

Evolution of Athletic Records: Statistical Effects versus Real Improvements

DANIEL GEMBRIS*, JOHN G. TAYLOR[†] & DIETER SUTER[‡]

*University of Heidelberg in Mannheim, Germany, [†]Department of Mathematics, King's College London, UK, [‡]Department of Physics, University of Dortmund, Germany

ABSTRACT *Athletic records represent the best results in a given discipline, thus improving monotonically with time. As has already been shown, this should not be taken as an indication that the athletes' capabilities keep improving. In other words, a new record is not noteworthy just because it is a new record, instead it is necessary to assess by how much the record has improved. In this paper we derive formulae that can be used to show that athletic records continue to improve with time, even if athletic performance remains constant. We are considering two specific examples, the German championships and the world records in several athletic disciplines. The analysis shows that, for the latter, true improvements occur in 20–50% of the disciplines. The analysis is supplemented by an application of our record estimation approach to the prediction of the maximum body length of humans for a specified size of a population respectively population group from a representative sample.*

KEY WORDS: Records, athletics, estimation of maxima and minima

Introduction

Athletic competitions are conducted as tests of the abilities of individual athletes or teams, in order to produce a quantitative ranking of their capabilities. Besides this goal of fair ranking all competitions also include elements of chance. This is perhaps best demonstrated by the fact that many different kinds of gambling are based on the outcomes of athletic competitions. If reliable predictions of athletic results performances were possible, such markets would no longer be of interest. Nevertheless, a statistical analysis shows that it is possible to predict the range into which records are most likely to fall in future periods (Gembris *et al.*, 2002).

In previous work on the analysis of athletic top performance the annual best results have been regarded as drawn from an asymptotic extreme value distribution (Coles, 2004). If a value differed by more than a threshold value from the mean of this distribution, it can be regarded as an outlier. Furthermore, models for describing systematic trends (i.e. progress of athletic performance), which have to be distinguished from stochastic effects, have been proposed and applied. In this work the emphasis lies on the (analytical) description of the temporal development of records by means of order statistics for small sample sizes. Other

real-world applications of extremal statistics, for which a theoretical overview is given in Galambos (1978), has previously been presented in Coles (2004). Two examples for a biomedical application are the analysis of epilepsy data and the detection of tumours in MR images (Roberts, 2000).

Record Estimation

The evolution of records is determined by two factors: systematic trends (i.e. progress of athletic performance) and stochastic effects. The stochastic effects are best taken into account by considering the annual best results whose maximum up to the given point in time represents the record. We assume that these annual best results can be described as a strongly stationary i.i.d. (=independently identically distributed) random process, i.e. two consecutive values are independent and all random numbers obey the same statistical distribution. The records for which we try to find estimates may be either maxima (e.g. of measures of length for disciplines like jumping or throwing) or minima (typically of running times). We discuss here explicitly the case of a maximum, but include applications to the case of minima. Our goal is to determine the expectation value $\hat{x}_{\max}$ for the best result during a period B ,

$$\hat{x}_{\max} = E \left(\max_{i=1}^N (x_i) \right), \quad (1)$$

where the best result of *each year* is denoted as x_i .

Considering the annual best results as a stationary stochastic series, we describe the probability density distribution of the yearly results as $\rho(x)$ determined from a period A . The probability that the best result for a given year remains below x is $\int_{x''=-\infty}^x \rho(x'') dx''$ and the probability that for a value x' reached in a specific year the best results of all other years remain below this value is given by

$$\left(\int_{x''=-\infty}^{x'} \rho(x'') dx'' \right)^{N-1}.$$

The expectation value $\hat{x}_{\max}$ can thus be calculated as

$$\hat{x}_{\max} = N \int_{x'=-\infty}^{\infty} x' \left(\int_{x''=-\infty}^{x'} \rho(x'') dx'' \right)^{N-1} \rho(x') dx'. \quad (2)$$

By the normalization factor N it is taken into account that the best result can occur during any of the N years with equal probability. A corresponding formula can be found for the minimum¹ and similar formulae for the standard deviation of the estimations, which specify the confidence intervals (see the appendix).

Since the distribution $\rho(x)$ is stationary, the function $\hat{x}_{\max}$ increases monotonically with N . The expectation value shifts with the mean μ of the distribution ρ and scales linearly with the standard deviation σ of the distribution ρ (the former corresponds to an offset, the latter to a change of the units). Figure 1 represents graphically the evolution of $\hat{x}_{\max}$ with N assuming a Gaussian distribution, for $\mu = 0$ (effecting only the offset) and a standard deviation of $\sigma = 1$. In order statistics, the right-hand side (rhs) of equation (2) is interpreted as the expectation value of the extrema distribution belonging to ρ .

To obtain analytical results, we again assume the distribution ρ to be a Gaussian.² Exact analytical expressions for equation (2) exist then for small sample sizes N ($1 \leq N \leq 6$) (David, 1970). For the larger values of N , in which we are interested, it is possible to obtain

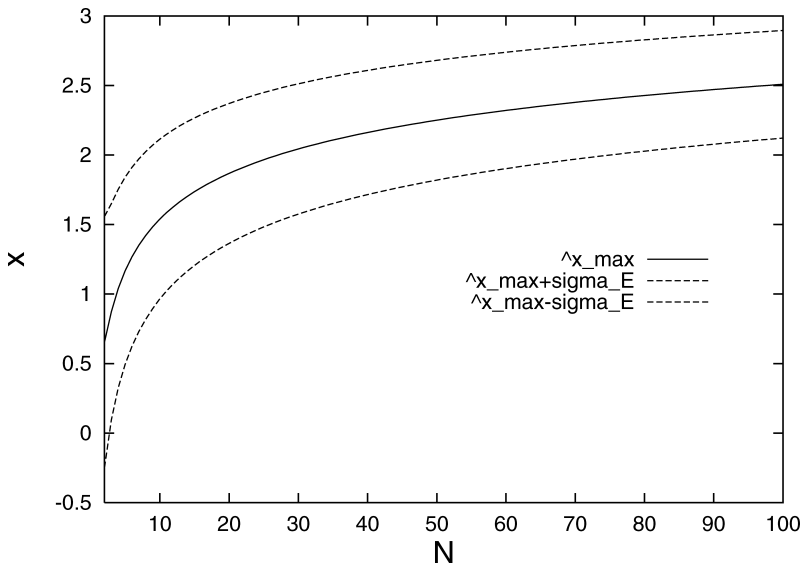


Figure 1. Increase of the estimate $\hat{x}_{\max}$, $\hat{x}_{\max} - \sigma_E$ and $\hat{x}_{\max} + \sigma_E$ for a Gaussian probability density function ρ with $\mu = 0$ and $\sigma = 1$

