be selected for analysis. This is known as the *peak over threshold* approach and is one of the two approaches taken in extreme value theory.

1. Obtain the data on the best performance of top athletes from the official website of the International Association of Athletics Federation (IAAF) in the Top Lists for the past 3 years. We use the data from the 3 years prior to the year being investigated in order to obtain a large sample which is still considered current, allowing for reliable estimations and predictions to be made.
2. Collate the data into a spreadsheet such as the most commonly used Microsoft Excel.
3. Order the data from best performance to worst performance for each of the three years.
4. Running events are measured by using time. Thus, the lower the time, the better the performance. To avoid estimation problems of the parameters and solve this discrepancy, convert running times to average speed by dividing the total distance by the time taken to cover the distance. No conversion is required for other events where better performances result in higher raw scores.
5. Compare the worst performances across the 3 years of data and use the best of these 3 worst performances as the threshold, deleting any data points with worse performance than this threshold. This is the *peak over threshold* approach which is one of the two approaches used in extreme value theory. This ensures that there are no 'holes' in data. In other words, if we didn't do this, then we might have missed a whole set of data that we cannot obtain ranging between the worst performance of one year and the worst performance of another year.
6. Combine the data for the 3 years into the same columns and order the data from the best performance to the worst performance.
7. Use the column of athlete names to remove entries where the **same** athlete has another entry with a higher score (due to participation in multiple years). If using Microsoft Excel, the 'Remove Duplicates' function is applicable as is it will remove entries lower down in the list first. As the data has been ordered in the previous step, only the top performance for each athlete is retained, as is required.

   It is a requirement in extreme value theory that the data points are individually and identically distributed (i.i.d.). Thus, an athlete's performance (in this case his best performance in the last 3 years leading up to an event) can only appear once.

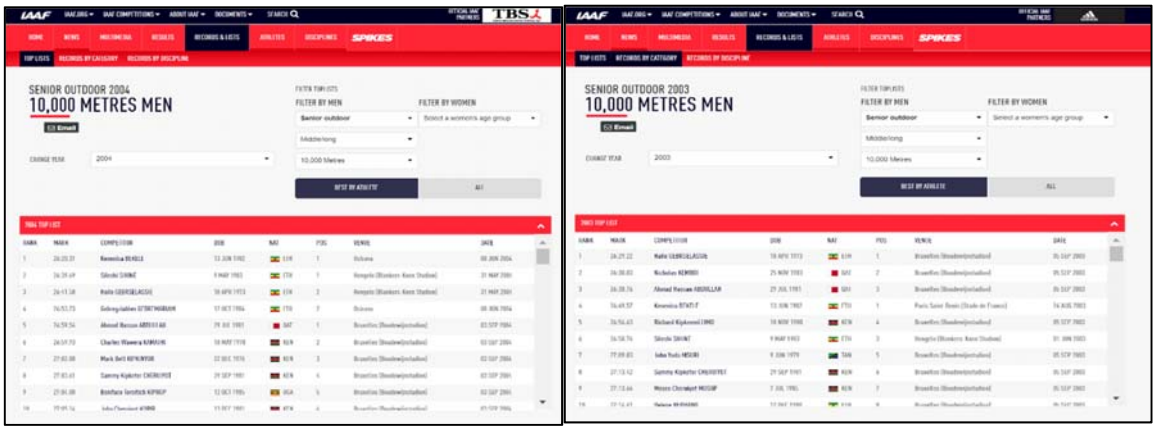
8. For some events (such as the 100m), data will occur in clusters because of the limited precision of measurement devices such as stopwatches. Because these clusters can cause problems in estimation, the data for these events needs to be 'smoothed' out.

   For example, suppose '**n**' top athletes run a personal best of 9.80 seconds in the 100m. This occurs not because the actual times are exactly the same, but because the electronic stopwatch is accurate to 0.01 of a second. We need to smooth these n results over the interval (9.795, 9.805) by:

$$d_j = 9.795 + 0.01 \times \frac{2j-1}{2n}, \;\; j = 1, \dots, n$$

For understanding purposes, outlined below is a step-by-step **example** determining the probability of breaking the world record in a 10,000 Metre Men's athletic event in 2005 assuming the participants are the top athletes in the world that are most likely in the sample of data taken from IAAF.

1. Go to IAAF Top Lists and collect data for 10,000 Metres Best by Athlete from years 2002, 2003 and 2004. This data is the best personal performance of the top athletes between 2002 and 2004 in the 10,000 Metres Men category.



