Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Journal of Computational and Applied Mathematics 169 (2004) 227–234

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.com/locate/cam

On survival times of sport records

Farzad Noubary[a,*], Reza Noubary[b]

[a] Department of Statistics, Penn State University, University Park, PA 16802, USA
[b] Department of Mathematics, Bloomsburg University, Bloomsburg, PA 17815, USA

Received 11 June 2003; received in revised form 27 November 2003

Abstract

Survival of sport records is investigated assuming that the number of attempts to break a record is governed by a non-homogeneous Poisson process. Explicit formulae for two practical cases are derived, and their applications are demonstrated using an example.

Keywords: Sports; Records; Survival times; Non-homogeneous Poisson process

1. Introduction

Let $R > 0$ and $S > 0$ be two random variables with respective distribution functions $F_R(.)$ and $F_S(.)$. Suppose R, the record in a given sport, is subject to set of events (attempts) S. Then the record breaks if the value of S exceeds (subceeds) R. The value of S is a function of the type of sport, number of participants, prize, training, environmental factors such as temperature, altitude, etc., and factors important to the athletes and the public. The value of R depends on factors such as the type and popularity of the sport, amount of rewards or prizes, number of formal competitions, etc. The probability of breaking a record is then

$$P(S > R) = p = 1 - \int_0^\infty F_S(x)\,\mathrm{d}F_R(x),$$

where p is the probability of breaking a record in a single attempt. (Note that the calculations presented here can easily be modified for sports where breaking a record corresponds to $S < R$, e.g., 100 m dash.) When applying this model, one is frequently interested in the probability of breaking a record in a specified interval, say $(0, t]$, where 0 represents the beginning of the period. Assuming

* Corresponding author.
E-mail address: farzad@psu.edu (F. Noubary).

0377-0427/$ - see front matter
doi:10.1016/j.cam.2003.12.022

that T is the length of time a record is (or will be) held, the probability of record being broken in the time interval $(0,t]$, denoted by $F_T(t)$, can be obtained as

$$F_T(t) = P(T \leqslant t) = 1 - P(T > t) = 1 - L_T(t),$$

where $L_T(t)$ is the survival function defined as $L_T(t) = P(T > t)$, $L_T(0) = 1$. If R is a record subject to a sequence of attempts $S_1, S_2, \ldots, S_n$, then $L_T(t)$ is given by

$$L_T(t) = \sum_{r=0}^{\infty} P(N(t) = r)\bar{P}(r), \tag{1}$$

where $\{N(t), t \geqslant 0\}$ is a general counting process of attempts occurring randomly in time and $\bar{P}(r) = P(\max(S_1, S_2, \ldots, S_r) < R)$, $r = 1, 2, \ldots, n$, with $\bar{P}(0) = 1$.

If we assume further that the occurrence of the attempts is governed by a homogeneous Poisson process with rate λ, then from (1) we obtain

$$L_T(t) = \sum_{r=0}^{\infty} (e^{-\lambda t}(\lambda t)^r / r!)\bar{P}(r). \tag{2}$$

(Note that $\bar{P}(r)$ can also be considered the probability of surviving the first r attempts.) If a further assumption is made that the attempts are independent and identically distributed random variables, then (2) reduces to

$$L_T(t) = \sum (e^{-\lambda t}(\lambda t)^r / r!)(1 - p)^r = \exp(-\lambda t p). \tag{3}$$

Thus, if the mean rate of attempts and the period of interest are given, then $L_T(t)$ can be calculated for any p. Hence the main problem for the situation described above is that of estimating the p, i.e. the probability of breaking a record in a single attempt. As a crude approximation, one may estimate p as a ratio of the number of records set to the number of attempts made.

Now consider a more general situation in which:

(1) a record is assumed to be a random variable with distribution function $F_R(.)$;
(2) the occurrences of attempts follow a stochastic point process $\mathbf{P}$ with a counting process $\{N(t), t \geqslant 0\}$ which develops over time and governs the values of attempts $\{S_n, n = 1, 2, \ldots\}$.
(3) $\mathbf{P}$ and $\{S_n\}$ are independent.

Here S_n denotes the nth independent and identically distributed random variable, realized at the moment $N(t)$ first reaches n. (S_n is the measured value of the nth attempt.) Note that for this situation breaking a record corresponds to the occurrence of the first record attempt (new high) in the S-sequence relative to the reference value R in the sense of the definition given below. In this approach, effects of other factors could be incorporated into $\mathbf{P}$ by increasing or decreasing the number of attempts in a given period. This could be done, for example, by using expert opinion regarding the importance or value of factors such as diet, increase in prize monies, new technology, etc. in terms of number of participants.