an approximation for $\hat{x}_{\max}$ by expanding the asymptotic maxima distribution ($N \rightarrow \infty$) described in Galambos, 1978) in a Taylor series as a function of $\ln \ln N$ (Gembris, 2002);

$$\hat{x}_{\max} = \mu + \sigma (a_0 + a_1 \ln \ln N + a_2 (\ln \ln N)^2), \quad (3)$$

where μ and σ are the mean and standard deviation of ρ . The coefficients a_i can be optimized for the relevant range of N . If, for example, $1 \leq N \leq 100$ the optimal coefficients are $a_0 = 0.818$, $a_1 = 0.574$ and $a_2 = 0.349$. With these coefficients and $\sigma = 1$, the error of the approximation of $\hat{x}_{\max}$ remains $< 0.06\%$ over the given interval. Analogously we approximate the standard deviation of $\hat{x}_{\max}$ as

$$\sigma_E = \sigma (b_0 + b_1 \ln \ln N + b_2 (\ln \ln N)^2). \quad (4)$$

The optimal parameters for the same range of N are $b_0 = 0.8023$, $b_1 = -0.2751$ and $b_2 = 0.0020$, resulting in an approximation error of less than 0.15% .

German and International Athletics Results

We have applied the estimation approach to two sets of athletics records, the German championships and the worldwide annual best results (Scherer, 1997; Steinmetz, 2000). The analysed data set for the German results, which we consider first, covers the top performances of men in 22 disciplines during the years 1973–1996.³ We divided this data set into a reference period (1973–1984), from which we calculated the mean annual best result μ for each discipline, as well as its variance σ^2 . These parameters then formed the starting point for an estimation of the best result over the subsequent period 1985–1996. Mean and variance of all 22 disciplines are listed in columns two and three of Table 1. As a check for consistency and the presence of trends, we calculated these parameters also for the second period; these results are shown in the fourth and fifth columns.

As a simplification we assumed the annual best results to be Gaussian distributed. More accurately these results themselves would have to be described by order statistics, but

Table 1. Estimation and comparison of best results in athletics. The mean and the standard deviation of the best values of all competitions that were carried out during the annual German Athletics Championship between 1973 and 1984 are the input information for the estimation of the 'best' best values between 1985 and 1996 (Scherer, 1997). Results are given as times in the notation h:min.s

Competition	μ for 1973-1984		μ for 1985-1996		σ	p_{ST}	Estimation	σ_E	Real result (within.... σ_E -interval? .. $2\sigma_E$ -interval?)
	h:min.s	resp. m	h:min.s	resp. m					
100 m	10.35		10.35		0.14	0.805	10.15	0.07	10.16 (y, y)
200 m	20.78		20.59		0.17	0.052	20.34	0.15	20.38 (y, y)
400 m	45.54		45.64		0.49	0.607	44.76	0.27	44.92 (y, y)
800 m	1:47.23		1:47.55		1.35	0.569	1:45.08	0.75	1:45.73 (y, y)
1500 m	3:40.00		3:42.46		4.25	0.082	3:36.82	1.11	3:37.36 (y, y)
5000 m	13:42.13		13:45.76		14.62	0.524	13:21.38	7.26	13:26.54 (y, y)
10000 m	28:32.05		28:27.96		23.76	0.552	28:00.89	9.12	27:57.98 (y, y)
Marathon	2:17:53.08		2:16:07.5		2:48.41	0.158	2:12:50.88	1:45.68	2:12:12 (y, y)
110 m hurdle	13.91		13.45		0.24	3.943E-5	13.58	0.11	13.05 (n, n)
400 m hurdle	49.41		49.14		0.76	0.493	47.63	0.62	48.02 (y, y)
3000 m									
Steeplechase	8:28.33		8:29.83		4.50	0.629	8:19.06	3.24	8:20.5 (y, y)
4 × 100	39.71		39.55		0.21	0.225	39.09	0.22	39.2 (y, y)
4 × 400	3:06.14		3:06.47		1.60	0.583	3:03.99	0.75	3:03.04 (n, y)
20km walking	1:29:33		1:25:13		2:47.92	2.578E-4	1:26:15	1:09	1:21:40 (n,n)
50km walking	4:03:59		3:55:54		4:20.12	0.001	3:57:25	2:18	3:45:57 (n,n)
High jump	2.24		2.30		0.04	0.017	2.35	0.04	2.38 (y, y)
Pole vault	5.36		5.55		0.09	0.001	5.60	0.09	5.70 (n, y)
Long jump	7.94		8.00		0.14	0.277	8.14	0.07	8.24 (n, y)
Triple jump	16.37		16.92		0.21	3.234E-4	17.02	0.23	17.31 (n, y)
Shot-put	19.75		20.20		0.26	0.030	20.76	0.35	20.53 (y, y)
Discus-throw	62.23		65.93		1.45	9.962E-5	65.96	1.31	69.02 (n,n)
Throwing the hammer	76.41		78.94		2.25	0.013	80.08	1.28	82.16 (n, y)

this could be a complex undertaking as outlined in the discussion section. Nevertheless the Gaussian distribution represents a very good approximation to these results (and the international ones) as has been verified using the Kolmogorov-Smirnov test (Press *et al.*, 1995). The quality of the approximation is quantified by a parameter $p_{KS} \leq 1$. If the samples were drawn from a true Gaussian population, the mean value of p_{KS} would be 0.5, compared to 0.56 for the measured data. Furthermore, distributions from asymptotic order statistics for the annual results would have broader tails compared with the Gaussian, which makes our estimations conservative, i.e. the estimation intervals would be larger if one used a non-Gaussian distribution. Performing the estimation with cut-offs at zero and at a physiological limit (Hammand & Diamond, 1997) showed that their effect can be safely neglected.

A comparison of columns 2 and 4 shows that changes of the mean leading to improvements prevail: in 15 of the 22 competitions an increase of the average best result occurred, in the other seven, a decrease. If no trends existed the number of increases and decreases would be very closely matched. To check for the presence of trends we calculated the probability that the average of the data for both periods is the same within the statistical uncertainty by applying a t-test. The resulting probabilities (p_{ST}) are given in the sixth column. 11 cases show a clear shift of the mean ($p_{ST} \leq p_{thr} = 0.1$); with one exception these changes were towards better performance. The shift is strongest for discus-throw, the 20 km- and 50 km-walking competitions, 110 m hurdles, triple jump and pole vault, indicating that systematic progress has been made in these events. In the other 11 disciplines the shifts are weak and apparently stochastic, with roughly half of these disciplines advancing.