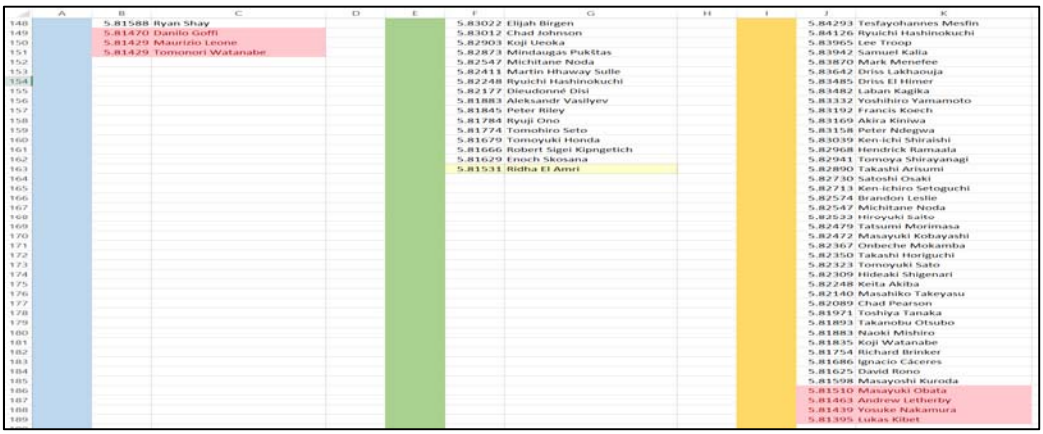
2. Collate the obtained data into a spreadsheet such as the commonly used Microsoft Excel. The data is already ordered from the best performance (lowest time) to the worst performance (greatest time) for each of the three years.

	A	B	C	D	E	F	G	H	I	J	K
1	2002	Mark	Competitor		2003	Mark	Competitor		2004	Mark	Competitor
2		26:49.4	Sammy Kipketer Cheruiyot			26:29.2	Haile Gebrselassie			26:20.3	Kenenisa Bekele
3		26:49.9	Assefa Mezgebu			26:30.0	Nicholas Kemboi			26:39.7	Sileshi Sihine
4		26:50.2	Richard Kipkemei Limo			26:38.8	Ahmad Hassan Abdullah			26:41.6	Haile Gebrselassie
5		26:50.7	Ahmad Hassan Abdullah			26:49.6	Kenenisa Bekele			26:53.7	Gebregziabher Gebremariam
6		26:52.9	John Cheruiyot Korir			26:56.6	Richard Kipkemei Limo			26:59.5	Ahmad Hassan Abdullah
7		27:05.9	Patrick Mutuku Ivuti			26:58.8	Sileshi Sihine			26:59.9	Charles Waweru Kamathi
8		27:06.2	John Yuda Msuri			27:09.8	John Yuda Msuri			27:02.0	Mark Bett Kipkinyor
9		27:20.2	Mebrahtom Keflezighi			27:13.4	Sammy Kipketer Cheruiyot			27:03.6	Sammy Kipketer Cheruiyot
10		27:25.6	Gebregziabher Gebremariam			27:13.7	Moses Cheruiyot Mosop			27:04.0	Boniface Toroitich Kiprop
11		27:26.1	Sileshi Sihine			27:14.6	Dejene Berhanu			27:05.1	John Cheruiyot Korir
12		27:26.3	Kamiel Maase			27:15.9	Boniface Toroitich Kiprop			27:07.3	Moses Ndiema Masai
13		27:35.0	Paul Biwott			27:17.2	John Cheruiyot Korir			27:09.6	Richard Kipkemei Limo
14		27:35.1	José Ríos			27:21.6	Paul Kosgei Malakwen			27:17.1	Nicholas Kemboi

3. Since the 10,000 Metre is a running event, it is measured using time. Consequently, a lower time is required to break the record. To ensure that a higher value is required to break the record, this time needs to be converted to average speed using the formula:

$$Average\ Speed = \frac{total\ distance}{total\ time}$$

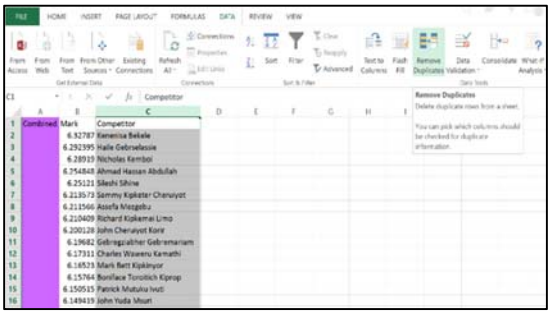
4. Compare the worst performances across the three years of data and use the best of these three worst performances as the **lower threshold**. This ensures that there are no 'holes' in data.



From the screenshot above, we are comparing the worst performances of the data sets of the each of the 3 years. The best performance of these three worst performances is used as the threshold. In this case, Ridha El Amri's average speed of 5.81531 m/s is the used LOWER THRESHOLD. All the worse performances (slower average speeds) than this lower threshold are removed from the data sets. In the above image, the deleted data points that fall short of the lower threshold are highlighted in red.