Definition. A first record with respect to R occurs at $t = t_{n_1}$ if (a) $S_i < R$ for all $i < n_1$, (b) $S_{n_1} > R$, in which case $T = t_{n_1}$ = the first record time, and $S_{n_1} = S_{(1)}$, the first record value. Alternatively,

$$T = \inf\left\{t : \max_{n \leqslant N(t)} S_n > R\right\}, \quad S_{(1)} = S_{N(T)}.$$

The rest of this paper will focus on calculating record breaking probabilities when the occurrences of attempts are governed by **P**, a Poisson process with time-dependent rate. A Poisson model seems reasonable since the number of attempts in two different time intervals or two different locations may be considered independent. Moreover, the probability of breaking a record may be assumed to be proportional to the number of attempts. Also, the probability of setting more than one record in a small number of attempts is negligible. Finally, because the number of attempts during different times of the year may be different (e.g. summer vs. winter, Olympic year vs. non-Olympic year), a Poisson process with time-dependent rate seems more appropriate. Furthermore, a time-dependent rate takes into account the change in the number of attempts over time.

2. Record survival

Suppose that **P** is Poisson with time-dependent rate $\lambda(t) > 0$ and let

$$\Lambda(t) = \int_0^t \lambda(u)\,\mathrm{d}u.$$

If $F_R(.)$ denotes the distribution function of R and $F_S(.)$ the distribution function of $\{S_n, n = 1, 2, \ldots\}$, then since $(T > t)$ if and only if $\max(S_1, S_2, \ldots, S_n) < R$, the required probability $P(T > t)$ is given by

$$P(T > t) = \int_0^\infty \sum_{n=0}^{\infty} \mathrm{e}^{-\Lambda(t)} \frac{(\Lambda(t))^n}{n!} (F_S(x))^n \,\mathrm{d}F_R(x). \tag{4}$$

(Note that $(F_S(.))^n$ is the distribution function of $\max(S_1, S_2, \ldots S_n)$.] Expression (4) can also be written as

$$P(T > t) = \int_0^\infty \exp[-\Lambda(t)(1 - F_S(x)]\,\mathrm{d}F_R(x). \tag{5}$$

If $R = R_0$ is given (e.g. R_0 is the present record), then

$$P(T > t \,|\, R = R_0) = \exp[-\Lambda(t)(1 - F_S(R_0))]. \tag{6}$$

Now, it is clear that for the general case, (5) will require knowledge of both $F_R(.)$ and $F_S(.)$. However, there are two important cases discussed below where calculations can be carried out with less information and without numerical integration. Taking the viewpoint that the strength of a record in a given sport is measured by the number of attempts required to break it, these cases and the assumptions made are reasonable.

Case A. Suppose that **P** has been observed throughout the time interval $(-\tau, 0]$, where 0 represent the present time. Suppose also that the largest value in this interval is used as a reference for determining further records. Then

$$F_R(x) = P(R < x) = \sum_{n=0}^{\infty} P[\max(S_1, S_2, \dots, S_n) < x \mid N(\tau) = n]P(N(\tau) = n)$$
$$= \sum_{n=0}^{\infty} e^{-\Lambda(\tau)} \frac{(\Lambda(\tau))^n}{n!} (F_S(x))^n = \exp[-\Lambda(\tau)(1 - F_S(x))] \tag{7}$$

and application of (5) yields

$$P(T > t) = \Lambda(\tau)[1 - \exp(-(\Lambda(\tau) + \Lambda(t)))]/[\Lambda(\tau) + \Lambda(t)]. \tag{8}$$

With confidence given by the right-hand side of (8), there will be no new maximum in $(0, t]$ greater than the one in $(-\tau, 0]$. Thus the survival probability depends only on the rate of attempts. This makes sense since many researchers attribute the breaking of records not only to individual attempts, but also to the total number of attempts. The total number of attempts, in turn, can be attributed to the increase in world population in general and the increase in population of athletes or participants in particular (see Section 4). Note that since $D(t) = \Lambda(\tau) + \Lambda(t)$ is non-decreasing, for any $h > 0$:

$$E(T) = \int_0^{\infty} P(T > t) = \int_0^{h} P(T > t) + \int_h^{\infty} P(T > t)$$
$$= \int_0^{h} \frac{\Lambda(\tau)}{D(t)} [1 - \exp(-D(t))] \, dt + \int_h^{\infty} \frac{\Lambda(\tau)}{D(t)} [1 - \exp(-D(t))] \, dt,$$

which implies that

$$\Lambda(\tau)[1 - \exp(-\Lambda(h))] \int_h^{\infty} \frac{dt}{D(t)} \leqslant E(T) \leqslant \Lambda(\tau) + \Delta(\tau) \int_h^{\infty} \frac{dt}{D(t)}.$$

Thus

$$E(T) < \infty \quad \text{if} \quad \int_h^{\infty} \frac{dt}{D(t)} < \infty.$$

It is interesting to observe that for a homogeneous Poisson process, the expected time for a record to break is ∞. Recall that this is a well-known result in the theory of records for independent and identically distributed random variables.