The predicted best results $\hat{x}_{max}$ for 1985–1996 are given in the seventh column of Table 1 together with the corresponding confidence interval given by σ_E in the eighth column. The last column of Table 1 gives the actual best result for this period and indicates whether it lies within an interval $\hat{x}_{max} \pm \sigma_E$ or $\hat{x}_{max} \pm 2\sigma_E$. Figure 2 allows one to compare the prediction with the actual results for the second period graphically. The lines mark the theoretical estimate and the $\pm\sigma_E$ confidence interval. The agreement between the actual and the theoretically predicted values is surprisingly good, even in some disciplines where trends appear to be present ($p_{ST} \leq 0.1$). Only four disciplines (discus-throw, 20 km and 50 km walking, and 110 m hurdle) fall clearly outside of the prediction interval. For all four disciplines, the t-test gave clear indications of an underlying trend ($p_{ST} \leq 0.001$) and in all four cases the actual results are better than the predicted values. These results make it clear that systematic progress was achieved for this set of disciplines.

An essential parameter for determining the accuracy of the estimation besides the shift in the mean is a change of the variance between both time intervals. A *decrease* of the σ increases the probability that the actual best results fall into the predicted confidence interval. An *increase* in σ leads to an underestimation of the best value when using data from the earlier period. This we can see for the 10000 m run (with a $>50\%$ increase of σ), where the actual best result is 3 seconds faster than our estimate but remains within one standard deviation, as well as for the 50 km walk. The latter additionally shows a significant decrease of the mean, but that alone cannot explain the underestimation.

In addition to the data of the German championship, we have analyzed the time series of the worldwide annual best results, that we obtained from a comprehensive database of athletic performances in the 20th century (Koponen, 2000). The number of athletes and competitions that is selected by this data set is significantly larger than for the German data set. We analyzed the data for the years 1980–1999, using the first half as a reference for predicting the record during the second half, as shown in Table 2. For an assessment of trends we also used data from the period 1970–1979. The mean and standard deviation for the reference interval are given in the second and third column (please note: the results for the running disciplines are given as average speeds, not times as in Table 1). As Table 2

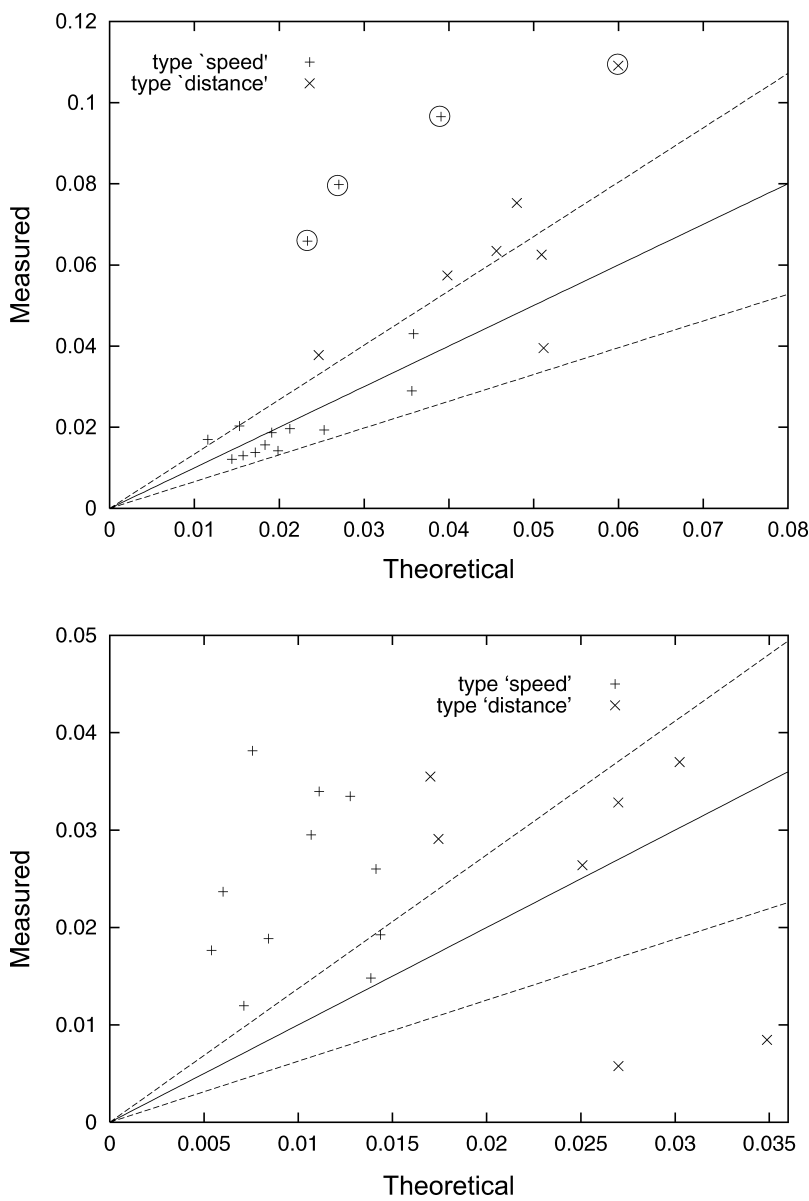


Figure 2. Top: comparison of actual and theoretical results for the relative differences between the best performances at the German Championship of the second 12 years interval and the mean annual best result of the first 12 years interval (x -axis: $\hat{x}_{best}/\mu - 1$, y -axis: $x_{best}/\mu - 1$). Jumping and throwing disciplines and running disciplines are labeled with different symbols. Best agreement is reached in the vicinity of the $y = x$ line. Only four disciplines (marked by circles) lie far outside of the confidence interval, indicating a clear systematic progress. The dashed lines (which are not symmetrical to the $y = x$ line) indicate the $\pm \sigma_E$ intervals. Bottom: analogous plot for the worldwide annual best results (10-year periods)

Table 2. Estimation results for the worldwide annual best results (Koponen, 2000) similar to Table 1. Here the results of running events are given as average speed (in meters per second) and the reference, respectively estimation, interval is 1980–1989 and 1990–1999. The values p_m obtained by a median two-sample test (see text) are stated as a measure for trends. A comparison of the two corresponding columns shows that stationarity of the annual best results is increasing: there are nine events with $p_m > 0.1$ for the first interval pair, but 12 for the second. The last column states a stationarity measure for the 20-year period 1980–1999 (*: To prevent the javelins from flying too far, a rule change was implemented in 1985)