In 2003, there might have been a lower average speed than that of Ridha El Amri (5.81531 m/s) which is the worst performance of the data set for 2003, but at the same time greater than that of the worst performance of one or both of the other years. If we haven't used the best of the worst times as the lower threshold, it could have led to 'holes' in the overall data, therefore interfering with the data analysis.

5. Combine the data from the three years into the same columns and order the data from the best performance (greatest average speed) to the worst performance (lowest average speed).
6. Once the data has been ordered from best performance to worst performance, remove all the athletes that appear more than once leaving only their best performance. This fulfils the requirement of extreme value theory that all data points are individually and identically distributed.



7. In the case of the 10,000 Metre event, the data is not clustered as the precision limitations of the stopwatches do not pose a significant problem as there is a greater range of time values. Therefore, there is no need to 'smooth' the data in this case.

STAGE 2: DATA ANALYSIS (with EasyFit) and GENERALISED PARETO DISTRIBUTION

In our model, we use the *peak over threshold (POT)* method to consider the distribution of exceedances over a certain threshold. We are interested in estimating the cumulative distribution function of F(x) of random variables *X* above a certain threshold *u*.

The distribution function Fu is called the *conditional excess distribution function*

$$F_u(y) = P(X - u \mid X > u),$$

where X is a random variable, u is a given threshold and X-u is any random data point exceeding the threshold *u* minus the threshold *u*

Applying this to our model,

- *X* is the best performance of any athlete (or even more generally anyone) in a specific athletic event in the past 3 years
- *u* is the best performance value out of the 3 worst performance values of top athletes in the data set taken from IAAF in each of the years
- *X-u* is all the best performances of an athlete in the past 3 years exceeding the threshold performance minus the threshold performance.

At this point extreme value theory (EVT) can prove very helpful as it provides us with a powerful result about the conditional excess distribution function which is stated in the following theorem:

For a large class of underlying distribution functions F the conditional excess distribution function Fu(y), for u large, is well approximated by

$$Fu(y) \approx G_{\xi,\sigma}(y), \qquad u \to \infty$$

where

$$G_{\xi,\sigma}(y) = \begin{cases} 1 - \left(1 + \frac{\xi}{\sigma} y\right)^{\frac{-1}{\xi}} & if\ \xi \neq 0 \\ 1 - e^{-\frac{y}{\sigma}} & if\ \xi = 0 \end{cases}$$

where y = X-u, the exceedances over the threshold

and $0 \leq y \leq xF - u$ if $\xi \geq 0$

and $0 \leq y \leq -\sigma\xi$ if $\xi < 0$.

(Pickands (1975), Balkema and de Haan (1974))

(Gilli, & Kellezi, 2006)

$G_{\xi,\sigma}(y)$ is the called the Generalised Pareto distribution (GPD) with 2 parameters

1. Shape parameter ξ
2. Scale parameter σ˜.

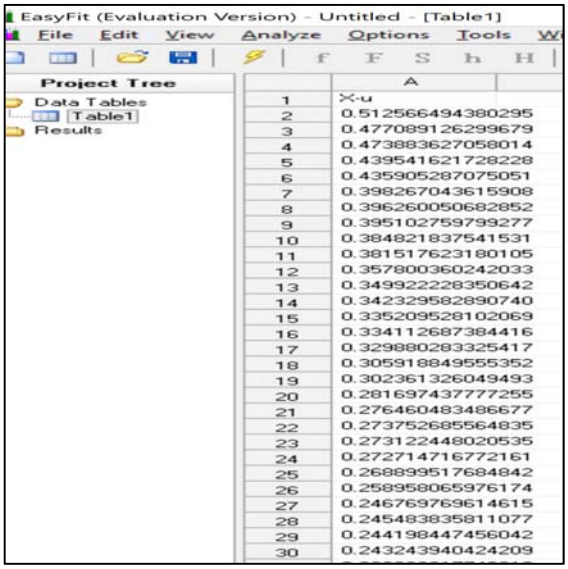
Using all the best performances of top athletes for the past 3 years exceeding the LOWER THRESHOLD established earlier, we can utilise the Generalised Pareto distribution to predict the probability of breaking the world record in the year in a specific athletic event following the 3 years for which data has been collected.

Because of the tedious process of estimating the parameters of the Generalised Pareto distribution, our model utilises the program EasyFit that can be downloaded from the website http://www.mathwave.com/ to estimate these parameters based on the data and fit a probability density curve. It also runs three Goodness of Fit tests: Kolmogorov-Smirnov, Anderson Darling and Chi-Squared. Based on the Kolmogorov-Smirnov goodness of fit tests for several different data sets from different athletic events, we have found that the Generalised Pareto distribution is ranked almost always in the top 3 best fits with a good fit, which is expected since we are using a peak over threshold approach.