Case B. This case is similar to case A except that here the exact number of attempts k that occurred in the past is known (but the time span covered may not be known). Here we are interested in breaking the mth record ($m \leqslant k$). For example, in a certain competition we may be interested in calculating the probability that a participant will break the third best world record. For this case we have

$$F_R(x) = \sum_{j=k-m+1}^{k} \binom{k}{j} [F_S(x)]^j [(1 - F_S(x)]^{k-j},$$

$$\mathrm{d}F_R(x) = m \binom{k}{m} [F_S(x)]^{k-m} [1 - F_S(x)]^{m-1}\, \mathrm{d}F_S(x)$$

and

$$P(T > t) = \int_{-\infty}^{\infty} \exp[-\Lambda(t)(1 - F_S(x))]\, \mathrm{d}F_R(x)$$

$$= m \binom{k}{m} \int_0^1 \exp[-\Lambda(t)u](1-u)^{k-m} u^{m-1}\, \mathrm{d}u.$$

For example, for $m = 1$ we have $F_R(x) = (F_S(x))^k$ and therefore

$$P(T > t) = k \int_0^1 \exp[-\Lambda(t)u](1-u)^{k-1}\, \mathrm{d}u$$

$$= k \frac{\exp[-\Lambda(t)]}{[\Lambda(t)]^k} \int_0^{\Lambda(t)} u^{k-1} \mathrm{e}^u\, \mathrm{d}u. \tag{9}$$

It is interesting to note that this case also covers the situation where the Lehmann alternative is satisfied by $F_R(.)$ and $F_s(.)$, That is,

$$F_R(x) = (F_S(x))^h \quad \text{or} \quad 1 - F_R(x) = (1 - F_S(x))^h$$

for some h. Here no reference is made to the past events, but R is assumed to behave as $\max(S_1, S_2, \ldots, S_h)$.

3. Examples

In an interesting paper Berry [1] has discussed the effect of population increase on the breaking of sports records. He has introduced the following exponential model for the growth of the world's male population:

Population in year $t = 1.6 \exp[0.0088(t - 1900)]$.

Let us assume that the number of attempts is proportional to the population size at time t. Then using (8) we have the following results:

(a) The best record of a period of length 100 years has 80% chance of surviving an additional 10 years.
(b) The best record of a period of length 50 years has 65% chance of surviving an additional 10 years.

(c) The best record of the period of length 10 years has 23.6% chance of surviving an additional 10 years.

Suppose now that attempts occur randomly throughout $(-\tau, t]$. If n_1 attempts occur in the interval $(-\tau, t]$ and n_2 attempts to occur in the interval $(0, t]$ then the probability of no record in $(0, t]$ is

$$P(\max(S_1, S_2, \dots, S_{n_1}) = \max(S_1, S_2, \dots, S_{n_1+n_2})) = \frac{n_1}{n_1 + n_2}.$$

Then corresponding to (a), (b), and (c) above the 10 years survival probabilities are, respectively, $\frac{100}{110} = 91\%$, $\frac{50}{60} = 83\%$, and $\frac{10}{20} = 50\%$.

Note that the record of the last 10 years may or may not be the same as the record of the last 20 years. This is one reason for the reduction in survival probability.

The example used above served as a demonstrating example. A more realistic situation should consider a model for the population of participants or even the population of participants who have the potential to break records in a given sport. Good examples of this include participants who qualify for the Olympics or participants who qualify for the Boston Marathon. Table 1 (see also Fig. 1) presents data regarding the times of the Boston Marathon together with the number of participants for the period 1970–2003. We chose 1970 because this was the year during which a qualifying time was introduced for participation. Using regression, the model given below was found to present the number of participants as a function of the year ($R^2 = 93.8\%$)

$$\text{Number of participants in year } t = -1294 + 1088t - 57.5t^2 + 1.25t^3.$$

In this model, the data for the year 1996 was replaced by the average of the two neighboring times since this was the 100th running of the Boston Marathon and more than 38,000 runners were allowed to participate.