Competition	μ	σ 1980-89	Estimation	Real result (within $\sigma_E/2\sigma_E$ interval)	p_m 1970-79 vs. 1980-89	p_m 1980-89 vs. 1990-99	p_c
100 m	10.037	0.035	10.091	10.215 (n, n)	0.1789	0.0027	<0.0001
200 m	10.017	0.083	10.144	10.352 (n, n)	0.6563	0.1789	0.0118
400 m	9.028	0.083	9.156	9.264 (n, n)	1.0000	0.0011	<0.0001
800 m	7.763	0.072	7.874	7.912 (y, y)	0.1789	0.6563	0.0748
1500 m	7.113	0.028	7.156	7.281 (n, n)	0.0011	0.1789	0.0002
5 km	6.3682	0.046	6.439	6.584 (n, n)	0.0011	0.6563	<0.0001
10 km	6.086	0.030	6.132	6.318 (n, n)	0.0011	0.0230	<0.0001
Marathon	5.492	0.030	5.538	5.595 (n, n)	0.00001	1.0000	0.0066
110 m hurdle	8.396	0.076	8.513	8.521 (y, y)	0.1789	0.1789	0.0016
400 m hurdle	8.449	0.039	8.510	8.551 (n, y)	n.a.	n.a.	0.3062
3000 m	6.125	0.042	6.191	6.306 (n, n)	0.1789	0.0230	<0.0001
Steeplechase	2.387	0.039	2.447	2.45 (y, y)	0.0004	0.8281	0.2659
High jump	5.921	0.116	6.100	6.14 (y, y)	0.00001	0.0230	0.0002
Pole vault	8.697	0.099	8.849	8.95 (n, y)	0.00001	0.6563	0.6365
Long jump	17.663	0.195	17.963	18.29 (n, n)	0.0011	0.1789	0.0085
Triple jump	22.385	0.392	22.989	23.12 (y, y)	0.1789	0.1025	0.2596
Shot-put	71.09	1.247	73.009	71.5 (n, n)	0.1789	0.0230	0.0765
Discus-throw							
Throwing							
the hammer	83.91	1.901	86.836	84.62 (n, n)	0.00001	0.6563	0.7083
Javelin throw*	93.436	6.394	103.280	98.48 (n, y)	1.0000	1.0000	0.4365

shows, the results for 11 of the 19 disciplines lie outside the $2\sigma_E$ interval around $\hat{x}_{\max}$ (fourth and fifth column).

As a check for trends we applied a linear trend analysis to the period 1980–1999, thus comprising both the reference and the estimation interval (last column, probabilities for $H_0: \rho = 0$, with ρ the slope of the regression line). In eight cases the development can be regarded as stationary, $p > 0.05$. As a further check we used the (non-parametric) median two-sample test, which does not assume that the underlying probability distribution is Gaussian. The values p_m , which are given in the sixth and seventh column, are a measure of the change in the median of the annual best results across the two respective intervals (1970–79 versus 1980–89 and 1980–89 versus 1990–1999). Comparing the two periods, it appears that stronger non-stationarities occur for the latter interval pair (nine events with $p_m > 0.1$), but they seem to decrease ($p_m > 0.1$ in 12 cases for the former interval pair). Given the increase in stationarity indicated by the median two-sample test one would rather expect an increase in the quality of the estimation, but on average that could not be confirmed. We found that an increase in variance is the reason for this apparent contradiction, analogously to the discussion of the estimations for the German Championship. While the average ratio of the standard deviations for the intervals 1973–1984 and 1985–1996 (which we calculated for comparative purposes) is nearly 1.0 it increases to 1.14 for the interval pair 1980–89/1990–99. Comparing the p_m s for two consecutive decades since 1930 averaged over all disciplines we found that the largest changes in performance occurred in the 1950s and 1960s (no corresponding Table or Figure).

Since a trend in the reference interval might lead to an artificially increased variance and thereby to deviations of the estimation as well as an increase in the width of the estimation interval we have repeated the analysis after a compensation for linear trends (using the commercial software package SAS). The mean μ has been replaced by the value of the regression line at the last (respectively latest) point in the reference interval; the standard deviation σ by the corresponding value provided by SAS. The differences in the estimation results have shown to be quite marginal. For the German championship, we found that the estimation for discus throwing, moved into the $2\sigma_E$ interval. For the annual best results, if no detrending is applied, five estimations are within the σ_E interval, four within the $2\sigma_E$ interval, and ten outside both. This changed to seven (within σ_E), one (within $2\sigma_E$) and 11 (outside both) in the case of (linear) detrending.

So far, we have used the estimation described in the previous section to predict the best result during a given period, but it can also be used to estimate the change of the record with time.⁴ Results obtained from applying this estimation to two world records are shown graphically in Figure 3. The observed development in the 110 m hurdle race competition is well compatible with purely stochastic changes, while for the 5 km race a clearly systematic progress is apparent.

In the German championships as well as in the worldwide best results, for all disciplines either the purely statistical model applies or an actual improvement is observed; no systematic decrease of the results (which would leave the record constant) could be observed.

Causes for Trends

The question of why trends occur in some disciplines but not in others cannot be answered by statistical considerations alone. Some of the apparent trends may again have a statistical basis, e.g. when the number of athletes increases (an effect which could also be described by our model). The effect of the increase in the number of participants is equivalent to a compression of the time axis: the statistical contribution to the increase becomes bigger. One