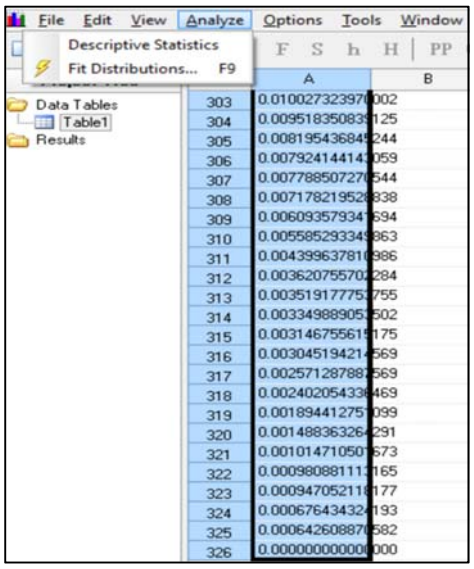
For understanding purposes, outlined below is a step-by-step EXAMPLE for analysing the data using EasyFit software to project the probability of breaking the world record in a 10,000 Metre Men's athletic event in 2005 assuming the participants are the top athletes in the world that are most likely in the sample of data taken from IAAF.

1. Copy the X-u column in Excel and paste it into the first column of Easy Fit

A	B	C	D
Combined	Competitor	Average Speed	X-u
	Kenenisa Bekele	6.32787	0.512566494380295
	Haile Gebrselassie	6.292395011	0.477089126299679
	Nicholas Kemboi	6.289189512	0.473883627058014
	Ahmad Hassan Abdullah	6.254847507	0.439541621728228
	Sileshi Sihine	6.25121	0.435905287075051
	Sammy Kipketer Cheruiyot	6.213572929	0.398267043615908
	Assefa Mezgebu	6.211565936	0.396260050682852
	Richard Kipkemei Limo	6.210408645	0.395102759799277
	John Cheruiyot Korir	6.200127723	0.384821837541531
	Gebregziabher Gebremariam	6.19682	0.381517623180105
	Charles Waweru Kamathi	6.17311	0.357800360242033
	Mark Bett Kipkinyor	6.16523	0.349922228350642
	Boniface Toroitich Kiprop	6.15764	0.342329582890740
	Patrick Mutuku Ivuti	6.150515413	0.335209528102069
	John Yuda Msuri	6.149418572	0.334112687384416
	Moses Ndiema Masai	6.14519	0.329880283325417
	Moses Cheruiyot Mosop	6.121224735	0.305918849555352
	Dejene Berhanu	6.117667211	0.302361326049493
	Mebrahtom Keflezighi	6.097003323	0.281697437777255
	Paul Kosgei Malakwen	6.091766369	0.276460483486677
	Simon Maina Munyi	6.089058571	0.273752685564835

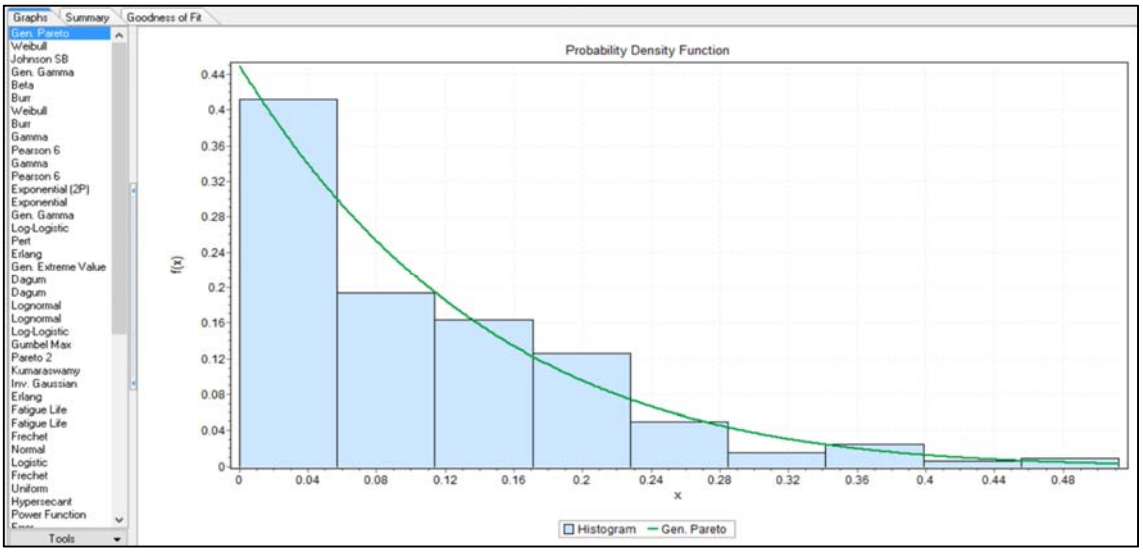


2. Highlight all the data in the column, click on the 'Analyse' tab and choose 'Fit Distributions'. EasyFit will fit many different distributions to the provided data ranking the distributions from the best fit to the worst fit by running several goodness of fit tests. It is expected for the generalised Pareto distribution (abbreviated to Gen. Pareto by EasyFit) to be among the best fits to the data. Indeed, for our example used, the generalised Pareto distribution is the best fit.