Using this model the survival probabilities for the next 5 (2003–2008) and 10 (2003–2013) years are respectively

$$P(T > 5) = 0.632 \quad \text{and} \quad P(T > 10) = 0.422.$$

Moreover,

$$P(T > 10 \mid T > 5) = 0.667.$$

4. Concluding remarks

In the analysis presented above, the number of attempts was used in a broad sense as the only factor contributing to the breaking of records. Similarly, Berry [1] used the male population of the world as the only predictor for Olympic winning times in several events. Using the coefficient of determination as a measure of fit, Berry found, for example, that $R^2 = 81.3\%$ for the Olympic winning times in the 100 m dash. For other events he found an R^2 value as high as 95.4%. One should expect even better results if the male population of the world is replaced in Berry's analyses by the population of participants or the number of attempts, since these are more precise measures of how many times a record is challenged. As a result, the number of attempts, used in our analyses, is a reasonable predictor to consider. Of course, there are many other factors that are also

Table 1
Data for Boston Marathon 1970–2003

Year	Winner	Time	Time (min)	Number of participants
1970	Ron Hill	2:10:30	130.50	1174
1971	Alvaro Mejia	2:18:45	138.75	1067
1972	Olavi Suomalainen	2:15:39	135.65	1219
1973	Jon Anderson	2:16:03	136.05	1574
1974	Neil Cusack	2:13:39	133.65	1951
1975	Bill Rodgers	2:09:55	129.92	2395
1976	Jack Fultz	2:20:19	140.32	2188
1977	Jerome Drayton	2:14:46	134.77	3040
1978	Bill Rodgers	2:10:13	130.22	4764
1979	Bill Rodgers	2:09:27	129.45	7927
1980	Bill Rodgers	2:12:11	132.18	5471
1981	Toshihiko Seko	2:09:26	129.43	6881
1982	Alberto Salazar	2:08:52	128.87	7647
1983	Greg Meyer	2:09:00	129.00	6674
1984	Geoff Smith	2:10:34	130.57	6924
1985	Geoff Smith	2:14:05	134.08	5595
1986	Rob de Castella	2:07:51	127.85	4904
1987	Toshihiko Seko	2:11:50	131.83	6399
1988	Ibrahim Hussein	2:08:43	128.72	6758
1989	Abebe Mekonnen	2:09:06	129.10	6458
1990	Gelinda Bordin	2:08:09	128.15	9412
1991	Ibrahim Hussein	2:11:06	131.10	8686
1992	Ibrahim Hussein	2:08:14	128.23	9629
1993	Cosmas Ndeti	2:09:33	129.55	8930
1994	Cosmas Ndeti	2:07:15	127.25	9059
1995	Cosmas Ndeti	2:09:22	129.37	9416
1996	Moses Tanui	2:09:15	129.25	38 708
1997	Lameck Aguta	2:10:34	130.57	10 471
1998	Moses Tanui	2:07:34	127.57	11 499
1999	Joseph Chebet	2:09:52	129.87	12 797
2000	Elijah Lagat	2:09:47	129.78	17 813
2001	Lee Bong-Ju	2:09:43	129.72	15 606
2002	Rodgers Rop	2:09:02	129.03	16 936
2003	R. Cheruiyot	2:10:11	130.18	17 567

significant, for example, introduction of prizes for the winners, etc. But as was pointed out earlier, one could account for these factors by increasing or decreasing $\Lambda(t)$ or by introducing additional terms.

Another interesting question is the limit of human ability and performance. This question relates to the idea of a possible ultimate record. In terms of what is discussed here, the ultimate record is the one that will survive forever, i.e. its survival probability is 1. Since every record will eventually be broken (provided that we can measure continuously, especially for increasing number of attempts) it is more practical to think of a survival time, e.g. 50 or 100 years and a survival probability

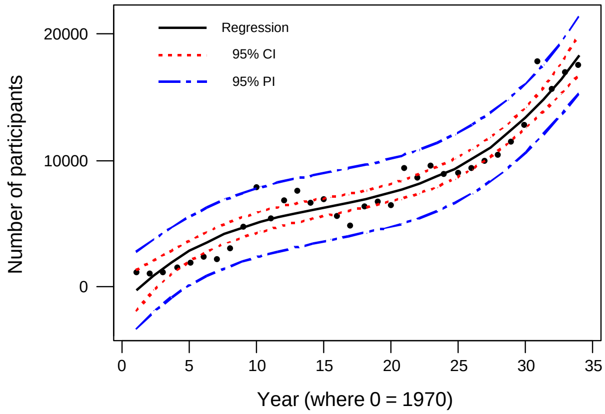

Fig. 1. Boston Marathon 1970–2003.

larger than, say 90%, for a record to be considered an ultimate record. The approach is under investigation.

Reference

[1] S.M. Berry, A statistician reads the sports pages, Chance 15 (2) (2002) 49–53.