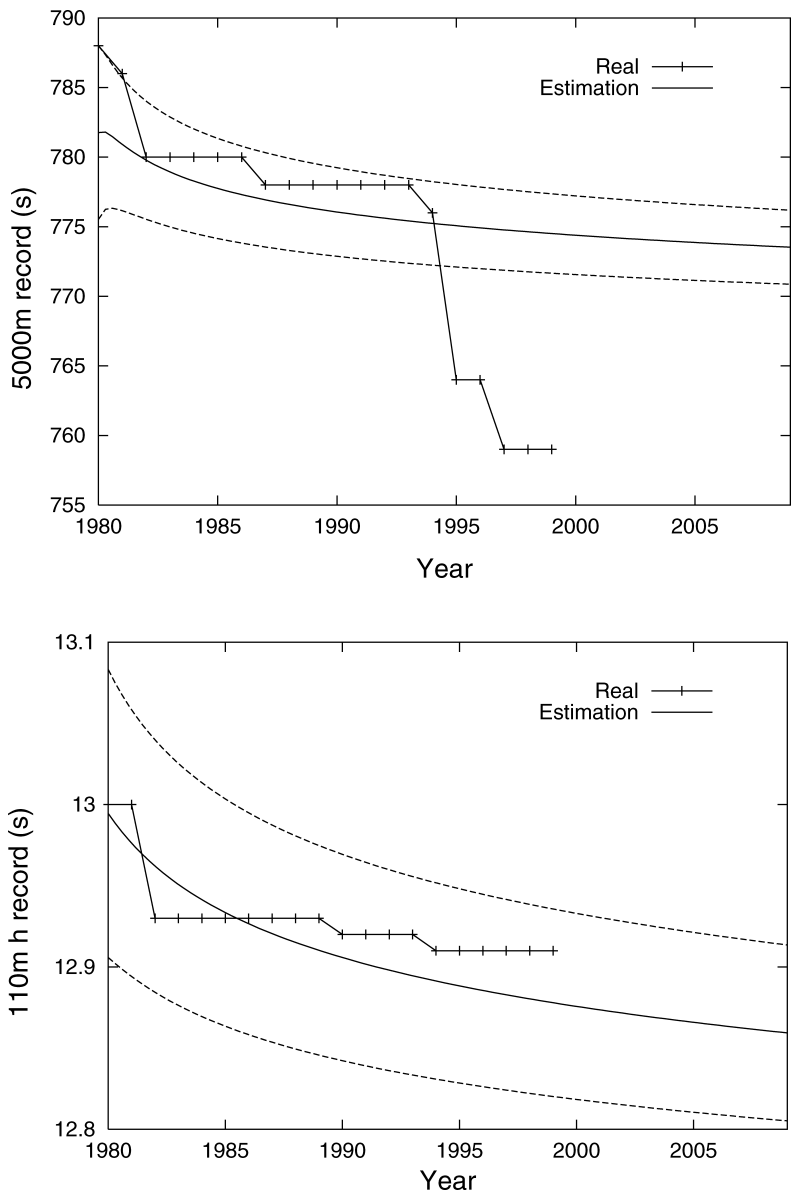


Figure 3. Record forecast obtained from equation (3). The upper panel shows the actual and the predicted world record development for the 5 km run, as an example for a non-stationary development, and the lower panel the same for the 110 m hurdle race. μ and σ have been obtained from the period 1980–1989

could expect this to have effected the Germany championships, due to the German reunification. As the observed stationarity suggests and as was also shown with the Chow test (see below) it had only a minor influence on the development (with the walking disciplines as exception). Due to the $\ln \ln$ dependence one would anyway not expect a strong change on the maximum performance (the number of inhabitants increased from about 60 million to about 80 million). For a more realistic modeling one would have to take into account the recruitment mechanisms, that result in the preselection of athletes for the championship.

This number (typically 10–30, and several 100 for the marathon run) can change independently from the population size. We also analysed data from the AAA championship in Great Britain (www.gbrathletes.co.uk), whose population was comparable to the one of Germany before reunification and did not have an analogous increase. There one also observes a very stationary evolution for the considered time interval 1973–1996 (striking exceptions are a long lasting trend for the triple-jump and even a negative (!) trend for the 5000 m and 10,000 m run).

On the worldwide scale, the increase in the overall population as well as higher participation from developing countries increases the sample sizes. Evidence for the contribution of those countries to the international performance evolution is provided by data from the Asian and African (biannual) championships (www.gbrathletes.co.uk) for the period 1973–2005 for the former and 1979–2004 for the latter, which show (positive) trends in most disciplines (although a flattening can be observed there as well).

True progress, which is due to improved training condition, the professionalization of many disciplines, or the elimination of psychological barriers is investigated thoroughly in specialized publications (see, for example, Hammond & Diamond, 1997; Grubb, 1998). In Grubb (1998), one can also find a description of a trend model and an estimation approach for the long-term development, but this approach requires the combination of results from different disciplines, which makes it more difficult to detect unusual developments in individual disciplines. The scaling (i.e. decrease) of performance in the running discipline with distance, a result of physiological limitations, has been studied by Savaglio & Carbone (2002).

Following a report (Jones, 2002) about our results (Gembris *et al.*, 2002) it was suggested in a letter to the editor that the observed sudden improvements in the mid-range running disciplines in the 1990s might be due to the use of new doping methods. We found further evidence of doping using the Chow test, which allows us to detect structural breaks in the data: the method divides the time series into two intervals and fits each segment with a line. It is then tested whether the difference between both lines is statistically significantly or not. If the significance values are low for all subdivisions this would be a hint that the trend is not adequately described by a linear model (which we observed in five cases for the German championship, and in four for the annual best results).⁵ Qualitative changes, characterized by a temporal decrease in performance, can be observed for the years 1988 and 1998 (with a small data basis though). The change in 1988 could be found both for the German championship and the international best results. This observation coincides with the prohibition of blood doping in that year. In 1998, an obligatory measurement of the hematocrit level was introduced. For the German results we found this effect in 1988 and the following years for the disciplines 100 m, 200 m, and 1500 m and for the disciplines 200 m, 400 m, 1500 m, 5 km, marathon, 110 m hurdle on the international level (based on a visual inspection one should also include long jump and 10 km).

Environmental factors might also have a systematic effect. In Jánosi & Bántay (2002) the influence of the geographical position of earth on the outcome of throwing events has been investigated, but this effect could not be identified in their data.

Discussion and Conclusion

As we have shown, apparent advances in athletics records can often be traced to purely statistical effects. Even in the absence of true progress, the increase in the sample size (number of competitions) implies that the record will improve. Potential reasons for non-systematic changes could be weather (wind, rain, temperature) or a bunching of sporting events in certain years: in many cases top athletes do not take part in national competitions,

because they want to save strength for international events or for the qualification of them. Tactical behavior can also have an effect, e.g. when athletes first run slowly and keep the decision of whether to sprint for the end of the race.

A comparison of retrospective 'predictions' of record changes with actual data allowed us to distinguish between disciplines in which true progress occurred from others where the evolution can be explained purely statistically. Our analysis of past performances showed that in the German championships, the majority of the disciplines exhibited little inherent progress. The number of disciplines in which true progress could be observed was significantly larger in the worldwide competitions. A more detailed analysis shows that the trends ('=true progress') appear to become smaller with time. An inspection of the names of the athletes who achieved the annual best performance or won the German championship did not reveal 'super athletes' dominating the development in a certain discipline. An exception is the performance of the German discus thrower Lars Riedel (a former GDR athlete) and the athlete Robert Ihly who dominated the 20 km walking discipline for years. The fact that the annual best results of any two successive years was hardly achieved by the same athlete provides evidence for the independence of the data. A further inspection revealed that former GDR athletes dominated the 20 km walking and 50 km walking disciplines in the championships immediately following the reunification.