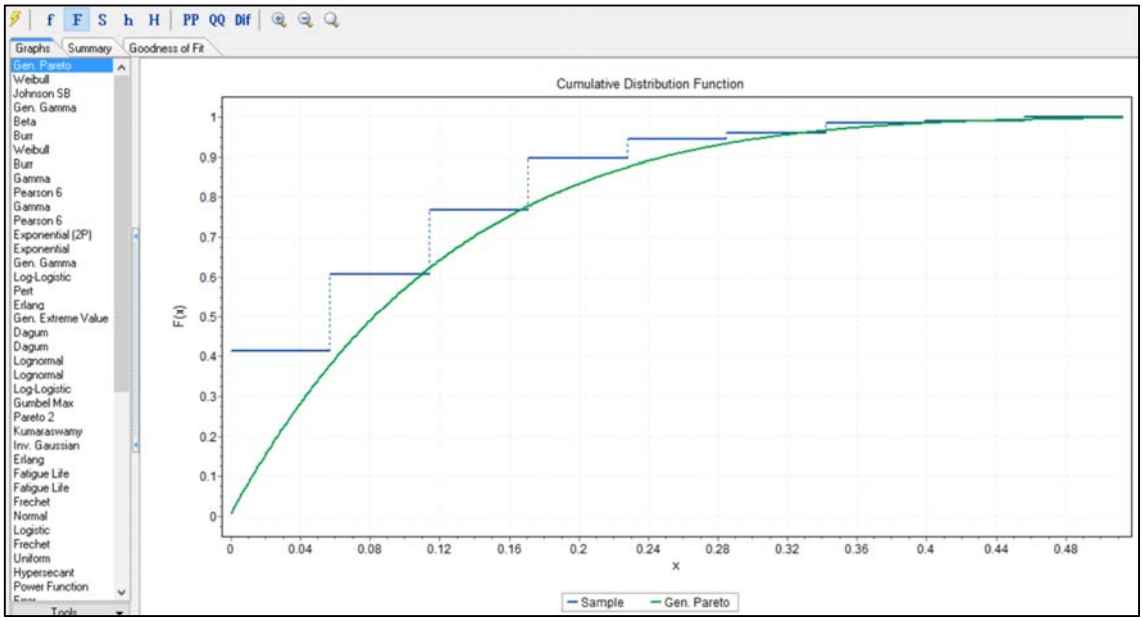


Distribution	Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
	Statistic	Rank	Statistic	Rank	Statistic	Rank
Beta	0.06448	9	1.8966	11	22.614	10
Burr	0.10405	25	96.432	50	N/A	
Burr (4P)	0.05307	3	0.89807	2	17.833	3
Cauchy	0.20284	43	21.117	41	80.181	40
Chi-Squared	0.67539	55	166.58	54	5065.8	48
Dagum	0.27302	49	70.96	48	248.22	45
Dagum (4P)	0.05815	6	5.616	24	N/A	
Erlang	0.1439	26	10.827	28	46.269	28
Erlang (3P)	0.08156	19	2.4737	14	26.208	13
Error	0.1933	42	14.507	39	70.786	39
Error Function	1	57	N/A		N/A	
Exponential	0.62522	53	144.62	52	8093.4	49
Exponential (2P)	0.05953	7	1.3066	6	21.192	7
Fatigue Life	0.14727	32	10.904	30	45.94	27
Fatigue Life (3P)	0.07514	15	1.523	9	24.527	12
Frechet	0.09763	24	4.4817	21	31.338	18
Frechet (3P)	0.08786	22	2.6672	17	34.047	20
Gamma	0.14637	29	11.025	31	46.418	29
Gamma (3P)	0.05804	5	0.99701	4	15.305	1
Gen. Extreme Value	0.07394	14	3.1222	19	30.314	17
Gen. Gamma	1	58	N/A		N/A	
Gen. Gamma (4P)	0.0721	13	1.5121	8	22.577	9
Gen. Pareto	0.04262	1	0.77317	1	19.375	4

From the screenshot above, the generalised Pareto distribution for our example used is the best fit according to both the Kolmogorov-Smirnov and the Anderson Darling goodness of fit tests. We suggest that the committee base the goodness of fit of the Kolmogorov Smirnov test. The rank can be checked by going to the 'Goodness of Fit' tab and clicking 'Kolmogorov Smirnov' to give the rank in ascending order.



This is the probability density function of the data used for our example. The green line is the generalised Pareto distribution fit for the probability density.

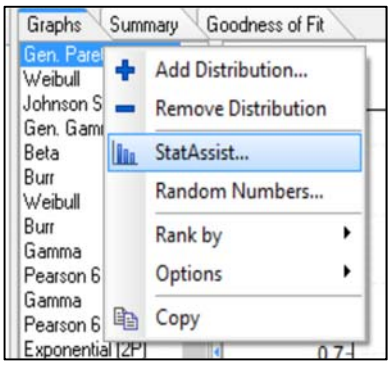


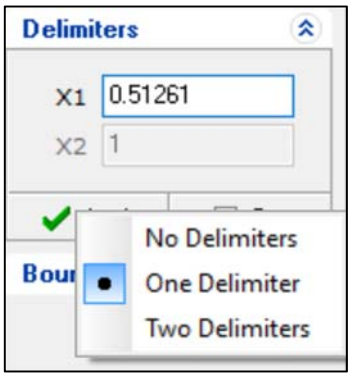
This is the cumulative distribution function and the green line is the generalised Pareto function fit.

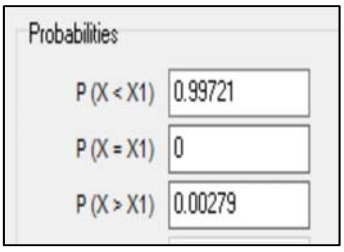
To view the probability density function select the f (lower case) at the top and to view the cumulative distribution function select F (upper case) at the top.

Note: In the process of fitting the generalised Pareto distribution, EasyFit estimates the two parameters ξ and σ and graphs the fit based on these estimates. This saves the committee from performing tedious mathematics to find the estimates themselves.

3. Once the generalised Pareto distribution is fitted
    a. Right-click on 'Gen. Pareto' and select 'Stat Assist'
    b. Select the 'Probabilities' tab
    c. Click on 'Delimiters' on the right hand side of the page
    d. Click on 'None'
    e. Select 'One Delimiter'
    f. Calculate the World Record value minus the threshold value (make sure to convert it average speed for a running evet)
    g. In 'X1' type in the calculated value for World Record minus threshold value



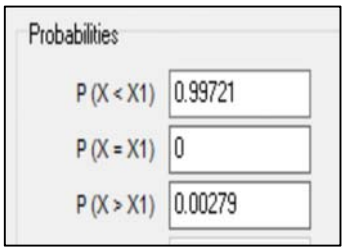




P(X>X1) gives the probability of a randomly selected top athlete from our large data set will break the World Record in an event based on his best performance in the last 3 years. Rephrasing for our example used, what is the probability of a top athlete in 10,000 metre event from our data set getting a higher average speed for the event than the average speed for the event in the Word Record.

P(X<X1) gives the probability of a randomly selected top athlete from our large data set will NOT break the World Record in an event based on his best performance in the last 3 years.

For our example of the 10,000 Metre Men event, we know for certain that within the 2002-2004 period, the WR set in 2005 had not yet been achieved. However, using the generalised Pareto distribution fitted by EasyFit and Stat Assist, we can project the probability that a single top athlete in this event, based on his best performance between 2002 and 2004, has to break the World Record.



RESULT = 0.00279 = 0.279%

The result obtained above is the projected probability of a top athlete in the 10,000 metre men event based on his best performance between 2002 and 2004, breaking the world record at the time, provided his best performance is at the event the committee is organising.

Now that we have calculated this probability of a single athlete breaking the record in his best performance, we will assume worst case scenario and use this probability as the probability of the athlete breaking the record at OUR event.

Stage 3: Binomial Distribution

In stage 2 we were able to extrapolate based on a comprehensive data set, and calculate an estimate of the probability that a given top athlete's **best performance through 2005**, which we assume occurs at our event, breaks the world record.

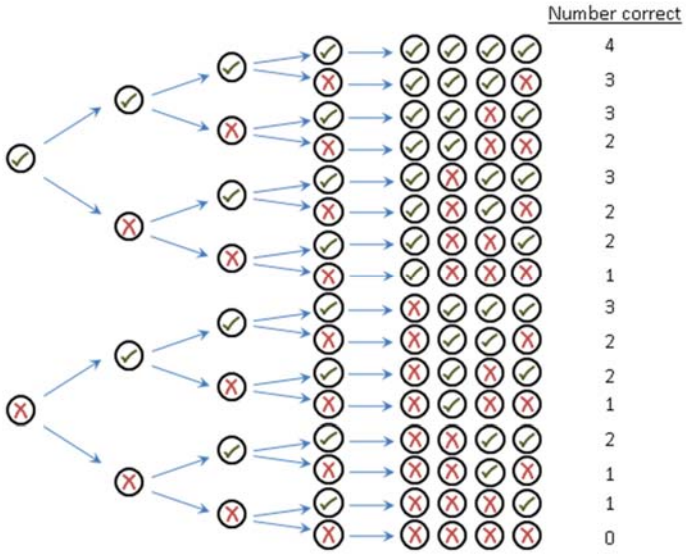
We then use a binomial distribution to factor in the number of athletes competing in the event to calculate the probability that the record is broken at the event.