We now turn to previous studies on the estimation of records in athletics: a work that discusses the effect of statistics on apparent progress internationally goes back to Tryfos & Blackmore (1985). They used their analysis to make predictions for three running disciplines that have also been analyzed by us: the 5 km, the 10 km and the marathon run (and in addition the 1000 m, 1 mile and 20 km run). In contrast to them, our analysis uses data from each year (rather than only record breaking events). This approach provides better estimates of the relevant statistical parameters and therefore more reliable forecasts of those changes that are due to the statistical effect. From a retrospective view, the running disciplines considered by Tryfos & Blackmore (1985) showed clear trends. As a result, their predictions significantly underestimated the actual evolution.

In Hernig & Klimmer (1980), a differential equation for the prognosis of records was presented that describes how records approach a given physiological performance limit by extrapolating from previous trends. A major difference to our approach is the phenomenological character of the equation, the larger number of fit parameters and the lack of an explicit formula for confidence intervals. In that paper, longer term developments are also analyzed.

In Barao & Tawn (1999), data from two disciplines, the women's 1500 m and 3000 m runs, have been analyzed using an alternative extremal (bivariate) analysis. The motivation was to question an unexpectedly good result in the 3000 m run. A further alternative approach for the analysis of record data has been described in Solow & Smith (2005). It presents a measure for quantifying the change by a new record in relation to the difference between the old record and the previous second best result. The analysis of the 1500 m and 3000 m runs has also been subject of Robinson & Tawn (1995) and the follow-up discussions in Smith (1997) and Robinson & Tawn (1997). For Robinson & Tawn (1995) it is central that the annual best results are the outcome of (asymptotic) order statistics. Our work, instead, aims at estimating maxima and minima in successive years (which could be denoted as second-order order statistics, since maxima of maxima are estimated). Robinson & Tawn also describe an elaborate model for trends. A main result of Smith (1997) is an equation that predicts the minimum of a set of results for a given N under the assumption of a Weibull distribution. We instead assume that the parent distribution is Gaussian, which represents a case of general importance.

While we have chosen to apply the method to the analysis of German athletics results, application to other athletics events should be straightforward. Rare events, such as the Olympic games, provide a smaller data basis, making it difficult to find a good estimate for the statistical parameters μ and σ . For data from other events, we refer the reader to www.gbrathletes.co.uk and www.leichtathletik-dgld.de.

Outside the world of sports, we employed our approach to estimate the body height of the tallest person of a population as a function of population size (see Figure 4). The data represent the national records of the countries Denmark, France, Germany, Netherlands, UK and USA (for the year 1998, see Table 3, obtained from different editions of the Guinness Book of Records, 1998, 1999) and the height of the tallest German draftee of the year 1996 (according to mustering data of the Bundeswehr). For the illustrative estimation in Figure 4 we used the parameters $\mu = 180$ cm and $\sigma = 8$ cm according to column four, where σ has been chosen such that about an equal number of points are above and below the estimation

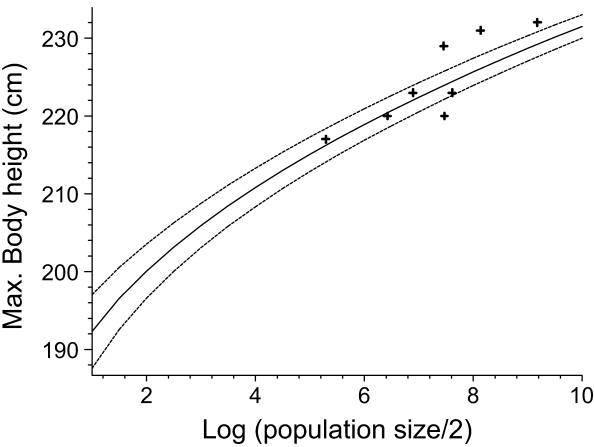


Figure 4. Estimated maximum bodysize for male populations of different sizes and the uncertainty of the estimation given by the ‘errortube’ (σ_E intervals). For the data points shown see Table 3. The estimation assumes that ρ is a Gaussian probability density function with parameters $\mu = 180$ cm and $\sigma = 8$ cm

Table 3. Data about the current records and mean values of the height of humans in different countries. The record values stem from different national editions of the *Guinness Book of Records*, 1998, 1999 where the year for which it was printed is given in brackets. This data has been used for the generation of Figure 4 (*: Group of 19 years old (male) mustering-candidates)

Country	Population size (july 1997) [CIA world fact book]	maximum body height	mean body height mean $\pm$ S.D.) [Grandjean, 1988]
Denmark	5, 305, 048	2.20 m (1998)	–
France	58, 609, 285	2.20 m (1988)	–
Germany	82, 071, 765	2.23 m (1998)	1.745 $\pm$ 0.072
(sub-set)*	193, 350	2.17 m (1996)	1.798 $\pm$ 0.069
Netherlands	15, 649, 729	2.23 m (1998)	1.825 $\pm$ 0.075
UK	57, 591, 677	2.29 m (1998)	–
USA	267, 954, 764	2.31 m (1999)	1.755 $\pm$ 0.071
World	5.8e9	2.32 m (1998)	–

curve. For a mathematically rigorous procedure for obtaining the model parameters more data points would be desirable.

Our approach of estimating the temporal development of maxima and minima would be useful for the definition of thresholds in the assessment of the relevance of small signals in a noisy environment (Gembris, 2005). In particular, we plan to apply this method to the analysis of brain activation in functional magnetic resonance (fMRI) (Kwong, 1995; Moonen & Bandettini, 2000). A main objective in this context is the detection of the onset of activation in 'real time' (Gembris, 2000). A neighboring possible field of application is that of 'stochastic resonance' (Gammaitoni *et al.*, 1998).

The approach should also be useful to obtain estimations for the quality of the best solutions of NP hard problems (employing density-of-states formalisms), like the traveling salesman problem (Beardwood *et al.*, 1958) and the optimization of road networks (Schweitzer *et al.*, 1997). An application to stock-market analysis also appears feasible in principle, either directly or indirectly: we verified that the prerequisite of stationarity is fulfilled at least for some second-line stocks. In these cases, the estimation could be part of a strategy for determining the best point in time of buying and selling stocks. For non-stationary time series it was proposed to use the so-called log-daily returns to obtain stationary data (Coles, 2004).

Our formulae imply that the effect of order statistics can be identified by plotting univariate data as a function of $\ln \ln x$, which might be relevant, for example, for the study of evolutionary processes.

Acknowledgements

We thank Professor Peter Grassberger (John von Neumann Institute of Computing, Research Center Jülich), Dr Wolfgang Meyer (statistician at the Central Institute of Applied Mathematics, Research Center Jülich) for useful discussions about this work, and Professor Paul Jansen (Central Institute of Applied Mathematics, Research Center Jülich) for technical support. From Günter Hernig (DGLD, www.leichtathletik-dgld.de) we received helpful information for the discussion part. We also thank the Guinness publisher in London, especially Mrs Amanda Brooks, for providing us with the contact coordinates of several national Guinness publishers and their respective employees who processed our requests. We are grateful to the base of the German Bundeswehr in Remagen for sending us the mustering data regarding the body height. D.G. is supported by the 'Elite support program for postdocs 2003' of the Landesstiftung Baden-Württemberg, Germany.