A Binomial Distribution is a series which occurs when a Bernoulli Trial is repeated 'n' times.

Notice our situation can be modelled as a Binomial Distribution with:

- Probability of a 'Success' = Probability estimate calculated in STAGE 2 = 0.279%
- Number of Trials = Number of Athletes competing in the committee's event

For example, the diagram on the right represents an event with 4 competitors, where a tick means they broke the WR while a cross means they failed to. Notice the only situation in which the WR was not broken is the bottom progression with 4 failures. To further generalise this, in a binomial distribution of any number athletes, the only situation in which the WR is not broken is when all 'n' athletes fail to achieve the needed result. Therefore, for a race with 'n' athletes:



$$Probability\ of\ WR\ being\ broken\ at\ our\ Event = 1 - (1 - P_{success})^n$$

where $P_{success}$ = the probability that a given top athlete will break the world record

and 'n' = the number of athletes competing

In this walkthrough we have considered the years 2002-2004 in anticipation of the 2005 Memorial Van Damme athletics event, in particular the 10km race. We have calculated $P_{success}$ = 0.279%, and we know that in our event 19 competitors will compete.

$$Probability\ of\ WR\ being\ broken\ at\ our\ Event = 1 - (1 - 0.00279)^{19}$$
$$= 0.0517$$
$$= 5.17\%$$

Using the above formula we see the probability that the record will be broken at our event is 5.17%.

Stage 4: Interpreting the Estimated Probabilities

At this point the committee now has a means to estimate the probability that a world record shall be broken during any one of their events. This applies to the 15km event as well as to each of the 40 track and field events.

Yet, a raw probability of occurrence still is not useful, as how does one decide at what probability insurance must or must not be purchased? For example, in our previous walkthrough how would the committee know if 5.17% is a probability worth of insurance or not?

This is where our research can now be applied. After applying our model repeatedly to multiple past events, a noticeable correlation between the **estimated percentage probability calculated by our model,** and the **actual result the following year** arose. So how was this correlation established?

- Our team trialled the model **twice** for three different events (200m, 5km, 10km).
- The first trial was upon the 3 years leading up to a year where we knew the **record WAS broken.**
- The second trial was upon the 3 years leading up to a year where we knew the **record WAS NOT broken.**
- For each trial we recorded our estimated percentage probability, and noticed that the estimates garnered before a successful year formed a distinct group as opposed to the estimates garnered before a failed year.
- From this correlation we were able to create a percentage range where the committee can input their own calculated percentage and determine whether or not insurance is necessary.

Interpretation of the Example

For example, the walkthrough provided in stages 1-3 is an example of a test for a SUCCESSFUL year. Our researchers knew that at the 2005 Memorial Van Damme Athletics Event in Belgium, the 10km World Record was broken by Kenenisa Bekele. Therefore, the estimated probability of the world record being broken in this event based off our data from 2002-2004 (calculated to be 5.17%), is recorded and categorised as a **WR breaking probability.**

RESULTS

Place	Athlete	Nationality	Time	
1	Kenenisa Bekele	Ethiopia	26:17.53	WR
2	Boniface Kiprop	Uganda	26:39.77	
3	Samuel Wanjiru	Kenya	26:41.75	WJR
4	Nicholas Kemboi	Kenya	26:51.87	
5	Sammy Kipketer	Kenya	26:52.60	
6	Mark Bett	Kenya	26:52.93	
7	Zersenay Tadesse	Eritrea	27:04.70	
8	Bernard Kipyego	Kenya	27:26.29	
9	John C Korir	Kenya	27:30.46	
10	Denis Ndiso	Kenya	27:54.71	
11	Stephen Koech	Kenya	27:55.92	
12	Ayad Lamdassen	Morocco	27:59.39	
13	John Yuda	Tanzania	28:05.86	
14	Bernard Chepkok	Kenya	28:11.45	
--	Tariku Bekele	Ethiopia	DNF	
--	Roberto Garcia	Spain	DNF	
--	Martin Keino	Kenya	DNF	
--	Geoffrey Kipngeno	Kenya	DNF	
--	El Hassan Lahssini	France	DNF	

However, we also required a non-WR breaking probability for the 10km event, and hence chose to apply our model to 2014, a year where the record was not broken. The exact same process was used, with data from 2011-2013 analysed, plotted, fitted, extrapolated and manipulated to produce an estimated probability of the WR being broken in 2014. The probability of this was 0.87%, and was recorded as a **non – WR breaking probability.** This was the process that was applied to the 10km Mens' Running event. We then proceeded to **evaluate both a WR breaking probability and a non – WR breaking probability for various other events**, recording our results.

Correlation Between the Estimated Outcome and Actual Outcome		
EVENT	SUCCESS	FAIL
10km Mens	5.17%	0.870%
5km Mens	5.98%	1.25%
200m Mens	11.6%	0.00432%

From analysis of these events, it is clear that for an estimated probability above 5% the world record will likely be broken in the following year, however if the probability is below 1% it is unlikely that the record will be broken. If a value between 5% and 1.5% is obtained by a committee, values above or equal to 3% should be rounded up to 5% and values below 3% should be rounded down to 1%.

Therefore, if the calculated probability is above or equal to 3%, it is likely that the world record will be broken and the committee should purchase insurance, however if the probability is below 3%, it is unlikely that the world record will be broken and insurance should not be purchased.

Advantages and Limitations of the Model

The method used to determine whether the world record will be broken has various advantages and disadvantages.

A limitation of the model is that for some events (such as long jump) where the world record has not been broken for a significant amount of time, records are not provided on the main IAAF website for the three years prior to the record being broken. This means that records would need to be obtained either from archived IAAF data or from a different source, where it may be presented in an alternate format. This increases the difficulty in collating and analysing the data for a year where the record was broken when attempting to compare the probability of the world record being broken prior to its actual historical occurrence and prior to the current event.

An advantage of our model is that it is able to produce a percentage probability which our team has found to correlate heavily with the actual occurrences and non-occurrences of the world record being broken. This probability is based on a sample of current top athletes which should be a fair representation of the athletes participating in the event.

Furthermore, our method may be applied to previous years where the outcome is now known (as we have shown above), meaning the model's validity and the decision-making scheme can be tested for each event before it is used to determine whether or not insurance should be purchased.

REFERENCES

Bionic Turtle,. (2008). *Extreme Value Theory (EVT)*. Retrieved from https://www.youtube.com/watch?v=o-cpu1IH3tM

Chen, L. (2014). *Predicting the Limits of Records in Athletics*. Retrieved from http://www.d.umn.edu/math/graduates/documents/ChenTR.pdf

Einmahl, J., & Smeets, S. (2009). *ULTIMATE 100M WORLD RECORDS THROUGH EXTREME-VALUE THEORY*. Retrieved from http://poseidon01.ssrn.com/delivery.php?ID=032102104004096027126070007030071028042048049042095026105087108064086084113099121000043045125052037037114120071069122083082117118061035009009030100002099068083111063062037121120003019124093066097001027086100113094122096070080025029018022124099098076&EXT=pdf

Gilli, M., & Kellezi, E. (2006). *An Application of Extreme Value Theory for Measuring Financial Risk*. Retrieved from http://www.unige.ch/ses/dsec/static/gilli/evtrm/GilliKelleziCE.pdf

Heisler, K. (2009). *MODELING LOWER BOUNDS OF WORLD RECORD RUNNING TIMES*. Retrieved from http://www2.stetson.edu/~efriedma/research/kheisler.pdf

IAAF: IAAF - International Association of Athletics Federations. (2016). *iaaf.org*. Retrieved 17 April 2016, from http://www.iaaf.org/home

Magnus, J., & Einmahl, J. (2008). *Records in Athletics Through Extreme-Value Theory*. Retrieved from http://cdata4.uvt.nl/websitefiles/magnus/JRM082.pdf

Schneider, U. (2016). *An Introduction to Statistical Extreme Value Theory* (1st ed.). Retrieved from https://www.google.com.au/url?sa=t&rct=j&q=&esrc=s&source=web&cd=4&cad=rja&uact=8&ved=0ahUKEwjKzdy6ipbMAhULopQKHY1IAbsQFgg2MAM&url=https%3A%2F%2Fwww2.meteo.uni-bonn.de%2Fprojekte%2FSPPMeteo%2Fwiki%2Flib%2Fexe%2Ffetch.php%3Fid%3Dcops_2a%26cache%3Dcache%26media%3Devt_cops2.pdf&usg=AFQjCNEExHMFlmXe9sEeBYIVWldR0zpbzQ&sig2=RaalwYfF5jBpu2L_H20Pjg

Technologies, M. (2016). *Distribution Fitting FAQ*. *Mathwave.com*. Retrieved 17 April 2016, from http://www.mathwave.com/articles/distribution_fitting_faq.html#q5