Note: supplementary material is available from the author Daniel Gembris on request. A recent paper that might also be of interest for the reader is: Einmahl, J. H. J. & Magnus, J. R. (2006) Records in athletics through extreme-value theory, Discussion paper 83, Tilburg University, Center for Economic Research. Available online at: <http://ssrn.com/abstract=936923>.

Notes

¹The formula for the minimum is identical to equation (2), except that the inner integral runs from x' to infinity.

²In other words, we have chosen the normal distribution as the probability density function: $\rho(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, with μ denoting the mean and σ the standard deviation.

³Results from the time before 1973 have been affected too strongly by trends and have therefore been excluded from our analysis. From 1973 on, recordings also have a higher accuracy, indicating both technological progress and the need for a better discrimination between the best and the second best results, or more

generally between two consecutive ranks. Since the athletic results of women show a trend until the early 1980s, they have not been considered either. Our analysis has been performed for all but one competition. The exception concerns Javelin-throwing, for which the regulation was modified in 1985.

⁴For the record estimation, the value of N has been determined that generates the record for the year 1980 from the μ and σ for the period 1980–1989. This N , N_0 , is then used as an offset in the estimation for the period 1990–2009 ($N' = N_0 \dots N_0 + 19$).

⁵A visual inspection of the time series revealed ‘outliers’ for the following disciplines: hammer (1993, German), 1500 m (1990), 5 km (1999), 10 km run (1999), 110 m hurdle (1981), shot put (1989, 1994) (for the world wide annual best results).

References

- Barao, M. I. & Tawn, J. A. (1999) Extremal analysis of short series with outliers: sea-levels and athletics records, *Journal of the Royal Statistical Society C-App.*, 48, pp. 469–487.
- Beardwood, J., Halton, J. H. & Hammersley, J. M. (1958) The shortest path through many points, *Proceedings of the Cambridge Philosophical Society*, 55, p. 299.
- CIA World factbook (2005) <http://www.odci.gov/cia/publications/factbook/index.html>.
- Coles, S. (2004) *An Introduction to Statistical Modeling of Extreme Values* (London: Springer).
- David, H. A. (1970) *Order Statistics*, ISBN 471-19675-4 (New York: Wiley).
- Galambos, J. (1978) *The Asymptotic Theory of Extreme Order Statistics* (New York: Wiley).
- Gammaitoni, L., Hänggi, P., Jung, P. & Marchesoni, F. (1998) Stochastic resonance, *Reviews of Modern Physics*, 70, pp. 223–287.
- Gembris, D. (2005) *Patentschrift* DE 198 21 468B4.
- Gembris, D., Taylor, J. G., Schor, S., Frings, W., Suter, D. & Posse, S. (2000) Sliding-window correlation analysis with reference vector optimization for functional magnetic resonance imaging in real time (FIRE), *Magnetic Resonance in Medicine*, 43(2), pp. 259–268.
- Gembris, D., Taylor, J. G. & Suter, D. (2002) Sports statistics-trends and random fluctuations in athletics, *Nature*, 417, p. 506.
- Grandjean, E. (1988) *Fitting the task to the Man*, ISBN 0-850-66192-7 (Taylor & Francis).
- Grubb, H. J. (1998) Models for comparing athletic performances, *Journal of the Royal Statistical Society D-Statistics*, 47, pp. 509–521.
- Guinness Book of Records* (1998, 1999) (Guinness Publishers).
- Hammond, K. A. & Diamond, J. (1997) Maximal sustained energy. budgets in humans and animals, *Nature*, 386(6624), pp. 457–462.
- Hernig, G. & Klimmer, H. (1980) *Theorie und Praxis der Körperkultur*, 29(6), pp. 404–412. <http://www.gbrathletes.co.uk>.
- <http://www.leichtathletik-dgld.de/>.
- Jánosi, I. M. & Bántay, P. (2002) Statistical test of throwing events on the rotating Earth, *European Physical Journal B*, 30, pp. 411–415.
- Jones, N. (2002) *New Scientist*, 174, p. 12.
- Koponen, M. (2000) The track and field athlete of the century according to the statistics. CD, <http://www.atkpike.com/Athlete.htm> (Atk-Pike, Asikkala, Finland)
- Kwong, K. K. (1995) Functional magnetic resonance imaging with echo planar imaging, *Magnetic Resonance Quarterly*, 11(1), pp. 1–20.
- Moonen, C. T. & Bandettini, P. A. (2000) *Functional MRI (Medical Radiology)* (New York: Springer).
- Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B.P. (1995) *Numerical Recipes in C, The Art of Scientific Computing*, 2nd edn (New York: Cambridge University Press).
- Robinson, M. E. & Tawn, J. A. (1995) Statistics for exceptional athletics records, *Applied Statistics - Journal of the Royal Statistical Society C*, 44(4), pp. 499–511.
- Robinson, M. E. & Tawn, J. A. (1997) Letters to the Editor, statistics for exceptional athletics records, *Applied Statistics - Journal of the Royal Statistical Society C*, 46(1), pp. 127–128.
- Roberts, S. J. (2000) Extreme value statistics for novelty detection in biomedical signal processing, *IEEE Proceedings-Science, Measurement and Technology*, 147(6), pp. 363–367.
- Savaglio, S. & Carbone, V. (2002) Human performances: scaling in athletic world records, *Nature*, 404, p. 244.
- Scherer, K. A. (1997) *Chronik Handbuch, Sportgeschichte in Zahlen (History of Sports in Numbers)* (Munich: Chronik Verlag).
- Steinmetz, F. (2000) *Deutsche Meisterschaften, 1898–1999, Band 1-5* (G. Hornberger).

- Schweitzer, F., Ebeling, W., Rosé, H. & Weiss, O. (1997) Optimization of road networks using evolutionary strategies, *Evolutionary Computation*, 5(4), pp. 419–438.
- Solow, A. R. & Smith, W. (2005) How surprising is a new record? *American Statistician*, 59(2), pp. 153–155.
- Smith, R. L. (1997) Letters to the Editor, Statistics for exceptional athletics records, *Applied Statistics – Journal of the Royal Statistical Society C*, 46(1), pp. 123–127.
- Tryfos, P. & Blackmore, R. (1985) Forecasting records, *Journal of the American Statistical Association*, 80, pp. 46–50.

Appendix

Estimation of the Minimum and the Variance, and the Effect of ‘Prehistoric’ Records

In this section we first state the equation for the estimation of the minimum, which is analogous to equation (2), and then present an equation for the variance of the extrema distribution.

The equation for the minimum corresponding to equation (2) reads as follows:

$$E\left(\min_{i=1}^N(x_i)\right) = N \int_{x'=-\infty}^{\infty} x' \left(\int_{x''=x'}^{\infty} \rho(x'') dx'' \right)^{N-1} \rho(x') dx' \quad (5)$$

$$= N \int_{x'=-\infty}^{\infty} x' \left(1 - \int_{x''=-\infty}^{x'} \rho(x'') dx'' \right)^{N-1} dx'. \quad (6)$$

Back to the case of the maxima, one finds that equation (2) can be reformulated as

$$\hat{x}_{\max} = E\left(\max_{i=1}^N(x_i)\right) = \int_{x'=-\infty}^{\infty} x' \tilde{\rho}(x') dx' \quad (7)$$

using the definition of a new probability density function for the distribution of the maxima

$$\tilde{\rho}(x') = N \left(\int_{x''=-\infty}^{x'} \rho(x'') dx'' \right)^{N-1} \rho(x'). \quad (8)$$

The function $\tilde{\rho}$ is a normalized probability density function as may be shown by a single integration by parts, and describes the distribution of the maximum values of the N random variables x_i . We want to state explicitly here, that no assumptions on ρ enter the derivation of equation (2) and equations (6) and (7). Therefore these equations are true in general. Eventual discrepancies between theory and experiment could always be reduced to the choice of ρ or its parameters.

For the variance of the maxima distribution one can immediately show that

$$E\left(\left(\max_{i=1}^N(x_i)\right)^2\right) = \int_{x'=-\infty}^{\infty} x'^2 \tilde{\rho}(x') dx'. \quad (9)$$

Together with equation (7) this can be used to determine the standard deviation of the estimated value, σ_E , which we use as measure for the reliability of the estimation.

For the situation that a record r of the ‘prehistory’ of interval A should be taken into account, the following equation holds (analogously in the case of the minimum):

$$\hat{x}_{\max} = \int_{-\infty}^{\infty} \max(r, x) \tilde{\rho}(x) dx, \tag{10}$$

or resolving the max function:

$$\hat{x}_{\max} = \int_{-\infty}^r r \tilde{\rho}(x) dx + \int_r^{\infty} x \tilde{\rho}(x) dx. \tag{11}$$

We have considered the effect of r on the record estimation for Figure 2 instead as described in Note 4. It was verified that both approaches give almost identical results.

Confidence Levels

We next present our computational results for the probability for random values to fall outside the estimated confidence interval around the estimated maximum. The probabilities that the maximum of one set of random elements exceeds the value of $\hat{x}_{\max} + \sigma_E$ or $\hat{x}_{\max} + 2\sigma_E$ are denoted by p_{σ} and $p_{2\sigma}$. A calculation of these probabilities is necessary, since extremal distributions are non-Gaussian in general, and the probabilities can therefore not a priori be expected to coincide with those of a Gaussian function. For a given mean, number of random elements and standard deviation σ , the corresponding estimates $\hat{x}_{\max}$ and σ_E can be calculated using equations (2) and (9), evaluated by a Riemann sum (for reliable numerical results the summands have to be sorted before the addition). This numerical approach is not restricted to the Gaussian case, but can be applied regardless of the distribution.

The probability that the maximum of a set of random values exceeds the estimated maximum $\hat{x}_{\max}$ by σ_E resp. $2\sigma_E$ is, for $N = 12$ (and arbitrary μ and σ), $p_{\sigma,u} = 0.155$ respectively $p_{2\sigma,u} = 0.033$. The probabilities for the cases that the true maximum is smaller than the estimate by more than σ_E or $2\sigma_E$ are $p_{\sigma,l} = 0.154$ and $p_{2\sigma,l} = 0.011$. These values are close to those obtained from a Gaussian with $p_{\sigma,l} = p_{\sigma,u} = 0.159$ and $p_{2\sigma,l} = p_{2\sigma,u} = 0.023$. This suggests that the maxima distribution is well described by a Gaussian function for small N (but this similarity is not explicitly utilized).

The Approximation Formula (Equation (3))

A central theorem of extreme-value statistics (Galambos, 1978) says that for certain distributions of random variables X_i , among them the Gaussian distribution, the following relation holds:

$$\lim_{N \rightarrow \infty} P(Z_N < a_N + b_N x) = H_{3,0}(x), \tag{12}$$

where P denotes probability, $Z_N = \max(X_1, X_2, \dots, X_N)$ and

$$H_{3,0}(x) = \exp(-e^{-x}) \tag{13}$$

is the cumulative density function of the asymptotic maxima distribution. For the coefficients, one finds in the case of Gaussian distributed X_i [Galambos 78]

$$a_N = (2 \ln(N))^{1/2} - 1/2 \frac{\ln(\ln(N)) + \ln(4\pi)}{(2 \ln(N))(1/2)} \tag{14}$$

and

$$b_N = (2 \ln(N))^{-1/2}. \tag{15}$$

The desired expectation value is then given by

$$E(\max_{\rho}) = a_N E(\max_{dH}) + b_N = 0.57722 \cdot a_N + b_N, \quad (16)$$

where $E(\max_{dH})$ represents the expectation value of the asymptotic maximum distribution ($\frac{\partial}{\partial x} H_{3,0}$). This already represents an analytical expression of the expected maximum for arbitrary N (approximation deteriorating for small N). Expanding equation (16) in $x = \ln(\ln(N))$ around $N = 24$ gives, for $\mu = 0$ and $\sigma = 1$,

$$\begin{aligned} E(\max_{\rho}) = & 0.9273 + 0.5981 \cdot x + 0.2901 \cdot x^2 - 0.0016 \cdot x^3 \\ & + 0.0078 \cdot x^4 + 0.0003 \cdot x^5 + O(x^6). \end{aligned} \quad (17)$$

For small N and thus small x ($x \approx 1$), the first three terms are clearly dominating. It is therefore possible to drop the higher order terms, which results in equation (3) given above. This approximation remains valid for $N \lesssim 10^9$. Since equation (16) holds in the limit of large N , it is advantageous not to choose the coefficients of the $\ln \ln$ approximation formula (equation (3)) as the expansion coefficients, but to determine them by means of a fitting procedure (see parameters a_i).

An approximate standard deviation for the maximum estimate is obtained by scaling the standard deviation of $\frac{\partial}{\partial x} H_{3,0}$ ($= 1.2825$) by b_N .